



People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research



AKLI MOHAND OULHADJ UNIVERSITY -BOUIRA-

Faculty of Exact Sciences

Department of Computer Science

Master Thesis

In Computer Science

Speciality: Information Systems and Software Engineering -ISIL-

Theme

Rumor Detection Platform

Supervised by

- BADIS Lyes

Realized by

- CHETTIH Rakan Karim
- GRIB Abderrahmane

2025/2026

Acknowledgments

First and foremost, we express our deepest gratitude to Almighty God for granting us the strength, health, patience, and perseverance necessary to complete this work. His guidance and blessings have been a constant source of support throughout this journey.

We would like to extend our sincere appreciation to our families for their unconditional love, continuous encouragement, and unwavering support. Their sacrifices, understanding, and belief in our abilities have been the cornerstone of our success and a constant source of motivation.

Our sincere thanks go to our supervisor, Dr. Badis Lyes, for his valuable guidance, insightful advice, constructive feedback, and continuous support throughout the realization of this thesis. His dedication and encouragement have greatly contributed to the completion of this work.

We are also deeply grateful to all our teachers and professors who have shared their knowledge and expertise with us over the years. Their efforts have played a significant role in our academic and personal development.

We would like to extend our appreciation to the members of the jury for dedicating their time to evaluate this work and for their valuable comments and suggestions.

Finally, we would like to thank the entire staff of the Computer Science Department for providing a supportive academic environment and for their efforts throughout our years of study.

Dedications

In the name of Allah, the Most Gracious, the Most Merciful. First and foremost, I thank Allah Almighty for His guidance and assistance throughout this journey, which without His grace would never have been completed.

I dedicate this work to my dear mother, who was never just a source of support, but the greatest strength I have ever known, in my life as a whole and in my academic path in particular, especially after the passing of my father, may Allah have mercy on his soul.

My father, Tahar Allah Yerhamo. No words I write could ever repay what he gave me. He was the father who taught me to smile and to laugh in the face of hardship. Even now, I still see his smile clearly, and I wonder how proud he would have been to see me reach this moment. May Allah have mercy on him and reward him for every single thing.

To my kind sister Ikram, who every single day pushed me, encouraged me, and reminded me why I started, and to my brothers Abdelhamid, Mounir, Oussama, and Hakim, who stood by me and showed me the true meaning of family when I needed it most.

To my colleague, my partner through it all, and my friend, Abderrahmane thank you for your patience, your support, and for never letting me give up. This work is as much yours as it is mine. May Allah bless you.

To Dr. Badis Ilyas, you were more than a supervisor. You were a teacher, a guide, and a friend who walked with us from the first lecture all the way to this final page. Thank you for your time, your wisdom, and your dedication.

To my friends, professors, and everyone in the Computer Science Department who supported us, your kindness meant more than words can say. May Allah bless you all.

Chettih Rakan Karim

Dedications

By the grace of Allah, the Most Gracious, the Most Merciful, to whom all praise and gratitude belong. By His will and guidance, this work was made possible. I dedicate these humble pages:

To my irreplaceable beloved parents, for their unconditional love, endless patience, silent sacrifices, and prayers that never ceased. No words could ever repay what you have given. May Allah preserve you, reward you abundantly, and grant you health and happiness.

To my dear brother and two sisters, for their affectionate presence, their encouragement, and their pride that kept me going. You have been a quiet strength throughout this journey. May Allah reward you immensely.

To Rakan, my dear friend first and foremost and my partner in this work. You gave your time, your patience, and your energy without hesitation. Whatever this work is worth, you are part of its value. May Allah reward you and bless your path ahead.

To our supervisor Dr. Badis Lyes, who was more than a guide. Your dedication, patience have left a mark that goes beyond these pages. May Allah reward you for every effort you gave.

To all those who contributed, directly or indirectly, to the completion of this work, with a word, a piece of advice, or a prayer, we thank you from the bottom of our hearts. May Allah make this work beneficial, accept it as a sincere effort, and reward everyone who played a part in bringing it to life.

Grib abderrahmane

Abstract

The rapid spread of rumors across digital platforms poses serious threats to public health, political stability, and social trust. Existing detection approaches, whether manual fact-checking, automated AI models, or community moderation, each address the problem only partially, suffering respectively from lack of scalability, limited explainability, and vulnerability to manipulation. This thesis presents the design and implementation of *Rumify*, a hybrid rumor verification platform that integrates three complementary layers: an AI-based analysis module combining a transformer text encoder and a GRU-based source credibility model to generate preliminary credibility and certainty scores, a community-based verification module where user contributions are weighted by a dynamic reputation system, and an expert validation module that issues final authoritative verdicts in a transparent and traceable manner while automatically recalibrating community reputation scores based on the alignment of user evaluations with the final decision.. A key contribution of this work is an original trust-weighted ranking algorithm extending the IARR method with Bayesian averaging to robustly combine machine learning predictions with community credibility signals. A multi-model feedback quality estimation architecture is also proposed to evaluate user commentary based on text correctness, thoroughness, and reasoning quality. The developed prototype demonstrates the feasibility of integrating AI analysis, participatory verification, and expert oversight within a unified verification pipeline.

Keywords: Rumor detection, misinformation, hybrid verification, reputation systems, natural language processing, community-based verification, expert validation, Bayesian averaging, credibility assessment, information integrity.

Résumé

La propagation rapide des rumeurs sur les plateformes numériques constitue une menace importante pour la santé publique, la stabilité politique et la confiance sociale. Les approches existantes de détection, qu'il s'agisse de la vérification manuelle des faits, des modèles d'intelligence artificielle ou de la modération communautaire, ne répondent que partiellement au problème en raison, respectivement, de leur manque de passage à l'échelle, de leur explicabilité limitée et de leur vulnérabilité aux manipulations.

Ce mémoire présente la conception et l'implémentation de *Rumify*, une plateforme hybride de vérification des rumeurs intégrant trois couches complémentaires : un module d'analyse basé sur l'intelligence artificielle combinant un encodeur de texte de type Transformer et un modèle de crédibilité des sources fondé sur un réseau neuronal récurrent GRU afin de produire des scores préliminaires de crédibilité et de certitude ; un module de vérification communautaire dans lequel les contributions des utilisateurs sont pondérées par un système dynamique de réputation ; ainsi qu'un module de validation par des experts chargé d'émettre des verdicts finaux faisant autorité de manière transparente et traçable, tout en recalibrant automatiquement les scores de réputation de la communauté en fonction de l'alignement des évaluations des utilisateurs avec la décision finale.

L'une des principales contributions de ce travail est la proposition d'un algorithme original de classement pondéré par la confiance, étendant la méthode IARR à l'aide du moyennage bayésien afin de combiner de manière robuste les prédictions issues de l'apprentissage automatique avec les signaux de crédibilité de la communauté. Une architecture multi-modèles d'évaluation de la qualité des retours utilisateurs est également proposée afin d'évaluer les commentaires selon trois critères : la correction linguistique, l'exhaustivité et la qualité du raisonnement.

Le prototype développé démontre la faisabilité de l'intégration de l'analyse par intelligence artificielle, de la vérification participative et de l'expertise humaine au sein d'un pipeline unifié de vérification de l'information.

Mots-clés : Détection de rumeurs, désinformation, vérification hybride, systèmes de réputation, traitement automatique du langage naturel, vérification communautaire, validation experte, moyennage bayésien, évaluation de la crédibilité, intégrité de l'information.

ملخص

يمثل الانتشار السريع للشائعات عبر المنصات الرقمية تهديدًا خطيرًا للصحة العامة والاستقرار السياسي والثقة الاجتماعية. وتتعامل الأساليب الحالية لكشف الشائعات، سواء كانت تعتمد على التحقق اليدوي من الحقائق أو نماذج الذكاء الاصطناعي أو المراجعة المجتمعية، مع المشكلة بشكل جزئي فقط، إذ تعاني على التوالي من محدودية قابلية التوسع، وضعف قابلية التفسير، والتعرض للتلاعب.

تقدم هذه الرسالة تصميم وتنفيذ منصة rumify، وهي منصة هجينة للتحقق من الشائعات تعتمد على ثلاثة مستويات متكاملة: وحدة تحليل قائمة على الذكاء الاصطناعي تجمع بين نموذج ترميز نصوص من نوع Transformer ونموذج لمصادقية المصادر يعتمد على شبكة عصبية متكررة من نوع GRU لإنتاج درجات أولية للمصادقية واليقين؛ ووحدة تحقق مجتمعية تُوزن فيها مساهمات المستخدمين بواسطة نظام سمعة ديناميكي؛ بالإضافة إلى وحدة تحقق خبراءية تصدر أحكامًا نهائية موثوقة بطريقة شفافة وقابلة للتتبع، مع إعادة معايرة درجات سمعة المستخدمين تلقائيًا وفقًا لمدى توافق تقييماتهم مع القرار النهائي.

وتتمثل إحدى المساهمات الرئيسية لهذا العمل في اقتراح خوارزمية ترتيب أصلية تعتمد على الثقة، وتُعد امتدادًا لطريقة IARR باستخدام المتوسط البايزي، بهدف الدمج المتين بين تنبؤات التعلم الآلي وإشارات المصادقية المستمدة من المجتمع. كما تم اقتراح بنية متعددة النماذج لتقييم جودة تعليقات المستخدمين بالاعتماد على ثلاثة معايير أساسية هي: سلامة النص، والشمولية، وجودة الاستدلال المنطقي.

ويبرهن النموذج الأولي المطور على إمكانية دمج تحليل الذكاء الاصطناعي، والتحقق التشاركي، والخبرة البشرية ضمن مسار موحد للتحقق من المعلومات.

الكلمات المفتاحية: كشف الشائعات، التضليل الإعلامي، التحقق الهجين، أنظمة السمعة، معالجة اللغة الطبيعية، التحقق المجتمعي، التحقق الخبير، المتوسط البايزي، تقييم المصادقية، نزاهة المعلومات.

Contents

Contents	i
List of Figures	vii
List of Tables	ix
List of Acronyms	xi
General introduction	1
1 Rumors	4
1.1 Introduction	4
1.2 Rumors definition	5
1.3 Typologies of Rumors	7
1.3.1 Thematic Categories	7
1.3.2 Structural or Truth-Based Categories	8
1.3.3 Temporal Categories	8
1.4 Spread and Propagation of Rumors	9
1.4.1 Mechanisms of Spread	9
1.4.2 Role of Influencers and Bots	10
1.4.3 Platform Dynamics	10
1.4.4 Speed vs. Correction	11
1.5 Evolution and Mutation of Rumors	11
1.5.1 Mechanisms of Mutation	11
1.5.2 Digital Platforms and Rumor Mutation	12

1.5.3	Collective Reinforcement and Memory Distortion	13
1.5.4	Implications for Detection	13
1.6	Sociocultural Role of Rumors	13
1.6.1	Coping with Uncertainty and Anxiety	14
1.6.2	Meaning-Making and Social Narratives	14
1.6.3	Political Control and Resistance	14
1.6.4	Community Bonding and Group Identity	15
1.6.5	Cultural Variations in Rumor Dynamics	15
1.7	Consequences of Rumors	15
1.7.1	Public Health Threats	15
1.7.2	Political and Electoral Manipulation	16
1.7.3	Social Tension and Violence	16
1.7.4	Economic Disruption	16
1.7.5	Reputational Damage	16
1.7.6	Erosion of Trust in Institutions	17
1.8	Challenges in Addressing Rumors	17
1.8.1	Speed and Volume of Information	17
1.8.2	Variability in Rumor Structure	18
1.8.3	Source and Intent Ambiguity	18
1.8.4	Platform Limitations and Privacy Barriers	18
1.8.5	Adversarial Adaptation	18
1.8.6	Ethical and Legal Constraints	18
1.8.7	Human Behavior and Confirmation Bias	19
1.9	Conclusion	19
2	Rumor Detection Techniques and Systems	21
2.1	Introduction	21
2.2	Definition of Rumor Detection	22
2.3	Manual and Human-Centered Detection	23
2.3.1	Methodology and Workflow	23
2.3.2	Strengths of Manual Approaches	24
2.3.3	Limitations	24
2.3.4	Illustrative Examples of Fact-Checking Platforms	24

2.4	AI-Based Detection Models	26
2.4.1	Support Vector Machines (SVM)	26
2.4.2	Recurrent Neural Networks (LSTM/GRU)	26
2.4.3	Transformer-Based Models	26
2.4.4	Multimodal Detection Approaches	27
2.4.5	Limitations and Challenges	27
2.5	Hybrid and Collaborative Detection Models	28
2.5.1	Concept and Motivation	28
2.5.2	Real-World Examples	29
2.5.3	Strengths and Limitations	29
2.6	Community Moderation and Crowdsourcing	29
2.6.1	Key Features of Community Moderation	30
2.6.2	Real-World Examples	30
2.6.3	Benefits and Risks	31
2.7	Approaches to Combating Rumors	31
2.7.1	Technological Interventions	31
2.7.2	Education and Media Literacy Campaigns	32
2.7.3	Community-Driven and Platform-Based Moderation	32
2.7.4	Legal and Policy-Based Measures	33
2.8	Challenges in Rumor Detection	33
2.8.1	Volume and Velocity of Content	33
2.8.2	Linguistic and Cultural Complexity	33
2.8.3	Ambiguity and Lack of Ground Truth	34
2.8.4	Adversarial and Evasive Techniques	34
2.8.5	Platform Restrictions and Data Access	34
2.8.6	Ethical, Legal, and Social Dilemmas	35
2.9	Recent Advances and Research Trends	35
2.9.1	Transformer-Based Models and Deep Learning	35
2.9.2	Graph-Based and Propagation-Aware Models	36
2.9.3	Multimodal and Cross-Platform Detection	36
2.9.4	Early Detection and Weak-Signal Modeling	36
2.9.5	Explainability and Human-AI Interaction	36

2.9.6	Adversarial Robustness and Low-Resource Adaptation	37
2.10	Identified Gaps and Opportunities	37
2.10.1	Fragmentation of Solutions	37
2.10.2	Lack of Integration Between AI, Community, and Experts	37
2.10.3	Weak Multilingual and Regional Adaptation	38
2.10.4	Limited Explainability and Transparency	38
2.10.5	Dataset Bias and Evaluation Gaps	38
2.10.6	Ethics and Governance	38
2.11	Conclusion	39
3	Conceptual Framework	41
3.1	Introduction	41
3.2	General System Overview	42
3.2.1	User-System Interaction Flow	42
3.2.2	Use Case Diagram	43
3.2.3	Global Functional Workflow	44
3.3	General System Design and Overview	45
3.4	Community-Based Verification Module	47
3.4.1	Objectives and Role	48
3.4.2	Architecture and Integration	48
3.4.3	User Roles and Permissions	49
3.4.4	Credibility and Reputation Mechanism	50
3.4.5	Interface and Interaction Design	50
3.4.6	Advantages and Limitations	51
3.5	Expert Validation Module	53
3.5.1	Objectives and Responsibilities	53
3.5.2	Workflow and Integration	54
3.5.3	Expert Dashboard and Tools	55
3.5.4	Feedback Loop to AI and Community	56
3.5.5	Governance and Ethical Considerations	58
3.6	System Metrics Overview	59
3.6.1	Supplementary Metrics	60
3.7	Feedback Quality Estimation	61

3.7.1	Quality Criteria for Feedback Evaluation	61
3.7.2	Multi-Model Evaluation Architecture	62
3.8	Trueness Estimation Framework	65
3.8.1	AI Prediction Model	66
3.8.2	Integration with Ranking System	68
3.8.3	Summary	69
3.8.4	Rating System	69
3.8.5	Our Approach to a Rating System	76
3.9	Dynamic Reputation Computation Framework	79
3.9.1	Theoretical Algorithm	80
3.9.2	Static vs Dynamic Execution Strategies	80
3.9.3	Challenges and Limitations	81
3.9.4	Proposed Solutions	83
3.9.5	Business Rules	84
3.10	System Architecture	85
3.10.1	Core Application	85
3.10.2	Domain Layer	86
3.11	Global System Design	89
3.11.1	Data Stores	89
3.11.2	Search Service	91
3.11.3	Iterative Algorithm and Aggregation Execution	92
3.12	Conclusion	93
4	Implementation and Testing	94
4.1	Introduction	94
4.2	Work Environment	94
4.2.1	Hardware Environment	94
4.2.2	Software and Libraries	95
4.3	Application Presentation	95
4.3.1	Home Page	96
4.3.2	Browse and Search Interface	96
4.3.3	Rumor Submission Interface	97
4.3.4	Rumor Analysis and Verification Interface	99

4.3.5 Expert Dashboard	101
4.4 Current Limitations	103
4.5 Conclusion	103
General conclusion and perspectives	105
Bibliographie	108

List of Figures

1.1	Examples of true, false, and unverified rumors related to the Iran–Israel conflict.	7
1.2	Spread and propagation of rumors across digital platforms, from initial source to viral amplification.	10
1.3	Illustration of the stages in the evolution and mutation of a rumor.	12
2.1	Homepage of the Fatabyyano fact-checking platform.	25
2.2	Homepage of the U.S.-based fact-checking platform FactCheck.org.	25
3.1	General flowchart of user and system interaction.	42
3.2	General use case diagram.	43
3.3	Global functional workflow.	44
3.4	Service interaction.	45
3.5	High-level application architecture.	46
3.6	Ranking-involved services.	47
3.7	Illustration of Multi-model architecture for feedback quality estimation.	63
3.8	Bipartite rating network.	70
3.9	User’s reputation by number of ratings.	82
3.10	Illustration of overlap between sets.	83
3.11	Sequential algorithm runs in conjunction with aggregation operations.	84
3.12	Layered architecture for the core application.	85
3.13	Class diagram representing the domain.	86
3.14	Application layer services.	88
3.15	Detailed global system design.	89

3.16 Data-stores of the system.	90
3.17 Cache structure.	91
3.18 Close look at search service.	92
3.19 Iterative algorithm execution process.	92
3.20 Aggregation operation execution process.	93
4.1 Home page of the rumor verification platform.	96
4.2 illustrates the search interface.	97
4.3 illustrates the Browse	97
4.4 Rumor submission interface.	98
4.5 Ask a Question interface.	99
4.6 Rumor analysis and verification interface (Part 1).	100
4.7 Rumor analysis and verification interface (Part 2: Community reviews and expert verdict).	101
4.8 Expert dashboard showing pending rumors awaiting verification.	102
4.9 Expert dashboard showing the history of issued verdicts.	102

List of Tables

1.1	Distinctions between rumors and other forms of misleading information. . .	6
2.1	Comparison of AI-Based Rumor Detection Models	28
2.2	Strengths and limitations of hybrid rumor detection systems	30
2.3	Benefits and Risks of Community-Based Moderation Models	32
3.1	Advantages and limitations of the hybrid rumor detection platform.	52
4.1	Hardware Environment	95
4.2	Libraries and Tools Used	95

List of Acronyms

NLP	Natural Language Processing
IFCN	International Fact-Checking Network
WHO	World Health Organization
EPI-WIN	WHO Information Network for Epidemics
UNGP	United Nations Global Pulse
UN	United Nations
AI	Artificial Intelligence
SVM	Support Vector Machines
RNN	Recurrent neural network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Units
BERT	Bidirectional Encoder Representations from Transformers
AraBERT	Transformer-based Model for Arabic Language Understanding
CAMeLBERT	A collection of pre-trained models for Arabic NLP
UNESCO	United Nations Educational, Scientific and Cultural Organization
GNNs	Graph Neural Networks
GATs	Graph Attention Networks
XIA	Explainable Artificial Intelligence
API	Application Programming Interface
RoBERTa	Robustly Optimized BERT Pretraining Approach
XLM-RoBERTa	Cross-lingual Robustly Optimized BERT Pretraining Approach
CLIP	Contrastive Language-Image Pretraining
UML	Unified Modeling Language
IARR	Iterative Algorithm with Reputation Redistribution)

General introduction

The rapid expansion of digital communication technologies has fundamentally transformed the way information is produced, shared, and consumed. Social media platforms such as X (formerly Twitter), Facebook, and WhatsApp have become the primary channels through which billions of people receive news and make sense of world events. While this connectivity has brought undeniable benefits in terms of information accessibility, it has simultaneously created fertile ground for the uncontrolled spread of unverified and false information. Among the most persistent and damaging forms of this phenomenon are rumors, which are unverified claims that circulate rapidly across networks before their authenticity can be established. From false health advice during the COVID-19 pandemic, to fabricated political narratives during elections, to WhatsApp-fueled mob violence in rural communities, the real-world consequences of unchecked rumors are severe and well-documented.

Addressing this challenge is far from straightforward. Existing solutions each bring meaningful contributions but remain individually insufficient. Manual fact-checking platforms, such as Fatabyyano or FactCheck.org, offer high accuracy and contextual judgment, but cannot scale to the volume and velocity of content generated online. Automated AI-based systems, ranging from classical support vector machines to state-of-the-art transformer models like BERT, enable fast and large-scale detection, yet they suffer from limited explainability, poor performance in multilingual or culturally specific contexts, and vulnerability to adversarial manipulation. Community-based moderation approaches, as seen on Reddit or through X's Community Notes, democratize the process and provide real-time contextual input, but remain susceptible to coordinated bias, groupthink, and varying levels of user expertise. Perhaps most critically, these approaches are rarely inte-

grated with one another. Solutions tend to be fragmented, addressing only one dimension of the problem, whether content, propagation, or credibility, without forming a unified and accountable framework.

In response to these identified gaps, this thesis proposes the design and implementation of Rumify, a hybrid rumor detection and verification platform that integrates three complementary layers into a unified pipeline: an AI-based NLP analysis module, a community-based verification module, and an expert validation module. Rather than relying on any single approach, the system is designed so that each layer feeds into and reinforces the next. The AI module provides a rapid preliminary credibility assessment and certainty score. The community module enables participatory review, where user votes and feedback are weighted according to a dynamic reputation system based on an adapted iterative ranking algorithm. The expert module provides the final authoritative verdict, which automatically triggers a recalibration of user credibility scores across the community layer based on the proximity of their evaluations to the final decision. Among the original contributions of this work is the design of a custom trust-weighted ranking algorithm that builds upon and extends the IARR method by incorporating Bayesian averaging to combine machine learning predictions with community credibility signals. Additionally, a multi-model feedback quality estimation module is proposed to assess the intrinsic value of user commentary based on text correctness, thoroughness, and reasoning quality, rather than mere engagement metrics.

This thesis is organized into four chapters. Chapter 1 provides a foundational understanding of the rumor phenomenon itself, covering its definitions from interdisciplinary perspectives, its typologies, the mechanisms through which it propagates and mutates across digital platforms, its sociocultural functions, its real-world consequences, and the core challenges that make it difficult to address. Chapter 2 surveys the state of the art in rumor detection, examining manual, AI-based, hybrid, and community-driven approaches, along with recent research trends and a critical analysis of the gaps that motivate this work. Chapter 3 presents the conceptual design and architecture of the proposed platform, detailing the community-based verification module, the expert validation module, the ranking algorithm, the trueness estimation method, the feedback quality evaluation module, and the overall layered system architecture. Chapter 4 covers the implementation and testing of the platform, presenting the development environment, the tools and

technologies used, the main application interfaces, and the results of functional testing across the system's core components.

Rumors

1.1 Introduction

In today’s hyperconnected digital society, the speed with which information spreads is unprecedented. While this connectivity has revolutionized communication, it has also facilitated the rapid and uncontrolled diffusion of unverified or false information, commonly known as rumors. These rumors can emerge suddenly, evolve rapidly, and influence individuals, communities, and even governments before their authenticity can be verified. The prevalence of such rumors on social media platforms like X (formerly Twitter)[1] , Facebook, and WhatsApp has raised significant concerns across various domains, including public health, politics, economics, and crisis response.

Rumors are not a new phenomenon, as they have existed throughout history as a mechanism for individuals to make sense of uncertain or ambiguous situations. However, in the context of modern technology, the scale, velocity, and impact of rumor propagation have dramatically increased. A single misleading post can reach thousands, if not millions, within minutes, causing panic, misinformed decisions, or reputational damage.

This chapter aims to provide a foundational understanding of the rumor phenomenon. It begins by exploring various definitions of rumors from interdisciplinary perspectives, followed by a detailed look at typologies based on their themes, structure, and context. It also examines the mechanisms of rumor propagation, the ways in which rumors evolve and mutate, and their sociocultural significance. Finally, the chapter highlights the real-world consequences of unchecked rumors and the challenges in controlling their spread.

Understanding the nature and behavior of rumors is a critical first step in designing

effective detection and intervention systems. This sets the stage for the next chapter, which will delve into the methods and technologies developed to detect, classify, and mitigate rumors in digital environments.

1.2 Rumors definition

The term rumor refers to unverified or doubtful information that circulates informally among individuals, especially during periods of uncertainty or crisis. Scholars across disciplines, such as psychology, sociology, communication studies, and political science, have provided various definitions of the concept, each highlighting different aspects of its nature.

One of the most widely cited definitions comes from social psychologists Allport and Postman [2], who defined a rumor as “a specific proposition for belief passed along from person to person, usually by word of mouth, without secure standards of evidence being present.” [2] This definition emphasizes both the lack of verification and the social transmission nature of rumors. Similarly, DiFonzo and Bordia [3] define rumors as “unverified and instrumentally relevant information statements in circulation that arise in contexts of ambiguity, danger, or potential threat.” [3].

While rumors initially spread without verification, they are not inherently false, as some are eventually confirmed as accurate based on later developments or reliable sources. In general, rumors share several key features:

Unverified status: They lack confirmation or denial from authoritative sources at the time of circulation [3].

Relevance and urgency: Rumors often concern topics of emotional, social, or political significance [2].

Believability: They tend to include just enough plausibility to be accepted by the public without critical evaluation [4].

Informal transmission: They are passed through interpersonal networks, and today, predominantly through social media platforms [5].

It is also important to differentiate rumors from other forms of misleading or unverified information, As shown in Table 1.1

Term	Definition
Fake News	Deliberately fabricated information presented as news, usually with the intent to mislead.
Misinformation	Incorrect or false information spread without malicious intent.
Disinformation	False information shared with deliberate intent to deceive.
Hoaxes and Gossip	Often target individuals or trivial topics, but may also spread in patterns similar to rumors.
Rumors	Typically shared by individuals who genuinely believe them to be true, making them harder to detect or debunk.

Table 1.1: Distinctions between rumors and other forms of misleading information.

In the digital era, the distinctions between these forms have become increasingly blurred. A post may begin as a rumor and evolve into misinformation or disinformation depending on how it is interpreted, repeated, or manipulated. This complexity underscores the importance of analyzing rumors as a distinct and evolving phenomenon in the study of information integrity.

Example: We present an illustrative example of a true rumor, a false rumor, and unverified rumors in Figure 1.1. All the stories in this figure revolve around the ongoing conflict between Iran and Israel in 2024–2025.

In contrast, a post claiming “*Iranian missiles have hit Israeli airbases*” is *fact-checking worthy*, and its veracity was later confirmed to be **true** by multiple news agencies. A second post claiming that “*a viral drone explosion video shows Iran striking Israel*” qualifies as a *rumor*, but its veracity was proven to be **false**, as the footage originated from a 2022 event in Kyiv, Ukraine. Meanwhile, posts speculating that “*Russia forces will join Iran militarily*” or that “*ships are ablaze in the Gulf of Oman due to the war*” are **unverified rumors**; they circulate widely but lack official confirmation or reliable evidence at this time.



Figure 1.1: Examples of true, false, and unverified rumors related to the Iran–Israel conflict.

1.3 Typologies of Rumors

Rumors can take many forms depending on their subject matter, structure, context, and intent. Understanding these typologies is crucial for building effective detection systems, as each type may exhibit different linguistic patterns, spread behaviors, and psychological appeal. Scholars have proposed several classification schemes, which we organize here into three main categories: thematic, structural, and temporal.

1.3.1 Thematic Categories

This classification groups rumors by their content or domain of concern, particularly during crises and high-impact events:

Health-related rumors: These include false claims about diseases, treatments, or vaccines (e.g., “drinking hot water cures COVID-19”). Such rumors are especially dangerous during pandemics, where they may cause individuals to avoid proper medical care [6].

Political rumors: These often involve elections, scandals, or conspiracies and may aim to manipulate public opinion [7]. For instance, unfounded allegations of election fraud have had major political consequences in several democracies.

Crisis/disaster rumors: Spread during natural disasters, wars, or civil unrest. They may falsely claim safety threats or emergency procedures, often leading to panic [8].

Social rumors: These include unverified stories about celebrities, public figures, or trending events. Although seemingly less harmful, they contribute to erosion of trust in public discourse.

This typology helps researchers tailor detection tools to the linguistic and emotional tone typical of each domain.

1.3.2 Structural or Truth-Based Categories

From the perspective of accuracy and intent, rumors may be classified as:

False rumors: Fully inaccurate and later debunked by authoritative sources.

Misleading or ambiguous rumors: Contain partial truths, exaggerations, or decontextualized information [9]. A subtype of these misleading rumors involves article headlines, which exaggerate, omit, or distort information to provoke clicks or mislead readers. These may function as intentional misinformation or arise from poor editorial practices.

Satirical/spoof content: Includes jokes or parodies that are misinterpreted as real information [10].

Disinformation: Deliberate creation and dissemination of falsehoods for political or financial gain.

Some frameworks also distinguish between malicious vs. non-malicious rumors, based on whether the spreader intends harm [11].

1.3.3 Temporal Categories

Rumors may also be categorized based on their lifespan and relation to events:

Spontaneous rumors: Emerge unexpectedly in moments of uncertainty.

Event-triggered rumors: Appear in response to a specific event (e.g., a political assassination or natural disaster).

Persistent or long-term rumors: Often take the form of conspiracy theories and may survive for years, adapting to new events [5].

These rumors evolve over time through reinterpretation, mutation, or reinforcement within closed communities.

Understanding these classifications not only improves the ability of artificial intelligence (AI) systems to detect and categorize rumors but also offers insight into the social psychology behind their formation and spread.

1.4 Spread and Propagation of Rumors

The propagation of rumors has been deeply transformed by the evolution of communication technologies. In traditional contexts, rumors spread orally, through interpersonal networks in homes, streets, or workplaces. Today, this process has been accelerated and amplified by digital platforms, where a single message can go viral in seconds. Understanding how rumors propagate is crucial for designing effective detection and intervention strategies.

1.4.1 Mechanisms of Spread

Rumors typically follow a peer-to-peer transmission model, in which individuals share information with others they trust, often without verification [2]. In the age of social media, this model has scaled up dramatically. Platforms like X, Facebook, WhatsApp, and Telegram allow users to share content instantly with large and decentralized audiences. The anonymity, low-cost sharing, and emotional appeal of digital communication increase the likelihood of viral spread [12].

Additionally, the social reinforcement loop, in which individuals tend to believe information that aligns with their existing beliefs or is shared by peers, contributes to the persistence and amplification of rumors. This process is further exacerbated by confirmation bias, where users selectively consume or forward information that supports their worldview [13] As shown in Figure 1.2 .

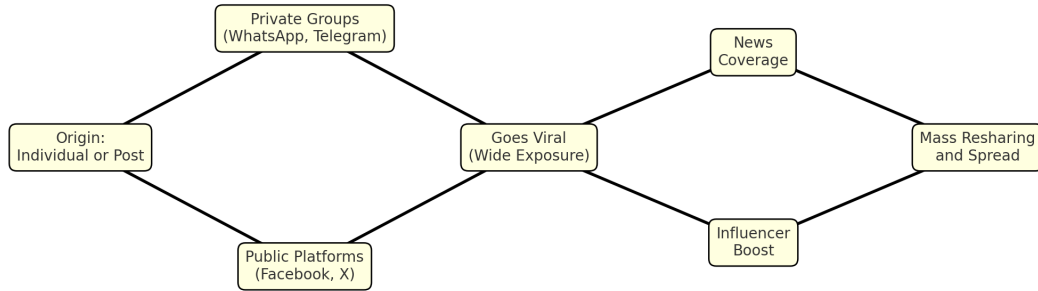


Figure 1.2: Spread and propagation of rumors across digital platforms, from initial source to viral amplification.

1.4.2 Role of Influencers and Bots

Certain users have disproportionate influence on rumor spread. These include:

- Influencers or public figures, whose posts are widely shared and trusted.
- Verified accounts, which often lend perceived legitimacy to the content.
- Bots and troll accounts can act as “super-spreaders” by amplifying rumors at scale. Shao et al. (2018) found that bots comprised only 6% of accounts sharing low-credibility content, yet were responsible for 31% of such content and frequently participated early in rumor cascades [14].

These actors often act as “super-spreaders,” accelerating the reach of a rumor beyond what would occur in a purely organic network.

1.4.3 Platform Dynamics

Different platforms contribute differently to rumor dynamics:

- X encourages fast, viral spread through retweets.

- Facebook supports threaded conversations and personalized newsfeeds, where rumors can persist within echo chambers.
- WhatsApp and Telegram use closed groups, making rumors harder to detect and track [15].

Each platform provides specific features, including visibility, speed, and group formation, that influence how rumors spread through the information ecosystem.

1.4.4 Speed vs. Correction

False information spreads faster, deeper, and more broadly than truthful information. In a large-scale analysis, Vosoughi et al (2018) found that false news stories on X (then known as Twitter) were 70% more likely to be retweeted and reached approximately 1,500 users about six times faster than true news[12]. While corrections and fact-checks do exist, they typically receive much less engagement and arrive too late to undo the initial damage.

This “first mover advantage” in rumor propagation emphasizes the need for early detection and real-time monitoring.

1.5 Evolution and Mutation of Rumors

Rumors are not static; they tend to evolve, adapt, and transform as they are transmitted across individuals and platforms. Much like biological organisms in a population, rumors mutate through the processes of reinterpretation, distortion, simplification, and embellishment. Understanding this evolution is essential for rumor detection systems, which must often deal with multiple versions of the same core message.

1.5.1 Mechanisms of Mutation

Every time a rumor is transmitted, whether through oral communication or online platforms, it may be altered, either deliberately or unconsciously. According to Allport and Postman [2], rumor transformation follows three key psychological laws:

Leveling: The message becomes shorter and more concise over time, with non-essential details dropped.

Sharpening: Certain elements are emphasized or exaggerated.

Assimilation: The content is modified to fit the beliefs, biases, or emotional needs of the communicator.

In online environments, this process is often accelerated. users may summarize, rephrase, translate, or meme-ify content, producing variant versions of the same original rumor. As illustrated in Figure 1.3, a single rumor can undergo various overlapping transformations, ultimately resulting in a highly mutated version.

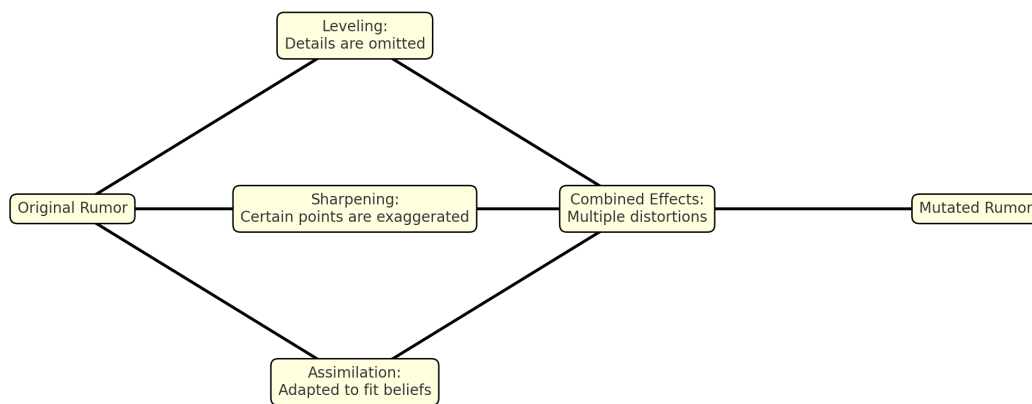


Figure 1.3: Illustration of the stages in the evolution and mutation of a rumor.

1.5.2 Digital Platforms and Rumor Mutation

On digital platforms, the mutation of rumors is amplified by:

- Character limits (e.g., X's 280-character limit)
- Visual reformatting (screenshots, videos, memes)
- Automated summarization or translation
- Language localization (rumors spreading across linguistic boundaries)
- Headline distortion and article framing (e.g., sensational titles that mislead or exaggerate)

A single rumor about a political figure, for example, may appear as a tweet, a meme, a video caption, or a forwarded WhatsApp voice note, each with small alterations that can distort the original meaning. In particular, misleading article headlines can be shared widely without the underlying content being read or critically evaluated. This headline-driven dissemination often strips away nuance, reinforcing misperceptions and catalyzing further rumor propagation.

This fragmentation of the original narrative makes it challenging for detection systems to match variants of the same rumor across platforms and formats [16].

1.5.3 Collective Reinforcement and Memory Distortion

As rumors evolve, they often become more emotionally charged, increasing their spreadability [17]. The process of collective reinforcement, in which users re-share modified versions, leads to the construction of false memories or socially accepted narratives, even when the original statement was inaccurate or had already been debunked.

This highlights a unique challenge: rumor detection is not just about verifying information once, but about tracking and updating evolving narratives in real-time.

1.5.4 Implications for Detection

For detection systems, rumor mutation implies:

- The need for semantic similarity models, not just exact matching.
- The importance of tracking propagation trees or clusters of rumor variants.
- The potential value of multimodal analysis (text, image, audio).

Without accounting for evolution, a detection system may miss emerging variants or mistake a modified rumor for a new one, reducing effectiveness.

1.6 Sociocultural Role of Rumors

Rather than being just communication errors or digital noise, rumors also fulfill important psychological and sociocultural functions. Across societies and throughout history, rumors have played a role in sense-making, community cohesion, resistance, and control.

Understanding these functions reveals why rumors persist even in the presence of corrective mechanisms and why individuals continue to share unverified information despite potential consequences.

1.6.1 Coping with Uncertainty and Anxiety

Rumors often arise in contexts where official information is scarce, delayed, or dis-trusted. In such cases, they function as a collective coping mechanism that helps individuals reduce anxiety and regain a sense of control. This is particularly evident during crises such as pandemics, wars, or natural disasters [3]. In these situations, individuals share rumors to fill the informational void or to prepare for perceived risks, even when the rumors are not verified.

1.6.2 Meaning-Making and Social Narratives

Rumors also help groups construct shared narratives about confusing or controversial events. These narratives may reflect:

- Cultural fears or values
- Collective grievances or suspicions
- Historical memory or trauma

For instance, in some societies, rumors about vaccine safety may be rooted in long-standing distrust toward public health institutions [18]. In this sense, rumors act as vehicles of social interpretation, enabling individuals to express views that may be marginalized or censored in mainstream discourse.

1.6.3 Political Control and Resistance

Rumors can be tools of both power and resistance. Authoritarian regimes may spread state-sponsored rumors to manipulate public opinion or discredit opponents [19], while marginalized communities may use rumors to challenge dominant narratives, especially when other forms of expression are restricted. In this way, rumors become a subtle form of political communication, embedded in asymmetries of power and trust.

1.6.4 Community Bonding and Group Identity

Sharing rumors often strengthens group identity. Whether in small friend groups or large online communities, circulating rumors may signal in-group membership or shared skepticism toward outsiders [20]. This dynamic explains why correcting a rumor can sometimes backfire, especially when it threatens the social cohesion of the group that believes it.

1.6.5 Cultural Variations in Rumor Dynamics

The interpretation and function of rumors also vary across cultures. In some contexts, rumors are seen as morally neutral or even socially necessary. In others, they are stigmatized. These cultural attitudes affect how individuals respond to rumors, whether they correct, ignore, or spread them [21].

Understanding these roles underscores the importance of not treating rumors solely as technical errors in communication systems. Any detection or intervention system must consider the human and cultural dimensions of rumor behavior.

1.7 Consequences of Rumors

While some rumors may appear harmless or trivial, others can have serious and far-reaching consequences at the individual, societal, and institutional levels. The effects of rumors are highly context-dependent, influenced by the nature of the rumor, the platform through which it spreads, and the population it reaches. Understanding these consequences is essential for justifying the need for robust detection and mitigation systems.

1.7.1 Public Health Threats

One of the most dangerous impacts of rumors is in the health sector, where misinformation can lead to real-world harm. During the COVID-19 pandemic, health-related rumors such as false cures, vaccine conspiracies, or fabricated statistics led to widespread confusion, vaccine hesitancy, and resistance to public health measures [6]. In some cases, individuals avoided medical care or engaged in unsafe behaviors based on rumors circulating in private messaging apps and social media.

In Algeria and the Arab world more broadly, such consequences are amplified by the widespread use of closed Facebook group chats, where rumors circulate invisibly to fact-checkers and reach large audiences before any correction is possible.

1.7.2 Political and Electoral Manipulation

Rumors have been widely used to manipulate political discourse and electoral outcomes. False claims about candidates, voter fraud, or foreign interference have been spread deliberately to erode trust in democratic institutions [22]. During critical elections, disinformation campaigns have been known to exploit rumors to polarize the electorate and suppress voter turnout, particularly in vulnerable or minority communities[23].

1.7.3 Social Tension and Violence

In some cases, rumors have been directly linked to real-world violence, including riots, mob lynchings, and hate crimes. For example, in India, false rumors spread via WhatsApp about child kidnappers led to multiple mob killings in rural areas [15]. These incidents demonstrate how digital rumors can escalate into mass hysteria when not addressed promptly and effectively.

1.7.4 Economic Disruption

Rumors can also have significant repercussions on financial markets and consumer behavior. Unverified information about a bank's solvency or a potential shortage of essential goods can provoke panic buying, abrupt stock sell-offs, or digital bank runs. In high-frequency trading environments, where algorithmic systems react to public sentiment within milliseconds, even slight shifts in information tone can produce considerable volatility. As Tetlock [24] demonstrates, media content and investor sentiment are closely correlated, with negative news triggering measurable changes in trading behavior and market performance.

1.7.5 Reputational Damage

Individuals and organizations can suffer serious reputational harm from unverified claims, especially when those claims are amplified online before they can be corrected.

This is particularly dangerous in the case of deepfakes, satirical misinformation, or rumors targeting public figures, where a false allegation can have lasting consequences even after it is disproven [25].

1.7.6 Erosion of Trust in Institutions

At a societal level, widespread rumor circulation contributes to a loss of trust in official sources of information. When rumors spread faster than verified statements, or when authorities fail to respond promptly and credibly, citizens increasingly turn to alternative and often unreliable information ecosystems. Studies have shown that exposure to disinformation correlates with reduced institutional trust and can deepen political polarization, thereby undermining democratic governance and crisis response mechanisms [26, 27].

These consequences demonstrate that rumors are not merely “noise” in the information environment, but act as active agents of disruption, capable of shaping decisions, behaviors, and systemic trust. Recognizing these risks is what motivates the development of proactive rumor detection platforms that integrate artificial intelligence, community participation, and expert verification.

1.8 Challenges in Addressing Rumors

Despite growing awareness of the threat posed by rumors, combating them remains a complex and multifaceted challenge. The rapid evolution of digital platforms, the emotional appeal of false information, and limitations in both human and automated detection systems create significant obstacles to rumor management. This section outlines the major challenges that researchers, developers, and policymakers face in addressing rumors effectively.

1.8.1 Speed and Volume of Information

Rumors now travel at the speed of real-time communication. Platforms like X and WhatsApp allow for instantaneous sharing, often before official sources can respond. A study by Vosoughi et al. [12] found that false news spreads faster and more broadly than

true news. The volume of content, which can reach millions of posts per hour, makes manual monitoring nearly impossible and calls for scalable, automated solutions.

1.8.2 Variability in Rumor Structure

Rumors can be presented in many different linguistic forms, styles, and modalities—from text and video to memes and voice notes. They may include exaggeration, sarcasm, partial truths, or decontextualized facts [10]. This variability poses difficulties for rule-based or static machine learning systems, which often struggle with non-standard language or creative formatting.

1.8.3 Source and Intent Ambiguity

One of the core difficulties in rumor detection is determining the intent and credibility of the source. Is the user maliciously spreading disinformation, unknowingly sharing misinformation, or simply joking? Context is often missing, making automated classification error-prone, especially without access to prior user behavior or network metadata [9].

1.8.4 Platform Limitations and Privacy Barriers

Private messaging platforms such as WhatsApp, Telegram, and Signal are common channels for rumor transmission, but their encrypted nature makes monitoring extremely difficult. Even public platforms have limitations due to API restrictions, content moderation policies, and concerns over user privacy, which can prevent researchers from collecting real-time data at scale [15].

1.8.5 Adversarial Adaptation

As detection systems improve, rumor spreaders adapt their tactics. For example, troll farms and bot networks may change language patterns, insert randomization, or obfuscate URLs to bypass filters. This “cat-and-mouse” dynamic makes it difficult to maintain long-term effectiveness in automated rumor detection [7].

1.8.6 Ethical and Legal Constraints

Efforts to combat rumors must navigate ethical concerns such as:

- Freedom of expression
- Right to privacy
- Avoiding censorship and bias

Overly aggressive content moderation or automated blocking systems may suppress legitimate dissent or minority voices, raising serious questions about fairness, transparency, and accountability [11]. Legal frameworks are often unclear or inconsistent across jurisdictions, complicating cross-platform interventions.

1.8.7 Human Behavior and Confirmation Bias

Perhaps the most difficult challenge is human psychology itself. Individuals are more likely to believe and share information that confirms their preexisting beliefs (confirmation bias), especially when it is emotionally charged [13]. Attempts to correct rumors are sometimes ineffective, or even counterproductive, if they are perceived as partisan or dismissive.

These challenges highlight the need for hybrid approaches that combine technological tools, community involvement, and expert judgment, as proposed in this thesis. A purely technical solution is unlikely to succeed without considering the social, legal, and cognitive dimensions of rumor behavior.

1.9 Conclusion

Rumors are a powerful and pervasive phenomenon that influence how individuals and societies interpret events, especially during times of uncertainty. As explored in this chapter, rumors are not merely falsehoods but complex social artifacts that fulfill psychological, communicative, and political functions. They emerge spontaneously, evolve dynamically, and spread rapidly, often adapting in form, content, and reach as they move through digital environments.

This chapter has provided a foundational understanding of the concept of rumors, including their definitions, typologies, propagation patterns, sociocultural functions, real-world consequences, and the multidimensional challenges they pose. These insights are

essential for designing effective systems that seek not only to detect rumors but also to understand and mitigate their impact.

Addressing rumors requires more than just identifying what is true or false—it demands an approach that accounts for human behavior, social context, and technological limitations. The complexity of rumor dynamics underscores the need for collaborative models that combine automated AI tools, community involvement, and expert judgment, which is the central focus of this thesis. Special emphasis is given to rumors propagated through articles and misleading headlines, which often appear credible and are widely shared without readers engaging with the full content. These forms of rumors pose unique challenges, particularly due to their structured language and viral framing tactics.

In the next chapter, we will delve into the state of the art in rumor detection, exploring the existing technologies, methodologies, and platforms used to identify and evaluate rumors. This review will also highlight techniques tailored to detecting misleading article-based content, laying the groundwork for the development of a hybrid detection system that merges machine learning with human insight.

Rumor Detection Techniques and Systems

2.1 Introduction

The increasing volume and velocity of information shared online have made social media a fertile ground for the rapid spread of rumors and misinformation. In this evolving digital ecosystem, false or unverified content can reach thousands—or even millions—of users before any form of correction or validation is applied. This phenomenon presents serious risks to public health, political stability, economic behavior, and social trust. To address these challenges, the field of rumor detection has become an essential area of research in both academia and industry. Early approaches relied heavily on manual verification carried out by journalists and domain experts. While these methods remain valuable for their accuracy and contextual sensitivity, they are limited by their reliance on human labor, which makes them slow and difficult to scale.

Recent years have seen the emergence of automated rumor detection techniques based on artificial intelligence, natural language processing, and social network analysis. These systems offer faster processing and broader coverage but often struggle with nuances such as sarcasm, cultural context, or conflicting information. Moreover, they can produce results that are difficult to interpret or trust without human input.

In response to these limitations, hybrid models have attracted increasing interest. These approaches combine the scalability and speed of AI with the interpretive capabilities of human judgment, facilitated through community participation, expert moderation, or both. Platforms such as X’s Community Notes and Facebook’s third-party fact-checking initiatives exemplify this shift toward collaborative verification.

This chapter provides a structured review of rumor detection systems and methods. It begins with the definition of rumor detection, followed by an examination of manual, algorithmic, and hybrid approaches. It also explores practical applications of rumor detection, analyzes community-driven moderation systems, and highlights the main challenges in the field. Finally, it presents recent research trends and identifies critical gaps that motivate the development of the hybrid platform proposed in this thesis

2.2 Definition of Rumor Detection

Rumor detection refers to the process of identifying and evaluating unverified or false information circulating among individuals, especially through digital communication channels such as social media, forums, and messaging platforms. The primary goal is to determine the credibility or veracity of a claim before it causes harm or spreads uncontrollably.

In academic literature, rumor detection is commonly situated within broader research areas such as misinformation analysis, fake news detection, and credibility assessment. According to Zubiaga et al. [16], rumor detection involves a multi-step process that includes monitoring social media content, identifying candidate rumors, analyzing their credibility using linguistic and social features, and classifying the information as true, false, or unverified.

A key feature of rumor detection is its focus on early identification, ideally before the rumor becomes widespread or causes damage. This distinguishes it from traditional fact-checking, which often occurs after false claims have already gone viral. As noted by Ma et al. [28], the time-sensitive nature of rumor detection adds complexity to the task, as early content may be sparse, vague, or lacking in context. There are several types of rumor detection approaches:

Manual rumor detection, which relies on journalists, fact-checkers, or expert institutions to verify claims through research and source validation.

Automated rumor detection, which uses computational techniques—especially machine learning and natural language processing (NLP)—to classify claims as true or false based on patterns in the text and social behavior.

Hybrid rumor detection, which combines both approaches by integrating algorithmic

mic predictions with human input from users, moderators, or professional fact-checkers [7].

Rumor detection systems often aim not only to classify claims but also to track how rumors evolve and spread, analyze the users or networks promoting them, and provide explanations or sources that support their assessments. The task is complicated by linguistic ambiguity, misinformation disguised as humor or opinion, and the sheer speed of content generation online.

As a result, rumor detection is increasingly seen as a multi-disciplinary challenge, requiring input from computer science, communication studies, psychology, and even ethics and public policy. Effective systems must be scalable, transparent, and adaptable to different languages, cultures, and platforms.

2.3 Manual and Human-Centered Detection

Before the rise of algorithmic solutions, rumor detection was primarily a manual process, carried out by professional journalists, researchers, and independent fact-checkers. This approach remains essential today and forms the foundation upon which many modern automated systems are built. Manual rumor detection involves identifying questionable claims, researching supporting evidence, and publishing verifiable conclusions. It prioritizes accuracy, transparency, and human judgment, particularly in contexts that require nuanced interpretation.

2.3.1 Methodology and Workflow

Traditional human-centered rumor detection usually follows a structured process:

- Monitoring media and social channels for emerging claims.
- Selecting claims based on virality, potential harm, or public interest.
- Investigating using credible sources, expert interviews, and document analysis.
- Publishing findings with clear explanations and references.

Organizations like Snopes, PolitiFact, FactCheck.org, and international coalitions such as the International Fact-Checking Network (IFCN) exemplify this approach. In the Arab

world, platforms like Fatabyyano and AkhbarMeter have emerged to fill similar roles, using expert teams to verify claims in Arabic across various topics.

2.3.2 Strengths of Manual Approaches

- **Contextual Depth:** Human fact-checkers can interpret sarcasm, cultural references, and local context.
- **Ethical Judgment:** They can apply value-based reasoning to sensitive or ambiguous claims.
- **Public Trust:** Well-established outlets often have reputations that enhance credibility.

2.3.3 Limitations

Despite their strengths, manual approaches face several challenges:

- **Scalability:** The sheer volume of content on social media exceeds human processing capacity.
- **Speed:** Verification can take hours or days, allowing rumors to spread in the meantime.
- **Labor Intensity:** Requires trained personnel, funding, and organizational resources.
- **Limited Reach:** Many fact-checking articles are read only after the rumor has gone viral.

As a result, manual detection, while reliable, is often reactive rather than proactive. These limitations have driven the push toward automated and hybrid solutions that can complement human expertise with algorithmic scale.

2.3.4 Illustrative Examples of Fact-Checking Platforms

fatabyyano.net: is an Arabic fact-checking platform launched in 2014, dedicated to verifying news and rumors on social media. Its name comes from a Quranic verse meaning

“verify”, reflecting its mission to combat misinformation. The platform classifies content as false, misleading, true, or satirical, and is one of the leading fact-checking initiatives in the Arab world, see figure 2.1.

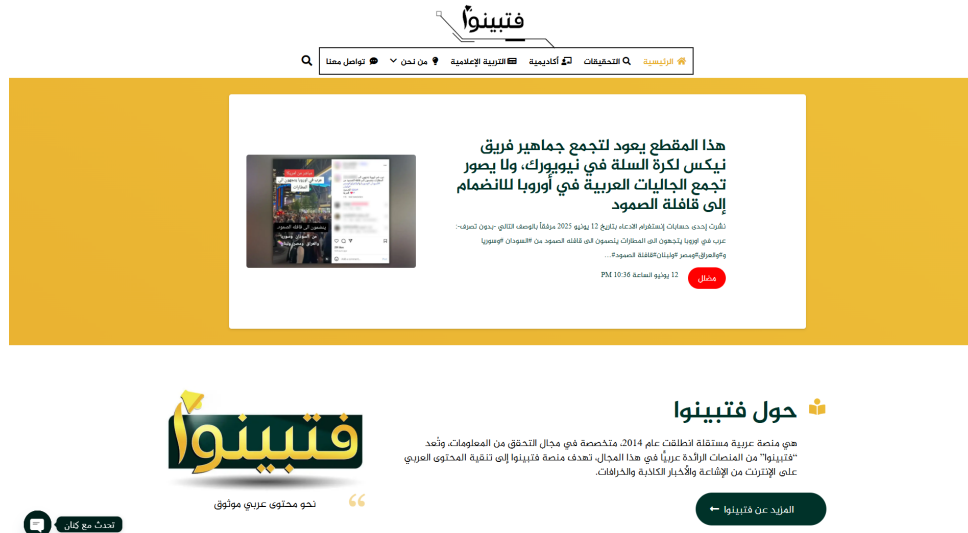


Figure 2.1: Homepage of the Fatabybyano fact-checking platform.

FactCheck.org: is a non-profit U.S.-based fact-checking website that aims to reduce deception and confusion in U.S. politics. It monitors the factual accuracy of claims made by political figures, major interest groups, and media outlets, see figure 2.2

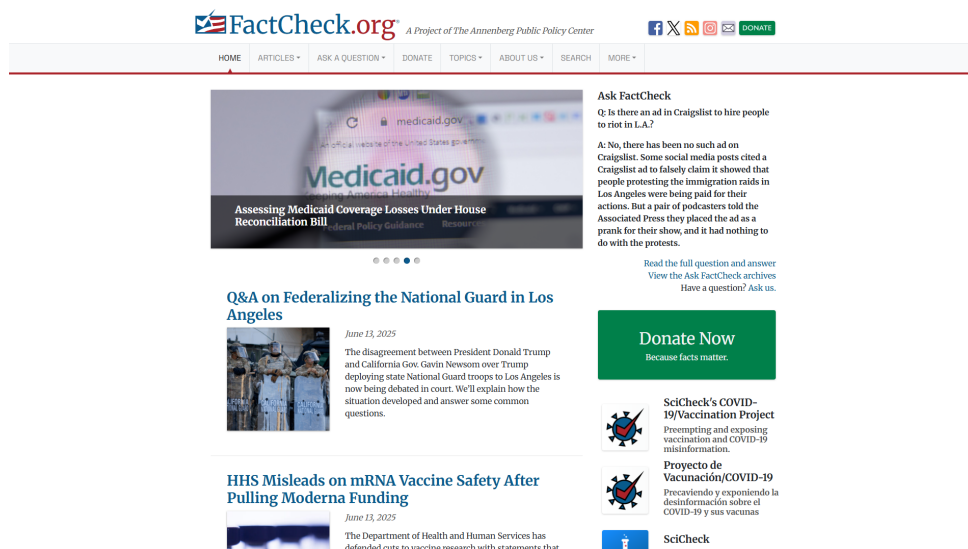


Figure 2.2: Homepage of the U.S.-based fact-checking platform FactCheck.org.

2.4 AI-Based Detection Models

The rise of large-scale online misinformation has led to the development of automated systems capable of detecting rumors in real time. Among the most prominent approaches are AI-based techniques that leverage supervised learning, deep learning, and transformer-based architectures to classify content as true, false, or unverified. While some detection systems operate in a fully automated manner, others incorporate human oversight by using AI models to assist fact-checkers or guide community moderation. This section outlines major AI techniques applied in rumor detection, including their capabilities and limitations.

2.4.1 Support Vector Machines (SVM)

Support Vector Machines (SVM) were among the earliest models used for rumor detection, especially for binary classification of short text. SVMs rely on feature engineering using linguistic, structural, and temporal cues extracted from posts. Features such as word frequency, punctuation use, posting time, and user metadata are often used to train classifiers that distinguish between rumor and non-rumor content [29]. Despite their simplicity and interpretability, SVMs often lack the flexibility to capture nuanced semantic patterns, particularly in informal or multilingual settings.

2.4.2 Recurrent Neural Networks (LSTM/GRU)

Recurrent neural networks (RNNs), particularly Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), have been widely used to model rumor propagation over time. These models can capture temporal dependencies in rumor-related conversations, making them suitable for classifying entire discussion threads or reply sequences [30]. LSTMs have shown success in early detection of rumors by learning patterns in sequential user engagement. However, they still struggle with long-range dependencies and often require extensive labeled data for training.

2.4.3 Transformer-Based Models

Recent advances in NLP have led to the adoption of transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT) for rumor detection

tasks. BERT enables deep contextual understanding by capturing both the left and right context of a word, making it more effective than previous models for detecting nuanced misinformation [31]. Fine-tuning BERT on rumor-labeled datasets allows models to distinguish between factual and misleading content with high accuracy. However, BERT's performance can degrade in low-resource languages or when handling code-switched text, and it lacks transparency in its decision-making process.

2.4.4 Multimodal Detection Approaches

Given that rumors often involve text, images, and videos, multimodal models have emerged as a promising approach. These systems combine textual features (from tweets, posts, or captions) with visual features (from images or video thumbnails) to improve classification accuracy. Models such as those built on the CrisisMMD dataset use convolutional networks alongside text encoders to assess credibility during crisis events [32]. While multimodal systems achieve better robustness, they are computationally expensive and require large annotated datasets across multiple modalities.

2.4.5 Limitations and Challenges

Despite their success, AI-based models face several limitations. Many systems lack explainability, making it difficult for human moderators or experts to validate decisions. Models often exhibit bias toward dominant languages and struggle with informal, multilingual, or sarcastic content. Furthermore, real-time deployment of these systems at scale requires both computational resources and dynamic adaptation to emerging rumor patterns. See Table 2.1 for a comparative summary of these model, typical use cases and limitations.

Table 2.1: Comparison of AI-Based Rumor Detection Models

Model	Input Type	Use Case	Limitation
Logistic Regression	Text	Simple rumor classification	Limited nuance, lacks contextual awareness
SVM	Text	Short-form rumor detection	Poor performance on long or ambiguous inputs
RNN / LSTM	Text sequence	Capturing temporal patterns in rumor spread	Requires large labeled sequence data
BERT / RoBERTa	Contextual text	Context-aware rumor verification	Computationally expensive, needs fine-tuning
Multimodal Fusion	Text + Image	Detecting rumors in combined content (e.g. fake images)	Requires multi-source inputs and complex fusion strategy

2.5 Hybrid and Collaborative Detection Models

As the limitations of both fully manual and fully automated rumor detection approaches have become increasingly apparent, hybrid systems have emerged as a promising middle ground. These systems aim to combine the speed and scalability of AI with the contextual awareness, critical reasoning, and ethical judgment of human participants. Hybrid approaches typically involve collaboration between AI models, fact-checking organizations, platform users, and sometimes external experts.

2.5.1 Concept and Motivation

Hybrid rumor detection systems are designed to:

- Use AI to identify suspicious content at scale.
- Involve users or moderators to confirm, flag, or contextualize claims.

- Enable expert review where necessary for high-impact or ambiguous cases.

This layered approach enables faster detection than manual-only systems while reducing the false positives or interpretability problems of black-box AI systems [7]

2.5.2 Real-World Examples

TikTok Fact-Checking and Content Moderation: TikTok employs a hybrid approach to misinformation detection. Its algorithms flag potentially misleading or harmful content based on engagement patterns and content analysis. These flagged posts are then reviewed by third-party fact-checking organizations and internal moderation teams. Verified false claims may be labeled, removed, or algorithmically suppressed. The platform also notifies users who previously interacted with the misinformation, exemplifying an evolving form of AI-assisted and human-guided rumor moderation.

Facebook Third-Party Fact-Checking: Meta partners with certified fact-checking organizations worldwide (including Fatabyyano in the Arab world). Suspicious posts are flagged by AI and reviewed by human checkers who determine whether they are true, false, or misleading. AI is then used to suppress reach or add warnings.

Google Fact Check Explorer: This tool aggregates fact-checking results from various sources. It relies on human-verified claims but uses AI to help match claims across different contexts and languages.

Microsoft’s NewsGuard Integration: NewsGuard provides credibility ratings for news websites using trained human analysts, and these ratings are integrated into browsers or platforms via APIs, often in combination with AI-based URL filtering.

2.5.3 Strengths and Limitations

The following table illustrates the strengths and limitations of hybrid and collaborative detection models 2.2:

2.6 Community Moderation and Crowdsourcing

In addition to expert fact-checking and automated systems, many platforms have adopted community-based models that rely on user participation to detect, flag, and

Strengths	Limitations
Combines human judgment with algorithmic scale	Dependent on user motivation and expertise
Enhances trust and transparency	May be vulnerable to coordinated manipulation or bias
More robust against complex or nuanced claims	Requires governance systems to resolve disagreements

Table 2.2: Strengths and limitations of hybrid rumor detection systems

moderate rumors and misinformation. These approaches are built on the idea that large, distributed communities can collectively identify false or misleading content faster and more accurately than centralized authorities or algorithms alone.

Community moderation and crowdsourcing differ from hybrid models in that users are the primary agents of detection and judgment, rather than complementing an AI system. While this approach democratizes rumor detection, it also presents unique risks related to bias, groupthink, and credibility.

2.6.1 Key Features of Community Moderation

Community-driven rumor moderation typically involves:

- Flagging and reporting mechanisms for suspicious or harmful content.
- Voting systems (e.g., upvotes/downvotes, likes/dislikes) to promote or demote content.
- Commentary and annotations to provide additional context or evidence.
- Reputation-based roles, where experienced users gain more authority (e.g., moderators).

These systems are often found in large online communities where content volumes exceed what platform staff or moderators can manage manually.

2.6.2 Real-World Examples

Reddit:

Reddit employs a layered moderation structure:

- Community moderators enforce rules at the subreddit level.
- Users vote on content visibility, and reports trigger reviews.
- Rules can differ widely between communities, reflecting local norms.

While not specifically designed for rumor detection, this system has been instrumental in identifying misinformation during major events like elections or pandemics.

Wikipedia:

Wikipedia uses crowdsourced editorial control, where volunteers edit and moderate content in real time. Vandalism and misinformation are often quickly corrected by the community, supported by citation policies and dispute-resolution guidelines. Wikipedia is frequently cited as a successful example of open-source knowledge verification.

Kialo:

Kialo is a debate platform that allows structured discussion of controversial claims. Users propose arguments for or against a claim and vote on relevance and quality. Though not focused on real-time rumor detection, it represents a structured model of crowdsourced credibility assessment.

2.6.3 Benefits and Risks

The following table illustrates the Benefits and Risks of Community-Based Moderation Models [2.3](#) :

2.7 Approaches to Combating Rumors

Informational ecosystems now require multifaceted strategies to counter the spread of rumors. While technical detection methods are crucial, broader social, educational, community, and legal approaches also play a fundamental role. This section outlines four major strategies:

2.7.1 Technological Interventions

These leverage machine learning, natural language processing, and automated systems to detect and flag rumors in real time. They are deployed within platforms (e.g., AI-

Benefits	Risks and Limitations
Scalable moderation with low operational cost	Risk of mob mentality or echo chambers
Community ownership increases platform trust	Potential for misinformation to be <i>popular but wrong</i>
Fast response to emerging claims	Varying quality of user input; lack of domain expertise
Transparency through open discussion and feedback	Vulnerable to coordinated manipulation or ideological bias

Table 2.3: Benefits and Risks of Community-Based Moderation Models

powered filters on Facebook, X) or as independent tools. Such solutions offer speed and scalability but often struggle with context, subtlety, and evolving misinformation tactics.

2.7.2 Education and Media Literacy Campaigns

Promoting critical thinking skills among the public is recognized as a sustainable way to combat misinformation. United Nations Educational, Scientific and Cultural Organization (UNESCO) recommends initiatives like integrating media literacy into school curricula, organizing public workshops, and supporting “information health” approaches that empower users to evaluate and verify content independently .

2.7.3 Community-Driven and Platform-Based Moderation

Human participation remains central to rumor control even in digital environments: Platforms empower users to flag content, contribute context, or vote on veracity. Notable programs include X Community Notes and Facebook’s third-party partnerships (e.g., with Fatabyyano in the Arab world).

These hybrid and crowdsourced methods enhance transparency but require careful governance to prevent manipulation and bias.

2.7.4 Legal and Policy-Based Measures

Many governments now use legal tools to deter or punish the deliberate spread of harmful rumors:

France’s Avia Law[33], Germany’s NetzDG[34], and the EU’s Digital Services Act mandate platforms to remove illegal content and enforce accountability.[35]

In Algeria, Article 196 bis of the Penal Code [36] criminalizes the “dissemination of false news likely to undermine security or public order,” with penalties including up to two years of imprisonment. Further amendments in early 2024 expanded the state’s power to prosecute “false news” that may destabilize institutions[37]. While intended to curb rumor spread, these laws risk infringing on freedom of expression and require transparent enforcement mechanisms.

2.8 Challenges in Rumor Detection

Despite the development of sophisticated models and collaborative systems, rumor detection remains a complex and evolving challenge. Technical limitations, human behavior, ethical constraints, and platform-specific barriers continue to affect the accuracy, speed, and reliability of detection systems. This section outlines the main difficulties that researchers, developers, and policy-makers face in designing and implementing effective rumor detection strategies.

2.8.1 Volume and Velocity of Content

Social media platforms generate millions of posts per minute, especially during breaking news events. Rumors can go viral within minutes, making real-time detection and response essential—yet extremely difficult. Manual methods are too slow, and even the most advanced AI systems may struggle to keep up with the speed and diversity of online discourse [12]).

2.8.2 Linguistic and Cultural Complexity

Rumors often contain sarcasm, slang, regional dialects, or coded language, making them difficult for algorithms to detect accurately. Additionally, multilingual content on

global platforms creates challenges for systems that are trained on English or high-resource languages only. This is particularly true in Arabic dialects, where differences between countries and informal spelling can reduce model performance. This limitation is particularly relevant in the Arabic context, where models such as Transformer-based Model for Arabic Language Understanding (AraBERT) and CAMeLBERT exist but coverage of Algerian Arabic dialect remains extremely limited in both training data and evaluation benchmarks.

2.8.3 Ambiguity and Lack of Ground Truth

Not all rumors can be definitively classified as true or false; many exist in gray areas characterized by partial truths, evolving narratives, or unverifiable content. Furthermore, ground truth labels in datasets are often delayed, inconsistent, or disputed, especially in political or sensitive topics. This affects both model training and evaluation. This limitation is particularly relevant in the Arabic context, where models such as AraBERT and CAMeLBERT exist but coverage of Algerian Arabic dialect remains extremely limited in both training data and evaluation benchmarks.

2.8.4 Adversarial and Evasive Techniques

Rumor spreaders constantly evolve their tactics to bypass detection systems. They may:

- Modify text to avoid keyword filters
- Share content via images or screenshots (invisible to text-based models)
- Use coded language or symbols

These behaviors require detection systems to adapt continually or incorporate multi-modal and semantic analysis techniques [7].

2.8.5 Platform Restrictions and Data Access

Many platforms restrict access to data Application Programming Interfaces APIs, limiting researchers' ability to collect comprehensive datasets. Closed networks like WhatsApp or Telegram present additional challenges due to end-to-end encryption, which pre-

vents third-party moderation. As a result, many detection systems are trained and tested only on public data, reducing their effectiveness in real-world environments [15].

2.8.6 Ethical, Legal, and Social Dilemmas

Efforts to detect and suppress rumors raise serious concerns about:

- Freedom of expression and the potential for censorship
- Bias in models or moderation teams
- Accountability in labeling information as false

Laws criminalizing misinformation (e.g., in Algeria or Singapore) can be controversial if applied unevenly or without transparency. These issues highlight the need for explainable AI, transparent governance, and human oversight in rumor detection systems [11].

2.9 Recent Advances and Research Trends

In response to the growing complexity and societal impact of rumors, the research community has made significant progress in developing more robust, adaptive, and explainable rumor detection systems. This section highlights recent trends and innovations from the last five years, focusing on improvements in AI models, data resources, system transparency, and real-world deployment

2.9.1 Transformer-Based Models and Deep Learning

The introduction of transformer architectures such as BERT, Robustly Optimized BERT Pretraining Approach (RoBERTa), and Cross-lingual RoBERTa (XLM-RoBERTa) has significantly enhanced the performance of rumor detection tasks. These models capture contextual meaning more effectively than earlier approaches and support fine-tuning for specific tasks such as stance detection, claim verification, or multilingual rumor analysis. Researchers have also proposed domain-adaptive pretraining, where models are first trained on general language tasks and then fine-tuned on rumor-related corpora to better capture misinformation patterns.

2.9.2 Graph-Based and Propagation-Aware Models

Recent studies emphasize the importance of propagation context—how rumors spread across users and platforms. Graph-based models such as Graph Neural Networks (GNNs) and Graph Attention Networks (GATs) allow systems to model rumor diffusion patterns, user credibility, and repost behavior [38]. These models excel in scenarios where textual cues are ambiguous but spread patterns are revealing.

2.9.3 Multimodal and Cross-Platform Detection

With misinformation spreading across diverse media formats, multimodal detection has become essential. Researchers now train models to analyze text, images, and videos together to detect manipulated content [39]. Tools like Contrastive Language-Image Pre-training (CLIP) enable semantic matching across modalities, helping systems detect visual memes, deepfakes, or misleading captions. Cross-platform systems also attempt to track rumors as they migrate from one platform to another (e.g., from Reddit to X), helping detect coordinated campaigns and viral misinformation chains.

2.9.4 Early Detection and Weak-Signal Modeling

To stop rumors before they go viral, researchers are building systems that can detect rumors in their earliest stages, sometimes with only a few data points. These models use temporal dynamics, weak social signals, or partial propagation graphs to make early predictions .

2.9.5 Explainability and Human-AI Interaction

A growing focus is placed on explainable rumor detection, where models offer interpretable reasoning to justify classifications. This is important for public trust and for integrating AI with human-in-the-loop systems, especially in journalism or policy domains . Techniques such as attention visualization and counterfactual reasoning are being tested to make outputs more transparent.

2.9.6 Adversarial Robustness and Low-Resource Adaptation

To defend against manipulation, some models are trained with adversarial examples or augmented with contrastive learning to enhance generalization. Other research focuses on low-resource settings, including dialect-rich languages like Arabic or misinformation in local communities, where labeled data is scarce.

2.10 Identified Gaps and Opportunities

Despite the rapid evolution of rumor detection technologies and the development of hybrid systems that combine automation with human oversight, several critical gaps remain in the field. These limitations present opportunities for innovation, particularly in the design of systems that are transparent, inclusive, multilingual, and adaptable across contexts.

2.10.1 Fragmentation of Solutions

Many current systems focus on only one aspect of rumor detection: content, propagation, or user credibility. Few offer a comprehensive solution that integrates multiple dimensions of rumor behavior into a unified framework. This fragmentation limits system effectiveness, especially in real-world environments where rumors are complex, adaptive, and multimodal.

2.10.2 Lack of Integration Between AI, Community, and Experts

Although hybrid models are becoming more common, most still lean heavily toward either algorithmic automation or manual verification. Few systems establish structured collaboration among AI models, platform users (community), and expert validators, despite clear evidence that each group offers distinct and complementary advantages. There is a need for a modular architecture that assigns meaningful roles to each party, allowing for shared responsibility and accountability in rumor evaluation.

2.10.3 Weak Multilingual and Regional Adaptation

Most research and datasets focus on English or high-resource languages, leaving Arabic, African, and regional dialects underrepresented. Existing models often fail to generalize across cultural contexts, leading to decreased performance and relevance in many parts of the world. This gap is particularly concerning in politically sensitive or low-literacy environments where rumors may spread unchecked.

2.10.4 Limited Explainability and Transparency

Many state-of-the-art models (e.g., deep neural networks) operate as black boxes, offering little insight into how decisions are made. This lack of transparency poses problems for trust, especially when false information is flagged or removed. There's growing demand for explainable AI (XAI) in rumor detection—where users, moderators, and policy-makers can understand the rationale behind classification outcomes.

2.10.5 Dataset Bias and Evaluation Gaps

Benchmark datasets are often outdated, imbalanced, or lack metadata about source credibility, temporal context, or cross-platform behavior. They may also introduce labeling bias, where annotations reflect the assumptions of fact-checkers rather than objective ground truth. Future systems must incorporate more diverse, context-rich, and dynamic datasets to achieve reliable performance.

2.10.6 Ethics and Governance

Technical progress has outpaced the development of norms, guidelines, and governance structures for rumor detection. Who gets to decide what counts as a rumor? What rights do users have when their posts are flagged or suppressed? These unresolved questions present both challenges and opportunities for designing systems that respect user rights while promoting informational integrity.

Strategic Opportunity: Addressing these gaps calls for a collaborative platform that:

- Integrates AI detection with community participation and expert validation

- Supports underrepresented languages and contexts
- Offers explainable outputs and transparent workflows
- Enables dynamic moderation without sacrificing civil liberties

The system proposed in this thesis would represent a significant advancement toward scalable, accountable, and human-centered rumor detection. To further enhance user experience and minimize redundancy, the platform will also incorporate a specialized search engine that helps users efficiently locate relevant rumors.

2.11 Conclusion

This chapter has provided a structured and comprehensive overview of current methods and systems for rumor detection, highlighting the evolution from manual, expert-driven approaches to advanced AI-based and hybrid detection frameworks. We began by defining the core concepts of rumor detection, distinguishing it from adjacent domains such as misinformation and disinformation analysis. Manual methods, despite their high accuracy and contextual awareness, were shown to suffer from issues of scalability and speed, which gave rise to the increasing role of automated approaches.

A range of technical models, including support vector machines and deep learning architectures such as LSTM and BERT, were examined in detail, along with multimodal systems that integrate text, image, and metadata inputs. However, these models face notable limitations in explainability, bias, and adaptability to diverse languages or cultural contexts. This gap has led to a growing emphasis on hybrid systems that combine AI capabilities with human judgment, both through professional fact-checking and community-based moderation.

Despite notable progress, several challenges persist. These include the linguistic and cultural complexity of rumors, the difficulty of establishing ground truth, and the evasive tactics used by rumor spreaders. Recent research efforts have focused on improving model robustness, multilingual support, explainability, and early detection capabilities. However, gaps remain, particularly in integrating AI, expert input, and user participation into a unified, transparent, and scalable framework.

In response to these challenges, the next chapter introduces a hybrid platform that

integrates machine learning with structured human oversight. Particular emphasis is placed on detecting misleading content in news articles and headlines, which has become a prevalent and influential form of rumor propagation. This integrated approach aims to support real-time analysis while maintaining transparency, cultural relevance, and user trust.

Conceptual Framework

3.1 Introduction

As was previously alluded to, the primary aim in realizing this project is the seamless blending of the distinct ways fake news and misinformation detection are tackled. We aim to provide quality assessments, commentary, and verdicts on the different and varying pieces of information available through our platform, which will make use of complex and robust systems that metricize the aforementioned criteria. In this chapter, we will explore these systems in detail and the theoretical foundations upon which they rest. We will also discuss the overall structure and behavior that amalgamate these sub-systems into one final coherent system.

3.2 General System Overview

3.2.1 User-System Interaction Flow

User-System Flowchart the following flowchart summarizes general user-system interaction.

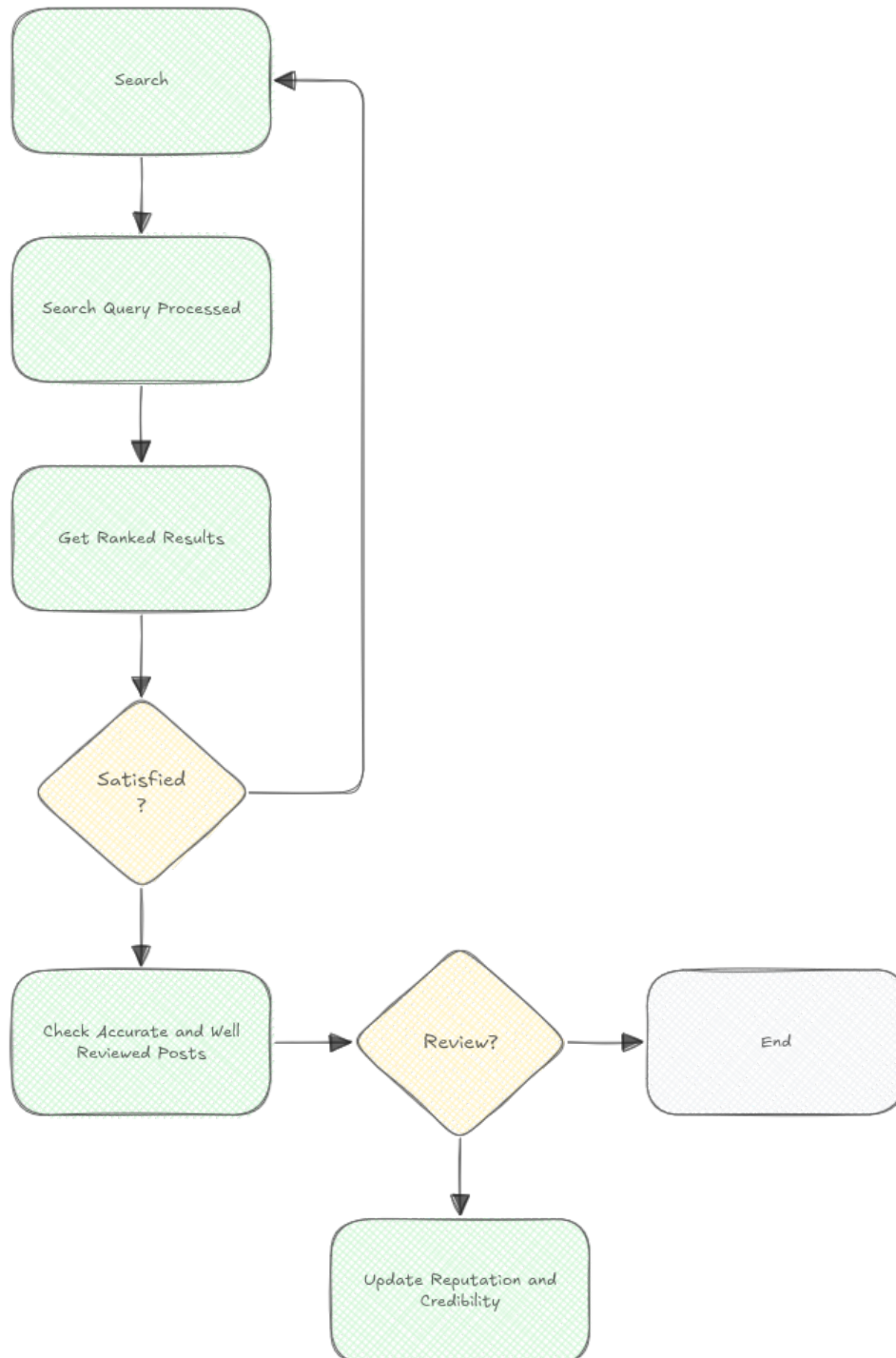


Figure 3.1: General flowchart of user and system interaction.

3.2.2 Use Case Diagram

System Use Cases The following Figure 3.2 provides a use case diagram describing a high-level view of the interactions between the different actors and the listed system's core functionalities. The diagram highlights the collaborative nature of the platform, where users contribute and consume information while moderators and experts provide oversight and validation. Together, these interactions support the platform's objective of assessing information reliability and facilitating the detection of misinformation through both automated analysis and human review.

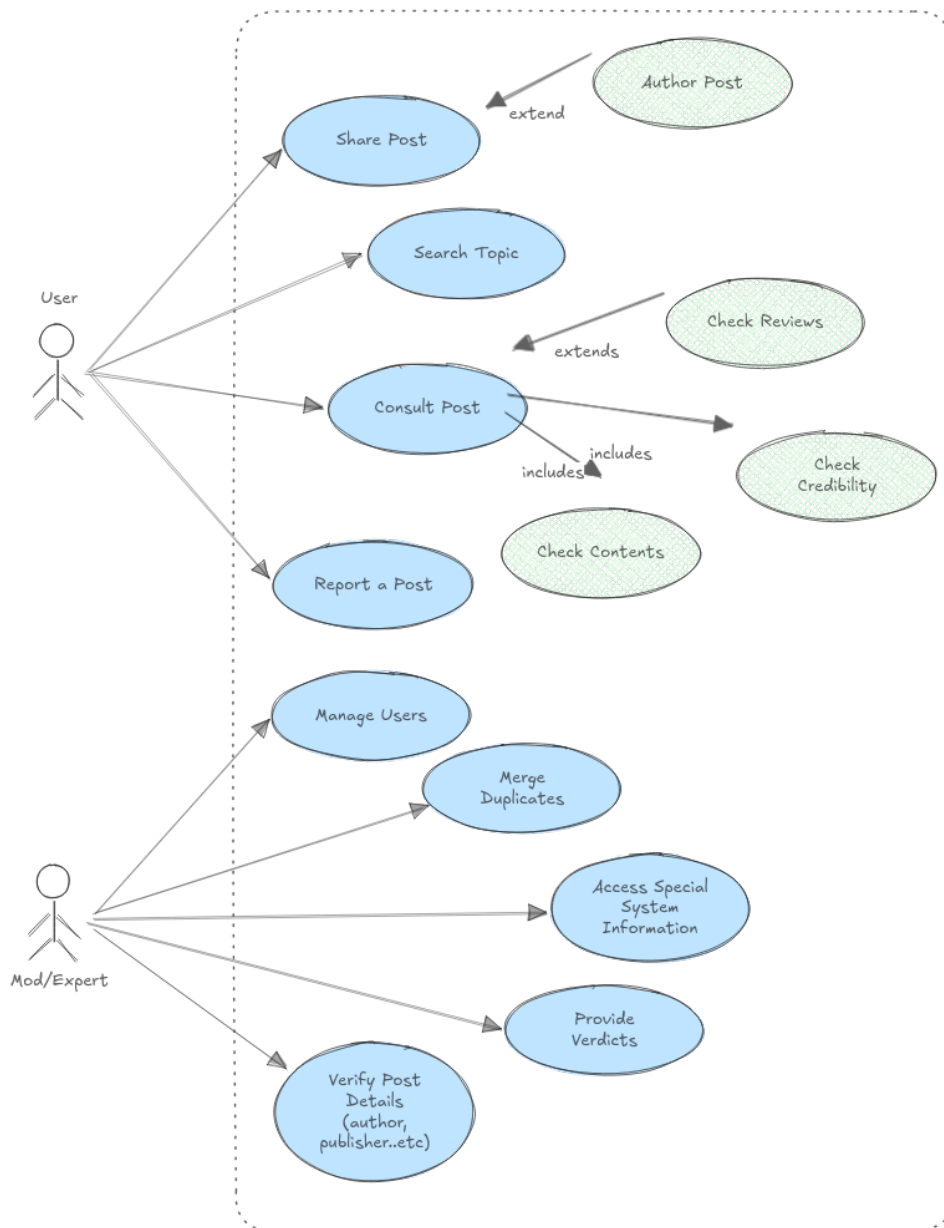


Figure 3.2: General use case diagram.

As can be observed in figure 3.2 , the system doesn't pose much complexity from a use case point of view, it offers, as will be seen later, instead complexity from a behavior and structure standpoints, for which we'll be trying to unravel.

3.2.3 Global Functional Workflow

The proposed platform follows a multi-stage verification workflow that combines automated analysis, community participation, and expert validation. The process begins when a user submits a rumor or query to the platform. An initial assessment is then generated using AI-based analysis techniques, providing a preliminary credibility estimation. The rumor is subsequently exposed to the community, where users can contribute ratings, discussions, and supporting evidence. Finally, domain experts review the available information and issue an authoritative verdict. This layered approach aims to leverage the strengths of different verification mechanisms while reducing the limitations associated with relying on a single source of assessment. Figure 3.3 illustrates the overall workflow of the proposed system.

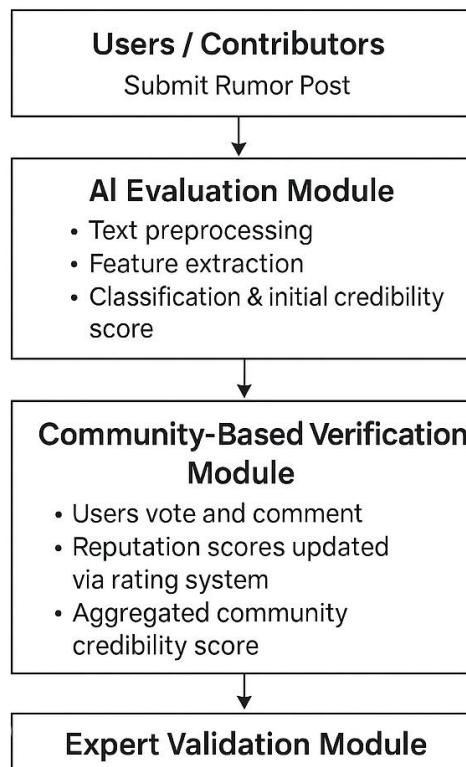


Figure 3.3: Global functional workflow.

3.3 General System Design and Overview

We'll assume having at our disposal a varying set of services, each contribute to respond to the use cases presented :

- NLP Services
- Ranking System Services
 - Rating System Service
 - Assisting Prediction Models Services
- Search Services
- Core Services

This is a high level dissection, following diagram 3.4 somewhat illustrates the relation and flow between these services.

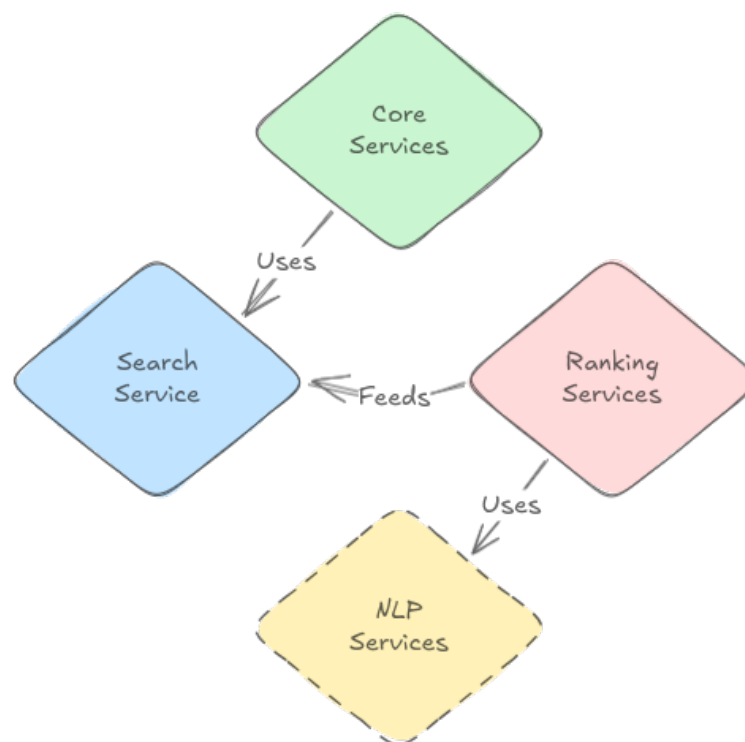


Figure 3.4: Service interaction.

Most suitably, this system should be modeled as a distributed one, as some parts and services would scale differently both in terms of production, and actual load. Figure

3.5 illustrates a high level system design representing a semi-monolith approach, further architectural overviews will be included later, when all parts and modules of the system are sufficiently gone through.

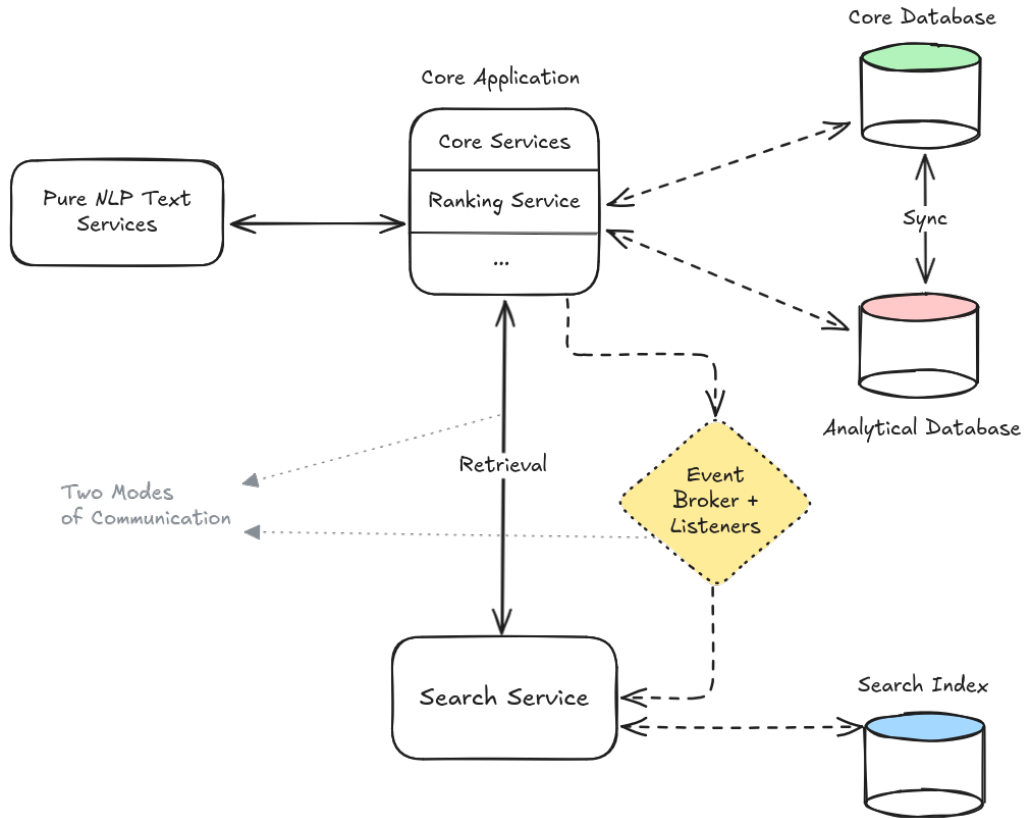


Figure 3.5: High-level application architecture.

At the core of this project are systems responsible for trueness and accuracy estimation, the services under the broad "Ranking" term are consolidated to achieve this, those systems in question are the global rating system and the assisting AI model this would later (with supplementary metrics discussed in later section) be fed to the search service as figure 3.6 shows, helping the user get the most accurate and most helpful results.

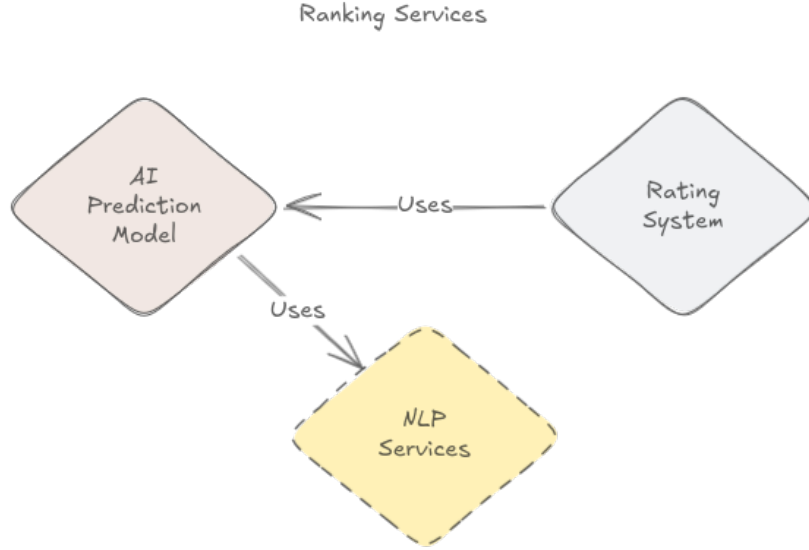


Figure 3.6: Ranking-involved services.

But before detailing the functioning of the assisting AI prediction model and the global rating system specifically, it is important to first introduce a generalized look of the community-based and expert validation modules. These components intervene after the AI stage in the actual workflow, they are presented here first to simplify the explanation and provide a clearer understanding of the overall verification process.

The community-based module highlights how collective reasoning and user participation contribute to evaluating and contextualizing assisting AI results, while the expert validation module demonstrates how final decisions are established through professional review and arbitration. This explanatory order helps contextualize the technical mechanisms of the AI and rating system, ensuring that their roles are understood within the complete hybrid architecture.

3.4 Community-Based Verification Module

The community-based verification module serves as the participatory core of the hybrid system. It transforms the verification process into a collective task, allowing users to contribute contextual knowledge, evaluate claims, and refine AI outputs through interaction. This layer bridges automated reasoning with social understanding, ensuring that verification remains inclusive, transparent, and adaptive to evolving information patterns.

3.4.1 Objectives and Role

The community-based verification module represents the participatory layer of the hybrid system, it's at the core of this project, assisted by AI prediction.

Its primary goal is to leverage the collective intelligence of users to evaluate and contextualize rumors detected by the AI module.

While automated models provide rapid large-scale detection, human users contribute social, cultural, and contextual insights that algorithms often miss—especially in multi-lingual or sensitive environments. This module aims to:

- Strengthen transparency and trust by allowing public participation.
- Facilitate rapid validation of AI-generated claims through collective voting and discussion.
- Detect local or culturally specific rumors that may not be covered by global datasets.
- Support the expert module by providing preliminary crowd consensus and evidence.

Comparable systems such as X Community Notes and Reddit's community moderation demonstrate that large groups of engaged users can collectively produce accurate judgments on information credibility when properly guided by reputation mechanisms and governance rules.

As previously shown in Figure 3.3, the community verification stage represents the second layer of the proposed verification process. This section focuses on its internal mechanisms and functionalities.

3.4.2 Architecture and Integration

This module is tightly integrated into the platform's workflow:

1. A user submits a new rumor or claim through the platform's interface.
2. The AI module automatically analyzes the content, assigns a preliminary credibility score, and classifies it according to predefined categories (e.g., politics, health, technology). Once evaluated, the rumor is published on the community interface along with the AI-generated score and relevant metadata.

3. Users interact with the post by voting on its truthfulness, commenting, or attaching supporting evidence such as references, screenshots, or URLs, affecting both the post's credibility and user's reputation scores, as managed by the platform's global rating system.
4. Experts then review the same post, considering both the consolidated AI output and the community feedback (votes and comments), and make a decision based on this evidence to validate or arbitrate the final verdict.

Conceptually, this layer acts as a social verification buffer between the AI's automated analysis and the expert's final decision. It transforms rumor evaluation into a transparent and participatory process.

3.4.3 User Roles and Permissions

The community-based module operates through a flexible role structure where each participant's influence depends on their credibility score, as managed by the platform's global rating system. Rather than assigning permanent user levels, the system continuously evaluates behavior and performance to determine the weight of each contribution.

Regular Users: Every registered user can view rumors, vote on their credibility, and provide supporting comments or sources. The impact of their actions depends on their personal score, which evolves with each verified or unverified contribution. Reliable users gradually gain higher credibility, giving their inputs greater influence, while inaccurate or biased actions reduce their weight in future evaluations.

Moderators: Moderators oversee discussions, manage abuse reports, and ensure the respect of community guidelines. They do not alter rumor evaluations directly but maintain the integrity of interactions and enforce participation rules.

Experts: The expert role is handled separately and will be detailed in the following section

This adaptive structure ensures that credibility is earned through continuous participation and accuracy, creating a self-regulating environment where every user's influence reflects their historical reliability rather than a static role.

3.4.4 Credibility and Reputation Mechanism

The credibility and reputation mechanism defines how reliability is quantified and propagated across the platform. Every component—users, moderators, experts, AI models, and rumor topics—is evaluated through a unified rating system that continuously updates credibility scores based on performance and interactions. Each user’s credibility score represents their historical accuracy and engagement quality. When a user votes on a rumor or adds feedback, the system has a way to later compare their action to the verified outcome (AI or expert verdict). Accurate or well-supported contributions increase credibility, while repeated incorrect or biased inputs decrease it. This evolving score determines the weight of each user’s future actions, allowing the platform to self-regulate participation quality over time. Likewise, rumor topics accumulate a trust index derived from the convergence or divergence of AI predictions, community feedback, and expert validation. By using a unified scoring logic across all entities, the system ensures fairness and transparency. Credibility is no longer a fixed attribute but a dynamic metric, allowing the community to evolve organically while maintaining accountability and resilience against manipulation.

3.4.5 Interface and Interaction Design

The interface of the community-based module is designed to facilitate transparent and efficient collaboration between users, moderators, and experts. It provides a unified space where all participants can access the same rumor information, system analysis results, and community interactions, while features and permissions adapt to each user’s role.

Each rumor topic appears as a post containing the initial claim, metadata (source, publication date, category, and system credibility score), and an interactive panel for community actions. Users can vote on the rumor’s truthfulness, add contextual comments, and attach evidence such as articles or media links. Their interactions are immediately recorded and reflected in the platform’s scoring system, updating personal and topic-level credibility metrics in real time.

The interface visually highlights aggregated community results and system feedback. Users can view system and expert evaluations and see the level of consensus among participants.

Experts access the same interface but with additional tools to review AI outputs,

ratings system output, analyze user discussions, and issue final verdicts. This shared environment reinforces the transparency of the process while clearly distinguishing decision authority.

Overall, the design encourages participation, accountability, and clarity, ensuring that every interaction contributes meaningfully to the collective verification process.

3.4.6 Advantages and Limitations

The community-based verification module introduces several advantages that strengthen the hybrid system's credibility and adaptability. By incorporating user participation, the platform benefits from collective intelligence, contextual awareness, and rapid feedback, which complement the precision of AI and the authority of expert validation as shown in table 3.1.

Advantages	Limitations and Risks
Transparency: Every step of the verification process is visible to users, from AI assessment to community discussion and expert verdicts.	Unequal Expertise: The accuracy of community evaluations can vary depending on participants' understanding of specific topics.
Adaptability: The community contributes real-time contextual knowledge, particularly valuable for local or culturally specific rumors.	Overload and Fatigue: Sustained participation in large-scale verification may lead to reduced attention or lower-quality feedback.
Self-regulation: Through the rating system, the platform automatically rewards accurate participation and reduces the influence of unreliable or malicious behavior.	Dependence on Reputation Calibration: If the scoring model is poorly tuned, the system might misjudge credibility levels or slow adaptation to new behaviors.
Scalability: Large volumes of content can be reviewed efficiently by distributing verification tasks among many users.	Bias and Manipulation: Coordinated or ideologically motivated groups may attempt to influence votes or discussions.
Engagement and Education: Public involvement increases awareness of misinformation mechanisms and promotes critical thinking among users.	

Table 3.1: Advantages and limitations of the hybrid rumor detection platform.

Despite these challenges, the integration of community participation—assisted by the AI prediction and balanced by the expert modules—creates a resilient verification ecosystem. It combines automation, collaboration, and professional oversight, ensuring that misinformation detection remains accurate, explainable, and socially grounded.

While the community-based module enhances transparency and leverages collective intelligence to evaluate rumors, it remains an intermediary stage in the overall verification cycle.

To ensure that every decision reaches a definitive and authoritative conclusion, the process continues with the expert validation module, where professional reviewers analyze the outputs of both the AI and community layers to establish the final verdict.

3.5 Expert Validation Module

The expert validation module represents the final point of convergence for all verification stages. It integrates the analytical precision of AI and the collective reasoning of the community into a structured, expert-driven decision process. Through professional oversight and methodological rigor, this component ensures that the final verdict of each rumor is both verifiable and accountable.

3.5.1 Objectives and Responsibilities

The expert validation module constitutes the final and authoritative layer of the hybrid rumor detection system. Its primary role is to ensure the accuracy, reliability, and accountability of the verification process by providing professional evaluation after the automated and community stages. This module serves several key objectives:

Final Decision Authority: The expert module delivers the conclusive judgment on each rumor. Experts review all information generated by previous stages—including the AI evaluation, aggregated community feedback, and contextual evidence—and determine the final status of the rumor as True, False, Partially True, or Unverified.

Quality Assurance: Experts act as a quality control layer that corrects possible errors from the AI model or community assessments. Their intervention guarantees that sensitive or ambiguous content is verified using validated reasoning and domain-specific knowledge.

Traceability and Transparency: Each expert decision is recorded, timestamped, and stored in the system database. This ensures full traceability of the verification process and allows the platform to explain or justify verdicts when necessary.

Knowledge Contribution: Expert verdicts form the basis of a growing knowledge base that is reused for AI retraining, system calibration, and community guidance. This continual feedback loop strengthens the accuracy and credibility of future detections.

Through these responsibilities, the expert module consolidates the results of the hybrid

model, ensuring that every rumor receives a professional and verifiable conclusion. It transforms the system from a participatory framework into a trusted decision-support environment, where human expertise validates both algorithmic and collective intelligence.

3.5.2 Workflow and Integration

The expert validation module operates as the final stage in the rumor verification pipeline. It is directly connected to the output of the community-based module and depends on the data generated during both the AI analysis and community interaction phases. This ensures that the expert’s assessment is informed by a complete view of the rumor’s life cycle. The verification workflow proceeds as follows:

Rumor Data Collection: Once a rumor has completed the community review phase, all related data—such as the AI-assisted system generated credibility score, comments, and reputation-weighted feedback—are automatically compiled into a structured report.

Expert Review Interface: The expert accesses the same rumor post as the community but with extended analytical tools. The interface displays AI predictions, aggregated community score through the global rating system, other statistics, and contextual metadata. This allows the expert to evaluate both quantitative signals (scores, agreement rates) and qualitative evidence (comments, sources, user reasoning).

Decision-Making Process: After reviewing all available information, the expert issues the final verdict, choosing one of the predefined outcomes: True, False, Partially True, or Unverified. The expert may also attach a short explanatory note highlighting the reasoning or referencing supporting material.

Validation and Storage: The verdict and its justification are validated by the system and permanently stored in the database. This decision becomes the official record of the rumor’s status and is reflected in the user interface for public transparency.

Feedback Integration: Once the verdict is recorded, the system updates all dependent modules:

- The AI module uses expert-labeled rumors as new ground-truth data for retraining and calibration.
- The rating system further adjusts user scores, outside of its usual cycle, this time according to the proximity of their evaluations to the expert’s final decision.

- The community interface updates the rumor’s state, clearly displaying the verified outcome.

This workflow ensures a closed and verifiable feedback loop, where the expert acts as the final authority while simultaneously reinforcing the learning capabilities of the AI and the self-regulation of the community. The integration of the expert module follows the same overall workflow illustrated previously in Figure 3.3. After the AI and community-based verification stages, the expert receives the aggregated outputs and performs the final validation step, closing the rumor evaluation loop.

This ensures coherence and continuity across all layers of the hybrid architecture.

3.5.3 Expert Dashboard and Tools

The expert dashboard serves as the main working interface of the validation module. It consolidates all rumor-related information into a single, structured environment that facilitates informed and traceable decision-making. Unlike the community interface, which focuses on participation and discussion, the expert dashboard is designed for analytical evaluation, decision accuracy, and evidence management.

Information Overview Panel

This panel displays the essential metadata of each rumor post, including:

- The original text or media of the rumor.
- Source information and publication date.
- The AI-generated credibility score and classification category.
- The community aggregated credibility score through user votes generated by the global rating system.
- Aggregated community results, including vote distribution and agreement rates.
- The reputation-weighted average of user credibility scores related to that post.

This overview provides the expert with an instant snapshot of both algorithmic and collective evaluations.

Analytical and Decision Tools

The expert interface integrates a set of analytical features designed to support objective and consistent judgment:

Comment Explorer: allows the expert to browse, sort, and highlight relevant user comments or attached sources.

Comparison View: presents a side-by-side visualization of AI analysis results and community consensus.

Contextual Indicators: display trend data, topic categories, or related previous cases to guide decision consistency.

Verdict Selector: an integrated control that enables the expert to choose one of the final states (True, False, Partially True, or Unverified) and write a short justification.

Validation and Publication Controls

Once a verdict is chosen, the expert can validate the decision using a secure confirmation step. The validated decision is automatically recorded in the database, timestamped, and made visible in the public interface. Each validation triggers a feedback update to both the AI and rating system modules, ensuring coherence across the platform.

Transparency and Traceability Features

To ensure accountability, all expert actions are logged. Each verdict can be reviewed later by system administrators or governance entities for auditing or methodological refinement. Experts may also view their historical decisions and accuracy statistics, allowing continuous improvement and transparency in decision-making.

3.5.4 Feedback Loop to AI and Community

The feedback loop is a fundamental mechanism that connects the expert module to the rest of the hybrid system. It transforms expert decisions into actionable data that improve both the AI's performance and the community's credibility calibration. Through this continuous exchange, the system evolves toward higher precision, fairness, and adaptability over time.

Feedback to the AI Module

Each expert verdict becomes a verified label that enriches the AI's training dataset. These labeled cases are periodically used to retrain or fine-tune the detection models, improving their ability to recognize linguistic patterns, propagation behaviors, and contextual cues associated with true or false information.

This process ensures that the AI progressively aligns with expert judgment, reducing classification errors in future detections. The expert feedback also supports the identification of recurring weaknesses in the model, such as bias toward certain topics or languages, allowing targeted model adjustments and improved generalization.

Feedback to the Rating System and Community

Once the expert decision is recorded, the rating system cycle is triggered once again comparing the final verdict with each user's previous evaluation. Users whose votes align with the expert's final decision gain credibility points. Those with contradictory or unsupported inputs lose credibility proportionally.

This automatic recalibration strengthens the reliability of the community's collective output, as user scores reflect their historical accuracy.

Additionally, the final verdict is published on the rumor's page, allowing users to view the outcome and the expert's short explanation. This transparency reinforces user trust, encourages responsible participation, and promotes learning through feedback.

System-Wide Impact

The feedback mechanism also contributes to the broader ecosystem of the platform by:

- Updating rumor trust indexes across the database.
- Improving the weighting model used in future community voting phases.
- Generating statistics and insights for administrative or research purposes.

Through this multidimensional feedback loop, the platform maintains a self-improving dynamic, where each completed rumor case contributes to a smarter AI, a more accurate community, and a more consistent verification process overall.

3.5.5 Governance and Ethical Considerations

The expert validation module plays a central role in ensuring the platform’s integrity and public trust. Beyond its technical function, it operates within a defined governance and ethical framework designed to guarantee fairness, accountability, and compliance with information ethics. This framework regulates how experts operate, how decisions are stored, and how transparency is maintained throughout the rumor verification process.

Ethical Guidelines for Expert Decisions

Experts are required to follow a strict code of conduct that emphasizes neutrality, transparency, and evidence-based reasoning. Decisions must rely on verifiable information sources and avoid any personal, political, or ideological influence. This ensures that the verification process remains impartial and that verdicts are based on objective analysis rather than subjective opinion.

Accountability and Traceability

All expert actions are logged and auditable. Each verdict is associated with a timestamp, decision history, and digital signature, allowing complete traceability of the reasoning process. This transparency enables future audits, peer reviews, or appeals if inconsistencies arise. It also reinforces public confidence in the system’s integrity by making decision-making processes explainable and verifiable.

Privacy and Data Protection

The system ensures that any data processed during verification—such as user comments, sources, or external links—is handled according to strict privacy and security principles. Sensitive user information is anonymized, and expert access to private data is restricted to what is strictly necessary for validation purposes.

Oversight and Continuous Improvement

A governance committee or supervisory entity may periodically review expert performance and decision consistency. Statistical analyses of expert accuracy, bias indicators, or response times are used to refine procedures and provide targeted feedback or training.

This oversight process contributes to maintaining high verification standards and fostering professional development among experts.

Transparency to the Public

The final expert verdicts are displayed openly on the platform, accompanied by a short justification. This public visibility strengthens user trust and provides educational value by showing how information is evaluated. It also demonstrates the collaborative balance between artificial intelligence, collective human reasoning, and expert knowledge.

By integrating these ethical and governance measures, the expert validation module not only guarantees the accuracy of rumor detection but also establishes a transparent and responsible decision-making framework. It ensures that the system's authority is rooted in professionalism, fairness, and accountability, making it both technically reliable and socially trustworthy.

The expert validation module completes the hybrid rumor detection architecture by providing the final and most authoritative layer of verification. Together with the AI and community-based modules, it forms a closed, self-improving loop where automation, collaboration, and expertise reinforce one another.

This integration ensures that each rumor is processed through successive layers of intelligence—algorithmic, collective, and professional—culminating in a verified and transparent outcome. The design not only guarantees accuracy but also establishes a foundation of trust between the system and its users.

3.6 System Metrics Overview

As can be recalled, one endpoint of the system is providing the user a search functionality that satisfies a number of metrics that form a criteria, those being:

- Trueness estimation being the main functionality
- Quality feedback and commentary assessing the accuracy of queried post
- Confidence and assertiveness of said post, speculative or opinion based posts are given less priority

This theoretical formula expresses 3.1 how a search rank could be evaluated :

$$S = w_c \cdot C + w_q \cdot Q + w_t \cdot T \quad (3.1)$$

C : Estimated certainty score of a post

Q : Estimated quality of feedback on the post

T : Estimated trueness score

S : Aggregate score of the post (denoted as α)

3.6.1 Supplementary Metrics

Certainty Metric

We consider a post's certainty score C to be a deciding factor, having damping effect on a post's search rank when low. ideally we would like to only challenge assertions, as ultimately those would be the most damaging as they can be taken as a matter of fact, as well as being easier to evaluate, take for example these two headline ideas:

1. *"Lionel Messi may find the idea of going back to Barcelona intriguing"*
2. *"Lionel Messi is going back to Barcelona"*

the first less helpful from a visitor's perspective, it is also hard to argue against even if it maybe false, creating an unwanted effect in our system. the second statement is an assertion that could definitely be evaluated, and the visitor has a high chance of getting provided a verdict on such rumor, leading to a satisfactory result. We won't dive into an implementation of assertiveness evaluation system in our project

Quality Metric

Alongside certainty, and the trueness metric, the aggregated commentary quality score of a post is also important. Posts with low trueness score but high quality commentary and feedback would aid the user in his decision and own evaluation.

The assessment of feedback quality therefore constitutes an important component of the proposed platform. The following section presents the criteria and evaluation framework used to estimate the quality of user-generated feedback.

3.7 Feedback Quality Estimation

In most modern social media platforms such as Facebook and Instagram, the evaluation of user-generated content (e.g., comments, feedback, or posts) is primarily based on interaction metrics, including the number of likes, shares, reposts, and replies.

While these indicators reflect popularity and user engagement, they do not necessarily represent the actual quality or reliability of the content. Misleading or poorly reasoned feedback may receive high levels of interaction due to emotional appeal, cognitive bias, or virality dynamics, rather than its informational value.

This limitation is particularly problematic in the context of rumor detection, where user feedback plays a critical role in shaping public perception and collective judgment.

Therefore, there is a need for a more robust and objective evaluation mechanism that assesses feedback based on intrinsic quality criteria, rather than relying solely on external interaction signals.

3.7.1 Quality Criteria for Feedback Evaluation

In this work, the quality of user feedback is evaluated based on three core criteria: text correctness, thoroughness, and sound reasoning with evidence. These criteria are selected to reflect the essential characteristics of meaningful, reliable, and useful feedback in the context of rumor verification.

Text Correctness

Text correctness refers to the linguistic and semantic quality of the feedback. High-quality feedback should be clearly written and easily understandable.

Key components:

- **Grammatical accuracy:** Correct use of syntax, tense, and sentence structure.
- **Spelling and punctuation:** Absence of basic writing errors.
- **Semantic clarity:** The meaning of the text is coherent and unambiguous.

This criterion ensures that the feedback can be reliably interpreted by both users and automated systems.

Thoroughness

Thoroughness measures the depth and completeness of the feedback. Valuable feedback should go beyond short or superficial statements.

Key components:

- **Coverage of the topic:** The feedback addresses relevant aspects of the claim or discussion.
- **Level of detail:** Inclusion of explanations rather than simple assertions.
- **Multiplicity of points:** Presence of one or more meaningful arguments or observations.

This criterion distinguishes between low-effort comments and well-developed contributions.

Sound Reasoning and Evidence

This criterion evaluates the argumentative quality of the feedback, with an emphasis on evidence-based reasoning.

Key components:

- **Logical structure:** A clear relationship between claims and their justifications.
- **Argument validity:** Reasoning that follows a coherent and rational logic.
- **Evidence support:** Use of facts, references, or external sources to support claims.

This dimension is particularly important in rumor detection, where unsupported opinions may contribute to the spread of misinformation.

3.7.2 Multi-Model Evaluation Architecture

To operationalize the previously defined criteria, the proposed system adopts a multi-model architecture inspired by the concept of Multi-Agent Argumentation and Grammar Integrated Critiquer (MAGIC). Within this framework, the evaluation process is decomposed into multiple specialized modules, each responsible for assessing a specific dimension of feedback quality.

Each module independently processes the input feedback and generates a partial score corresponding to its assigned criterion. These partial scores are subsequently aggregated to produce a final quality score, providing a comprehensive and structured evaluation of user feedback as illustrated in Figure 3.13.

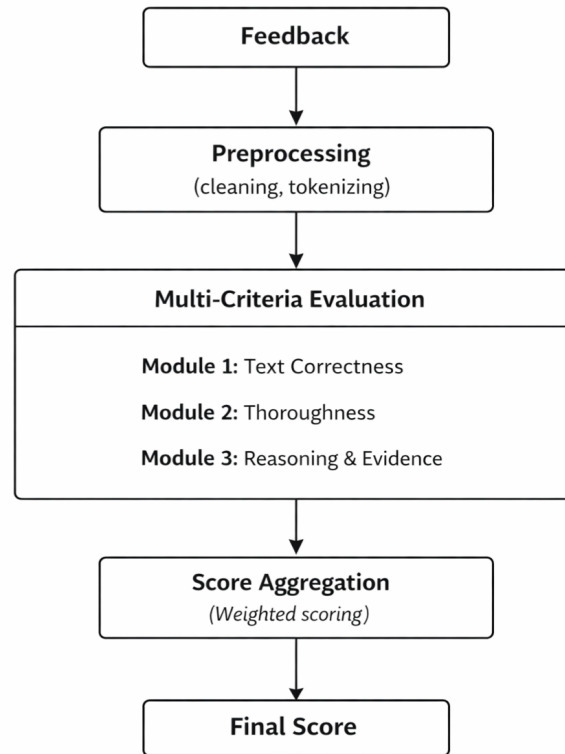


Figure 3.7: Illustration of Multi-model architecture for feedback quality estimation..

Module 1: Text Correctness Evaluator This module is responsible for assessing the linguistic quality of user-provided feedback, focusing on grammatical correctness, sentence structure, and semantic clarity. To perform this evaluation, automated writing assessment tools such as the Grammarly Writing Score API are employed to provide an objective measure of text quality.

To illustrate the behavior of this module, consider the following examples:

Example 1 (Correct text):

This claim is misleading because the data is outdated.

Example 2 (Incorrect text):

This claim misleading because data are outdated.

The first example is grammatically correct and semantically clear, and therefore receives a high correctness score. In contrast, the second example contains grammatical errors, including a missing verb and incorrect article usage, which reduce its clarity and result in a lower score.

Module 2: Thoroughness Evaluator This module measures the depth and completeness of the feedback. It evaluates whether the feedback provides sufficient detail, covers relevant aspects of the topic, and includes meaningful explanations.

The following examples highlight the difference between thorough and non-thorough feedback:

Example 1 (Thorough feedback):

This claim is misleading because the data is outdated and does not reflect recent studies published in 2023.

Example 2 (Non-thorough feedback):

This claim is misleading.

The first example includes both a justification and additional contextual information, indicating a higher level of completeness. The second example, however, is limited to a simple assertion without explanation, resulting in a low thoroughness score.

Module 3: Reasoning and Evidence Evaluator This module evaluates the argumentative quality of the feedback, focusing on logical reasoning and the presence of supporting evidence. It is conceptually aligned with fact-checking approaches, such as those based on the FEVER dataset.

The distinction between strong and weak reasoning is illustrated below:

Example 1 (Strong reasoning with evidence):

This claim is misleading because the data is outdated; according to WHO reports, more recent statistics show a different trend.

Example 2 (Weak reasoning without evidence):

This claim is misleading because I think it is wrong.

The first example demonstrates a structured argument supported by an external reference, which strengthens its credibility. The second example relies solely on subjective opinion without any supporting evidence, resulting in a significantly lower reasoning score.

Score Aggregation Each module produces a normalized score in the range $[0, 1]$. These scores are combined using a weighted aggregation function defined as follows 3.2 :

$$\text{Final Score} = w_1 \times \text{Correctness} + w_2 \times \text{Thoroughness} + w_3 \times \text{Reasoning} \quad (3.2)$$

where w_1 , w_2 , and w_3 represent the relative importance (weights) assigned to each evaluation criterion.

This modular and multi-criteria approach ensures that feedback evaluation is not solely based on linguistic quality, but also considers the depth of content and the strength of reasoning. Consequently, it provides a more reliable and meaningful assessment of feedback quality.

3.8 Trueness Estimation Framework

For our main metric, Evaluating trueness of a post would mean a blend of user ratings, and preconceived beliefs about the publishing sources and the shape of the post itself. To explicate the process in which we estimate trueness, we recall that two factors would be in play, to be exact those two are:

1. AI prediction model (automatic approach)
2. Rating System (collaborative approach)

Our goal is to consolidate these two approaches, allowing each to be weighed dynamically depending on the contextual importance. We propose a hypothetical equation (Equation 3.3) that combines both components to yield a satisfactory result:

$$T = w_1 R_1 + w_2 R_2 \quad (3.3)$$

T : Estimated trueness score

R_1 : Score derived from user ratings

R_2 : Score generated by the ML prediction model

w_1, w_2 : Context-sensitive weights for each source

We will first explain how we intend to compute R_1 and R_2 . Only then can we explore how to construct an appropriate equation that reflects their respective contributions accurately.

3.8.1 AI Prediction Model

As discussed previously, the proposed trueness estimation framework combines information originating from both automated and collaborative sources. The automated component is responsible for generating an initial credibility estimate for a submitted rumor before any community interaction takes place.

The purpose of this estimate is not to produce a final verdict, but rather to provide a prior belief that can later be refined through additional evidence, user feedback, and expert validation. Consequently, the system requires a prediction model capable of analyzing available information and producing a credibility score that serves as an initial assessment of the rumor’s veracity.

The following subsection presents one possible approach for generating this prior estimate and integrating it into the overall trueness estimation framework.

Prior Estimation via Content-Aware Source Credibility Model

To seed this process with an informed starting point, we propose a supplementary prior estimation model. This model ingests the full post text together with a learned, topic-conditioned history of the author and publisher, and outputs a continuous truthfulness score $p \in [0, 1]$. This score is directly used as the Bayesian prior probability of veracity in the refinement loop.

Rationale : A purely neutral prior (e.g., 0.5) ignores the strong signal present in an article’s linguistic content and the publisher’s track record. By initializing the refinement process with a data-driven prior, the main algorithm converges faster and is more robust to isolated, high-engagement falsehoods.

Model Architecture (Prior Generator) The prior model is a merged architecture composed of two complementary components:

- **Text Encoder:** a pretrained transformer model (e.g., RoBERTa) processes the post headline, body text, and publication metadata into a fixed-length representation vector v_{text} .
- **Source Credibility RNN:** a recurrent neural network (single-layer GRU) processes the chronological sequence of a source’s previously verified publications and produces a topic-conditioned credibility representation v_{cred} .

The vectors v_{text} and v_{cred} are concatenated and passed through a feed-forward neural network whose output is transformed by a sigmoid activation function to produce the prior probability

$$p = \sigma(f_{\text{FFN}}([v_{\text{text}}; v_{\text{cred}}])) \quad (3.4)$$

where $p \in [0, 1]$ represents the estimated probability that the post is truthful.

Source Credibility RNN Details

For each source (author or publisher), a rolling window containing the N most recent fact-checked posts is maintained. Each historical post is represented by:

- A topic embedding e_i , obtained from a fixed taxonomy (e.g., a fine-tuned topic classifier or LDA representation).
- An adjudicated truth score $y_i \in [0, 1]$.

The input at time step i is :

$$x_i = [e_i; y_i] \quad (3.5)$$

The sequence $\{x_1, x_2, \dots, x_N\}$ is processed by a GRU, yielding a final hidden state :

$$h_T = \text{GRU}(x_1, x_2, \dots, x_N) \quad (3.6)$$

To condition the credibility estimate on the current post’s topic, the topic embedding of the current post, denoted by e_{current} , is concatenated with the final hidden state:

$$v_{\text{cred}} = [h_T; e_{\text{current}}] \quad (3.7)$$

This design allows the model to emphasize source reliability within the specific topical domain relevant to the current post.

Training : The merged architecture is trained end-to-end using a corpus of fact-checked posts annotated with continuous truthfulness scores, where 0 denotes entirely false content and 1 denotes entirely true content.

The objective function may be formulated either as mean squared error

$$\mathcal{L}_{\text{MSE}} = \frac{1}{M} \sum_{i=1}^M (\hat{p}_i - p_i)^2 \quad (3.8)$$

or as the negative log-likelihood of a Beta-distributed target to encourage probability calibration.

To prevent information leakage, the credibility history associated with each source is constructed exclusively from posts published prior to the target post.

3.8.2 Integration with Ranking System

The proposed prior model provides the initial belief estimate used by the other part of the ranking model. Specifically,

$$P_0(\text{True}) = p \quad (3.9)$$

where p is the output of the prior generator.

Subsequent refinement stages update this belief through Bayesian evidence integration using user feedback, expert assessments, source corroboration, and other external signals.

Temporal Integrity : All training procedures and historical credibility sequences must respect chronological ordering. For any post published at time t , only source publications occurring before t may appear within the historical window. This ensures that the generated probability constitutes a genuine prior available at publication time.

3.8.3 Summary

The proposed prior estimation module complements the main rating algorithm next introduced, by providing an informed Bayesian prior that combines post content and source reliability. The architecture integrates transformer-based semantic understanding with a topic-conditioned credibility model derived from historical source performance.

Because the module is independent of the downstream refinement process, future improvements to the prior estimation architecture can be incorporated without modifying the evidence aggregation and Bayesian update mechanisms. The design therefore provides both practical utility and architectural modularity for future system evolution.

3.8.4 Rating System

There exists a plethora of rating and ranking systems around the internet, such as ones rating products, books, movies etc, some rating influence like the Klout Score [40] system a lot of work was done to give measured way of scoring and ranking objects of which we'll be discussing two.

Similar work

Graph Represented Systems : Of the rating systems there exists, are those that could be modeled as a graph. Say we have a set N of nodes, and L representing a set of links I representing relationships between nodes, influence in the case of the Klout system, citation for the PageRank, given a graph of nodes a number of mathematical models such as Markov's series and so on, could be applied to help provide adequate scoring for the participating objects.

Bipartite Network Represented Systems : This family of systems is what relevant to us, and the ones who'll be most suited to our needs. Said models are ones used to represent rating systems like mentioned before (movies, books, products...etc), many have endeavored to improve the detection of noise and spam, with user reputation based systems coming in different flavors, such as prediction models like Alqwadri et al.'s model [41] , or numerical methods like iterative refinement, the latter's type is the one we'll interest ourselves in. The Bipartite Network would normally have a set of raters R to one

side having a unidirectional relationship like “rate” towards a set of objects O like figure (3) shows.

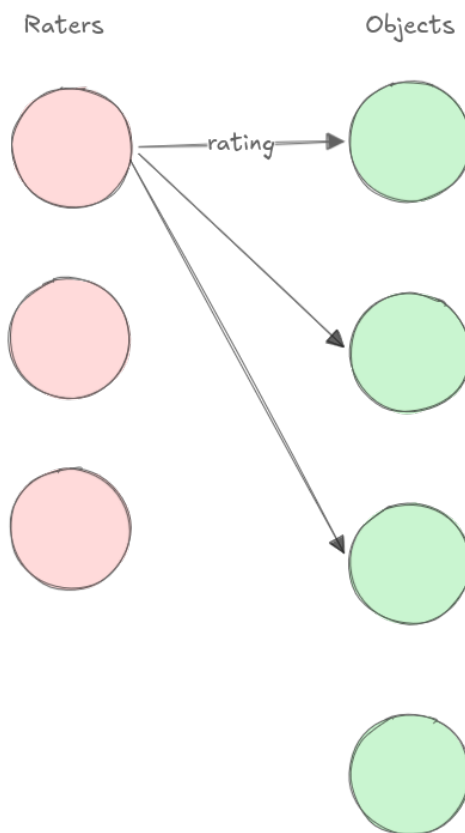


Figure 3.8: Bipartite rating network.

for any of the conventional usage of systems alike, no more variables are taken into account, and we could directly project them onto this representation.

This holds mostly true for our case, except for a few things, a rater may have a special link with each post, whether it be with the source (Bias for or against writer or journal), or the contents themselves (expertise/authority or bias in subject) we’ll look to first familiarize ourselves with existing solutions as a lot could be extrapolated.

A chunk of the proposed solutions aim to weigh user votes using a user reputation system and to estimate an object’s quality based on these votes. Typically, an *iterative approach* is adopted, whereby initial values for reputations from a set R and estimated qualities of rated objects O_r are established. These values are then iteratively adjusted until a convergence point is reached, updating both the reputation scores and object quality estimates simultaneously.

These approaches, which aim to minimize noise and mitigate spam in voting systems,

often follow a principle described as follows:

The distance between the estimated quality of an object and the weighted vote of a rater is used to adjust the rater's reputation accordingly. A large distance Δ negatively impacts the rater's reputation.

Take, for instance, this excerpt reported by Liao et al. [42], which describes a method called *Iterative Refinement (IR)* introduced by Laureti et al. [43], and which forms the foundation of our proposed solution:

“Under this framework, some methods have already been proposed. A representative one is called iterative refinement (IR) method [43]. In IR, a user's reputation is inversely proportional to the difference between his or her rating vector and objects' estimated quality vector (i.e., weighted average rating). The estimated quality of objects and reputation of users are iteratively updated until they become stable.”

This excerpt succinctly summarizes the guiding principle. In our work, we will rely on the methodology proposed in [42] to design a robust ranking and rating system.

Algorithm Used

We'll build upon different methods, to arrive at a tweaked version of iterative algorithm with reputation redistribution (IARR) [42] introducing some additional parameters to fix various deficiencies found in it, we'll finally incorporate our modifications tailoring to own specific needs.

Let us have:

O, U : The sets of objects and users, respectively.

Greek/Latin: Greek letters (e.g., α) are used to denote elements of the object set O , while Latin letters (e.g., i) denote elements of the user set U , for clarity and distinction.

Q_α : The estimated quality (trueness) score of object $\alpha \in O$.

R_i : The estimated reputation score of user $i \in U$.

O_i : The set of objects rated by user i .

U_α : The set of users who rated object α .

$r_{i\alpha}$: The rating given by user i to object α .

N.1 Arithmetic Mean Method

The simplest solution for estimating the quality of an object is to compute the arithmetic mean of the ratings it received. Equation 3.10 expresses this method:

$$Q_\alpha = \frac{1}{|U_\alpha|} \sum_{i \in U_\alpha} r_{i\alpha} \quad (3.10)$$

N.2 Pearson Correlation-Based Algorithm (CP)

We now explore an iterative method based on the CP approach. In this method, the estimated quality of an object is computed as a trust-weighted average of user ratings, and a user's trust level is evaluated using the Pearson correlation coefficient between their ratings and the estimated object qualities.

Step 1: Initial Trust Estimation : We begin by initializing user reputations as:

$$R_i = \frac{|O_i|}{|O|} \quad (3.11)$$

Where:

- $|O|$: Total number of objects.
- $|O_i|$: Number of objects rated by user i .

Step 2: Estimating Object Quality : The estimated quality Q_α of object α is calculated as a weighted average of ratings using user reputations:

$$Q_\alpha = \frac{\sum_{i \in U_\alpha} R_i \cdot r_{i\alpha}}{\sum_{i \in U_\alpha} R_i} \quad (3.12)$$

Step 3: Pearson Correlation Coefficient : The trustworthiness (correlation) score C_i for user i is computed using the Pearson correlation between their rating vector and the current estimated quality vector:

$$C_i = \frac{1}{|O_i|} \sum_{\alpha \in O_i} \left(\frac{r_{i\alpha} - \mu(r_i)}{\sigma(r_i)} \right) \left(\frac{Q_\alpha - \mu(Q_{O_i})}{\sigma(Q_{O_i})} \right) \quad (3.13)$$

Where:

- $\mu(r_i)$: Mean of ratings by user i .
- $\sigma(r_i)$: Standard deviation of ratings by user i .
- $\mu(Q_{O_i})$: Mean estimated quality of objects rated by user i .
- $\sigma(Q_{O_i})$: Standard deviation of those qualities.

The correlation $C_i \in [-1, 1]$, where:

- $C_i = 1$: Perfect positive correlation.
- $C_i = -1$: Perfect negative correlation.
- $C_i = 0$: No correlation.

Step 4: Updating Reputation : User reputation is updated as:

$$R_i = \begin{cases} C_i, & \text{if } C_i \geq 0, \\ 0, & \text{if } C_i < 0. \end{cases} \quad (3.14)$$

Step 5: Convergence Criterion : The iterative process is repeated until the estimated quality stabilizes. Let Q'_α denote the quality from the previous iteration. We compute the convergence metric Δ as:

$$\Delta = \frac{1}{|O|} \sum_{\alpha \in O} (Q_\alpha - Q'_\alpha)^2 \quad (3.15)$$

The iteration stops when $\Delta \leq 10^{-6}$.

N.3 IARR

Some problems could be encountered with the CP method alone which IARR addresses, a low number but a majority of low credible users could have significant say in deciding an object's score, whether by inflating or deflating, despite the credibility of such group being low. Furthermore the credibility of a user whose rating count is low is controlled, ensuring that a rating happening to be close to an object's estimated quality doesn't unduly raise the credibility of the user's with low number of votes, IARR adjustments resulting in Equations 3.16 and 3.17 are used to mend to these issues accordingly

While the CP method introduces trust-weighted aggregation through Pearson correlation, it can still suffer from critical weaknesses. For instance, a small group of low-credibility users may disproportionately influence an object's score — either inflating or deflating it — despite their limited trustworthiness. Additionally, a user with very few ratings might be mistakenly granted high credibility if their few ratings happen to align closely with the estimated qualities.

The IARR method introduces key adjustments to mitigate these issues. Specifically:

- It limits the influence of users with a small number of ratings.
- It amplifies or dampens reputational impact based on rating volume via logarithmic scaling.
- It integrates a maximum reputation normalization factor to prevent domination by low-credibility clusters.

The modified estimation of object quality Q_α under IARR becomes:

$$Q_\alpha = \left(\max_{i \in U_\alpha} R_i \right) \cdot \frac{\sum_{i \in U_\alpha} R_i r_{i\alpha}}{\sum_{i \in U_\alpha} R_i} \quad (3.16)$$

This formulation applies a normalization factor based on the **maximum reputation** among the raters of object α , limiting collective distortion by weak users.

The user credibility C_i is also refined as follows:

$$C_i = \frac{\log |O_i|}{\max_{j \in U} \log |O_j|} \cdot \frac{1}{|O_i|} \sum_{\alpha \in O_i} \left(\frac{r_{i\alpha} - \mu(r_i)}{\sigma(r_i)} \right) \left(\frac{Q_\alpha - \mu(Q_{O_i})}{\sigma(Q_{O_i})} \right) \quad (3.17)$$

Where:

- $\log |O_i|$ acts as a stabilizer, ensuring that users with very few ratings do not get overvalued.
- $\mu(r_i), \sigma(r_i)$ are the mean and standard deviation of ratings by user i .
- $\mu(Q_{O_i}), \sigma(Q_{O_i})$ are the mean and standard deviation of estimated qualities of objects rated by i .

These adjustments ensure a more stable and trustworthy iterative ranking system, minimizing outlier influence and improving convergence reliability.

$$R_i = \begin{cases} C_i^\theta, & \text{if } C_i \geq 0, \\ 0, & \text{if } C_i < 0. \end{cases} \quad (3.18)$$

with θ a given value greater than 1. The process is similar, taking into account the adjusted formulas.

N.4 IARRv2

The paper describes a proposed method called simply *proposed ER*, wherein additional penalty factors are introduced. These are introduced in case a select group of objects monopolize votes from a small pool of certain users, resulting in their weights being inflated and disproportionately larger than those of other users, as well as the monopolizing objects potentially getting a high score.

Furthermore, we detect a glaring incompatibility with our own model, when estimating the quality of objects, the IARR doesn't take into consideration the possibility of low scoring spam, a majority group of less credible users would be able to bring down an object's estimated quality, in fact it is made worse by multiplying with the maximum reputation score, as that would only serve to reduce the object's quality further.

We'll attempt to arrive at a solution. First, a provisional estimated quality $E_{i\alpha}$ is shown in formula 3.19, where the rating of user i is excluded.

$$E_{i\alpha} = \begin{cases} \frac{\sum_{j \in U_\alpha \setminus \{i\}} R_j r_{j\alpha}}{\sum_{j \in U_\alpha \setminus \{i\}} R_j}, & \text{if } |U_\alpha| \geq 2, \\ 0, & \text{if } |U_\alpha| = 1. \end{cases} \quad (3.19)$$

C_i is modified as formula 3.20 illustrates:

$$C_i = \frac{\log |O_i|}{\max_{j \in U} \log |O_j|} \cdot \frac{1}{|T_i|} \sum_{\alpha \in T_i} \left(\frac{r_{i\alpha} - \mu(r_i)}{\sigma(r_i)} \right) \left(\frac{E_{i\alpha} - \mu(E_{T_i})}{\sigma(E_{T_i})} \right) \quad (3.20)$$

with:

$$T_i = \{\alpha \mid \alpha \in O_i \text{ and } |U_\alpha| \geq 2\}$$

We show $\mu(E_{T_i})$ and $\sigma(E_{T_i})$ with the following formulas 3.21 and 3.22:

$$\mu(E_{T_i}) = \frac{1}{|T_i|} \sum_{\alpha \in T_i} E_{i\alpha} \quad (3.21)$$

$$\sigma(E_{T_i}) = \sqrt{\frac{1}{|T_i|} \sum_{\alpha \in T_i} (E_{i\alpha} - \mu(E_{T_i}))^2} \quad (3.22)$$

Another part of the proposed solution, shown in formula 3.23, involves substituting the maximum credibility level with a parameter that takes into account the number of raters for the object:

$$Q_\alpha = \frac{\log |U_\alpha|}{\max_{\beta \in O} \log |U_\beta|} \cdot \frac{\sum_{i \in U_\alpha} R_i r_{i\alpha}}{\sum_{i \in U_\alpha} R_i} \quad (3.23)$$

This modification addresses the issue described in the paper as follows:

“If there is a user with a high credibility level, there is the problem that even where the number of users giving a rating is low, the estimated quality will become high.” [44]

Ultimately, this formulation gives more importance to the number of voters rather than the reputations themselves. While it may be argued whether this yields better results, the matter remains somewhat subjective.

However, with all things considered, this formula still does not account for the possibility of low-rating spam. Therefore, we will pursue an alternative route for estimating the quality of objects.

3.8.5 Our Approach to a Rating System

we recall in 3.8 that estimating the object’s quality, for a general sense, or trueness in our case, requires taking two measures into consideration, an automatic prediction model that will constitute a “prior belief” and a verdict presented from a rating system, we also have unique metrics that need to be accounted for, so now that we’ve exposed ourselves to an existing rating system, we’ll now look to project it onto our application.

Incorporating specific metrics

Recollecting that each node from the user side has a unique relationship with the nodes from the object side in the bipartite network (where objects, in our case, posts), we define these relationships as follows:

- **Relationship with the contents of the post itself:**

- Bias in the subject matter (e.g., political bias, etc.)
- Degree of authority over the subject matter in a post, stemming from user expertise

- **Relationship with the owning sources of the post:**

- Bias for or against the writer and/or the journal

Let A and B be sets of authority and bias scores, where $A \subset [1, 2]$ and $B \subset]0, 1]$. Let $a_{i\alpha} \in A$ represent the authority score of user i over post α , and $b_{i\alpha} \in B$ the bias score of user i toward α .

We can now tune the previous formula for estimated quality to include these metrics in formula 3.24.

$$Q_\alpha = \frac{\sum_{i \in U_\alpha} R_i r_{i\alpha} a_{i\alpha} b_{i\alpha}}{\sum_{i \in U_\alpha} R_i a_{i\alpha} b_{i\alpha}} \quad (3.24)$$

Here, $b_{i\alpha}$ is a bias score and $a_{i\alpha}$ is an authority score for user i with respect to object α .

The bias score $b_{i\alpha}$ is initially calculated using a prediction model and then adjusted algorithmically before calculating Q_α to reflect the nature of the corresponding rating:

- If $b_{i\alpha} = 0.8$ (a negative bias) and the rating $r_{i\alpha}$ is *positive*, then the effective bias becomes $1 - 0.8 = 0.2$, which dampens the influence of the vote.
- If the rating $r_{i\alpha}$ is *negative*, then the bias score remains 1, allowing full effect.

We assume that a graph exists in which the nodes represent *fields of knowledge*, and the edges represent *relatedness* between fields. a post's addressed field is extracted beforehand using an NLP model, while each user's field is known or specified.

The distance between a user’s field and a post’s field (in the graph) is then translated into an authority score, where a distance of 0 corresponds to a maximum authority score of 2.

We will not delve into the technical details or procedure used to extract these values in this paper—we assume they are available.

Estimating trueness – Consolidating the two approaches

Previously, we discussed the lack of accounting for low-rating spam in earlier solutions. We now aim to expand our approach to accommodate the two main components in trueness estimation. To this end, we introduce a method called *Bayesian averaging*, which will help address the remaining challenges.

Bayesian Average This passage from Wikipedia provides a suitable explanation:

“A Bayesian average is a method of estimating the mean of a population using outside information, especially a pre-existing belief.” [45]

The formula for the Bayesian average is given as:

$$\text{Bayesian Average} = \left(\frac{v}{v+m} \right) R + \left(\frac{m}{v+m} \right) C \quad (3.25)$$

In a typical rating system:

- R : average rating
- v : number of votes
- m : minimum number of votes required to be considered reliable
- C : prior belief (external prediction)

This method blends two signals: the observed average rating and a prior-belief value. When the number of voters v is small, the result is biased toward the prior C . As v increases beyond m , the estimate becomes more reflective of the actual average rating R .

Utilizing Bayesian Averaging We have at our disposal an initial trueness score from a prediction model that constitutes a *prior belief*. Using Bayesian averaging, we can improve our estimation of the trueness score, as shown in formula 3.26:

$$Q_\alpha = \left(\frac{\sum_{i \in U_\alpha} R_i}{\sum_{i \in U_\alpha} R_i + \text{Average} \left\{ \sum_{j \in U_\beta} R_j \right\}} \right) R + \left(\frac{\text{Average} \left\{ \sum_{j \in U_\beta} R_j \right\}}{\sum_{i \in U_\alpha} R_i + \text{Average} \left\{ \sum_{j \in U_\beta} R_j \right\}} \right) P \quad (3.26)$$

Here, instead of using the number of votes, we rely on the reputation or trustworthiness scores:

- $\sum R_i$: the total credibility (reputation) of users who rated the current post α ,
- R : the average weighted rating from those users,
- P : the prior-belief trueness score predicted by the model,
- $\text{Average} \{ \sum R_j \}$: the average total reputation values over network objects j , used as a minimum threshold m .

This approach adapts the Bayesian averaging principle to trust-based systems. If there are enough credible voters for a post, the estimated score Q_α is primarily driven by the actual weighted average R . If credibility is low or noisy, the result is pulled toward the prior prediction P , offering a more robust estimation in sparse or biased conditions.

To conclude this section, we've improved and integrated some the rating models previously mentioned, wherein instead of bringing down the score, we make use of additional context in the shape of the prediction model score, acting a reference point in case of lack of sufficient user credibility. We ultimately have integrated and leveraged both automatic and manual approaches.

3.9 Dynamic Reputation Computation Framework

The previous sections presented the models used to estimate rumor trueness and user reputation. However, applying these models in a large and continuously evolving platform introduces challenges related to scalability and computational efficiency. As users, ratings,

and rumors are constantly added, repeatedly recomputing the entire network becomes impractical.

To address this issue, the proposed system adopts a dynamic reputation computation framework based on selective updates and incremental refinement. The following sections describe the theoretical algorithm, compare static and dynamic execution strategies, and present the solutions and business rules that ensure efficient and consistent reputation computation.

3.9.1 Theoretical Algorithm

The preceding sections introduced a theoretical formulation of the proposed algorithm. While conceptually sound, a direct and static implementation of this formulation presents scalability limitations, particularly when applied to large and continuously evolving networks.

3.9.2 Static vs Dynamic Execution Strategies

Two primary approaches can be considered when deploying the algorithm:

Static Approach

In the static approach, the algorithm loop is executed over the entire network whenever an update is required. Given that the network serves as the computational medium, this implies recomputing all relevant user reputations and post accuracy scores globally, even when only a subset of the network has changed. While straightforward, this approach is computationally expensive and does not scale efficiently.

Dynamic Approach

The dynamic approach introduces selective execution. Instead of recomputing the entire network, the algorithm loop is applied only to a relevant subnetwork on demand. This significantly reduces computational overhead and enables more responsive updates in environments with frequent changes.

Dynamic Execution Model

In systems characterized by a continuous influx of new users, ratings, and posts, the dynamic approach is more appropriate. The expected behavior is as follows:

- When a sufficiently large set of new ratings is introduced (as determined by system-defined thresholds), the associated posts and all participating raters are extracted to form a localized network.
- The algorithm loop is executed on this sub-network, updating:
 - User reputation scores
 - Post accuracy scores

Key Assumptions

1. **Non-static Reputation:** A user's reputation score is continuously evolving. Even in the absence of new ratings, changes in the evaluated posts influence the user's reputation.
2. **Approximate Post Accuracy:** Post accuracy scores may temporarily diverge from the true global state. This occurs because updates are only triggered upon sufficient new rating activity. Consequently:
 - Historical ratings are not immediately re-evaluated when user reputations change.
 - Alternatively, a potential design choice may restrict posts to consider only new ratings, ignoring retrospective reputation changes. Simply put, in this case, a user's reputation is taken into account at the moment of evaluation of a post, and any subsequent change in reputation, barring any revaluations, does not affect a post estimated quality. This is not what we will do.

3.9.3 Challenges and Limitations

Execution Frequency and Parallelism

Determining when to trigger the algorithm loop and by extension the aggregation operation is non-trivial. Key considerations include:

- Thresholds for “sufficient” updates
- Avoiding overlapping executions
- Managing concurrency and resource contention

Reputation Estimation Constraints

User reputation is derived from a correlation between:

- The global user rating vector
- The global post accuracy vector

In a partial network:

- Access to global vectors is limited
- Reputation estimates become context-dependent and potentially inconsistent

A naive recomputation from global sets introduces:

- Redundant calculations
- Prohibitive complexity, potentially approaching $O(n^k)$

More over, as global vectors grow in size a convergence point is expected, as a result:

- a user’s reputation risks becoming stagnant at some point.

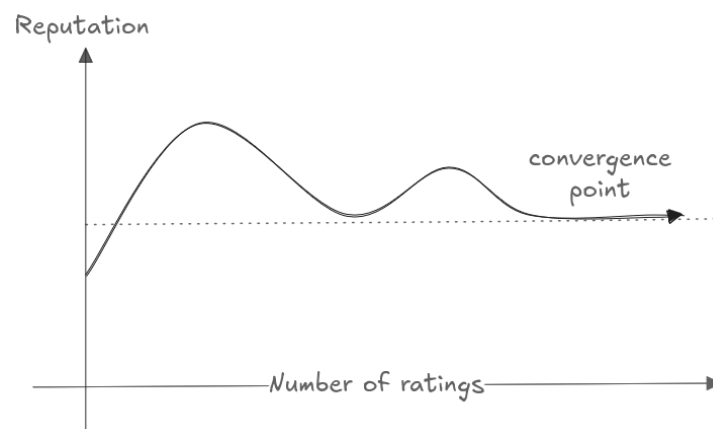


Figure 3.9: User’s reputation by number of ratings.

3.9.4 Proposed Solutions

We will discuss some measures that can be taken to address the potential worries, as either solutions or the enforcement of some rules.

Use of Aggregated Data

To mitigate the lack of global context, we introduce precomputed aggregates that approximate the global network state.

- Aggregates store summarized representations of ratings, posts, and user statistics
- Redundant elements are explicitly labeled and handled accordingly
- Overlap between aggregates and local subnetworks is expected

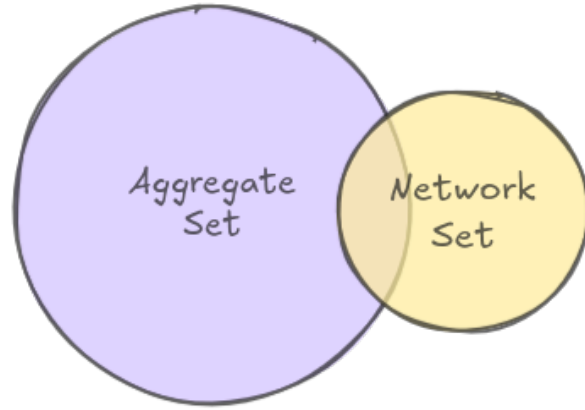


Figure 3.10: Illustration of overlap between sets.

Additionally, we will have to make use of the following adjusted formulas for mean 3.27 and variance 3.28 taking into account overlapping sets:

$$\bar{z} = \frac{n\bar{x} + m\bar{y} - k\bar{w}}{n + m - k} \quad (3.27)$$

$$S_z^2 = \frac{nS_x^2 + mS_y^2 - kS_w^2}{n + m - k} + \frac{n\bar{x}^2 + m\bar{y}^2 - k\bar{w}^2}{n + m - k} - \bar{z}^2 \quad (3.28)$$

- with X, Y are two sets being combined, $W = X \cap Y$ is the overlap and $Z = X \cup Y$ is the combined dataset
- n, m, k being the sample sizes for X, Y and W respectively
- $\bar{x}, \bar{y}, \bar{w}, \bar{z}$ are the means of X, Y, W, Z respectively
- $S_x^2, S_y^2, S_w^2, S_z^2$ are the variances of X, Y, W, Z respectively

Use of Specialized User Vectors

We can limit the effective global user sets to not include older interactions.

3.9.5 Business Rules

the system enforces the following rules

1. Accuracy-related scores are computed based on internal logic and always require a network construct.
2. Each (User, Post) pair is associated with exactly one rating. Historical ratings are not retained.
3. Entities (users, posts, ratings) are labeled if they exist within aggregation sets, these aggregated value of members that have been updated are stored to be used for overlap resolution.
4. Algorithm loop executions are strictly sequential to prevent overlapping runs and race conditions. In addition a bounded number of executions is allowed before forcing aggregation, preventing excessive loss of global context.

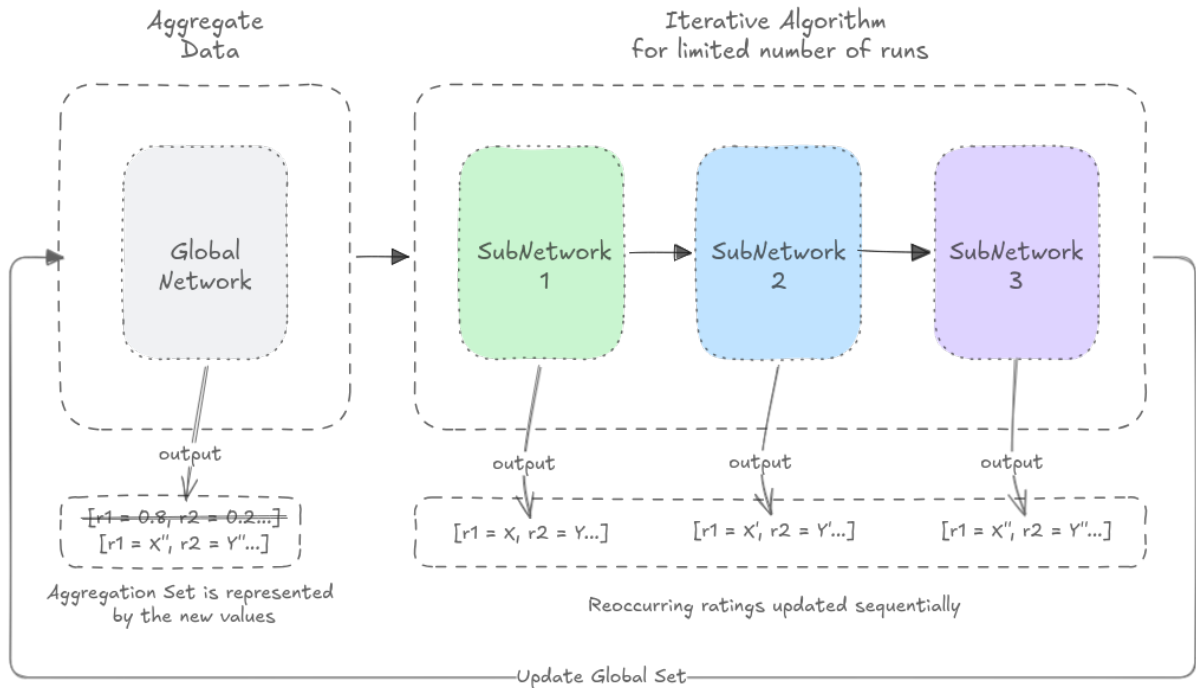


Figure 3.11: Sequential algorithm runs in conjunction with aggregation operations.

5. The iterative algorithm logic is proprietary and encapsulated within the system core.

The dynamic execution model, supported by aggregation mechanisms and strict business rules, provides a scalable alternative to static global computation. While it introduces approximation and consistency trade-offs, these are controlled through architectural constraints and iterative refinement strategies. After detailing some business rules that the system enforces, we are perfectly led to the next section which will describe the system design aspects.

3.10 System Architecture

Thus far, we have detailed the components of the system as well as some business rules to be enacted, we will describe now an architecture and the necessary infrastructure and dependencies that allow these components to exist.

3.10.1 Core Application

The core part of the semi monolith follows a layered architecture to enforce separation of concerns and maintain modularity.

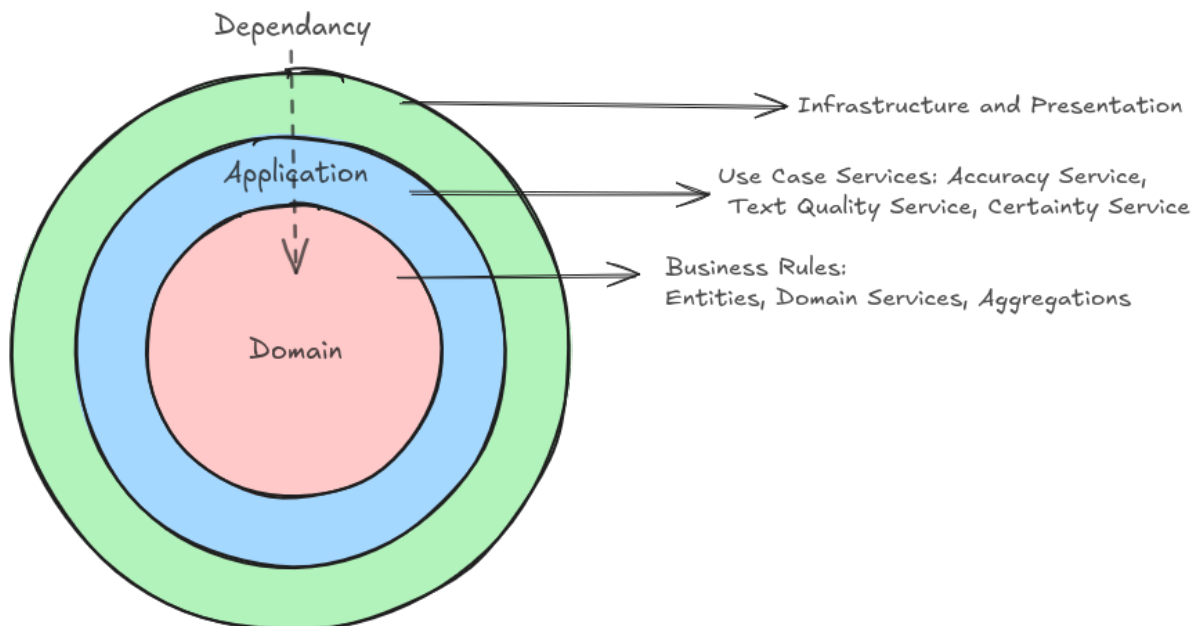


Figure 3.12: Layered architecture for the core application.

3.10.2 Domain Layer

- The domain layer encapsulates the close to unchanging core enterprise-wide business logic and rules
- It is independent of external dependencies

Here is the class diagram:

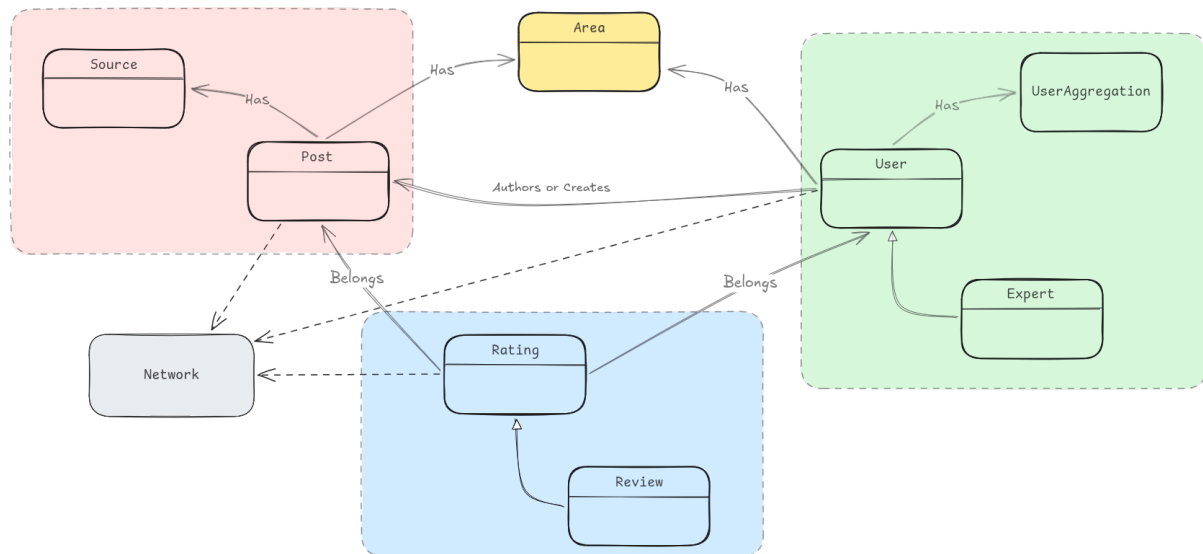


Figure 3.13: Class diagram representing the domain.

Entities

User

- Can create and/or potentially author posts
- Can issue exactly one rating per post
- Has an associated area of expertise
- Reputation is computed externally
- Linked to a user-specific aggregation set

Expert a specialized user, that can verify posts, and provide verdicts on them in his respective area of expertise.

Post

- Belongs to one or more areas of expertise
- Associated with a source (e.g., journal, author)
- Stores accuracy and certainty scores
- Evaluation logic is external

Rating

- Mutable and may belong to aggregation sets
- Can include a review with a quality score
- Maintains an internal correlation metric with the associated post

UserAggregation

- Represents aggregated ratings and posts for a user, a summary of a user's global post and rating vectors
- Stores statistical measures (e.g., mean, standard deviation)
- Handles overlap resolution logic

Network

- Defines the execution scope of an algorithm loop
- As it is the only medium through which the algorithm is run, it encapsulates significant portions of business logic

Domain Services

HandleNewRatingsService This is the service containing the iterative algorithm logic, it is a pure domain service meaning it does not interact with data manipulation, and outer layers as a whole.

- It runs on an extended network construct containing lost context from previous runs and newly added data. It provides updates values
- Requires a global summary if a user's interactions in the form of userAggregation

Application Layer

The application layer orchestrates domain logic to fulfill specific use cases.

- Coordinates services and workflows
- Acts as an intermediary between domain and infrastructure

A snapshot of how some these services are structured in relation the domain layer is given in Figure:

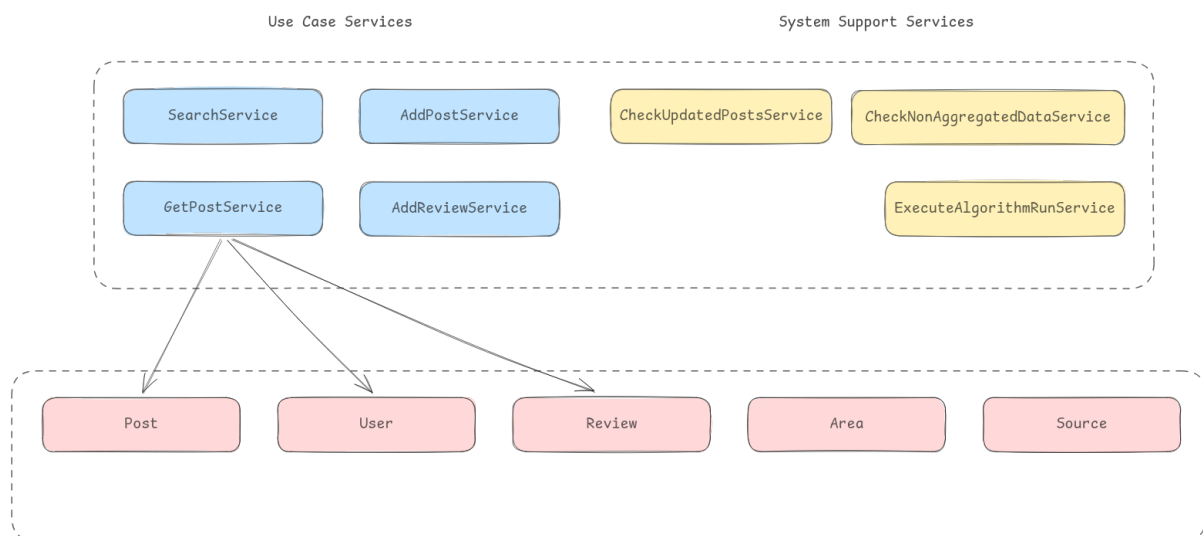


Figure 3.14: Application layer services.

Use Case Services These are the services that map out to uses cases,

Support Services These services are as the name implies supporting services, either to a use case service or for system-wide support functionality

Infrastructure and Presentation Layer

Infrastructure Handles external concerns:

- Data storage and retrieval
- Integration with external APIs, in our case with other parts of the distributed system

The presentation layer is responsible for exposing system functionality to users via interfaces

3.11 Global System Design

The global system architecture is shown in Figure 3.15. The design follows a service-

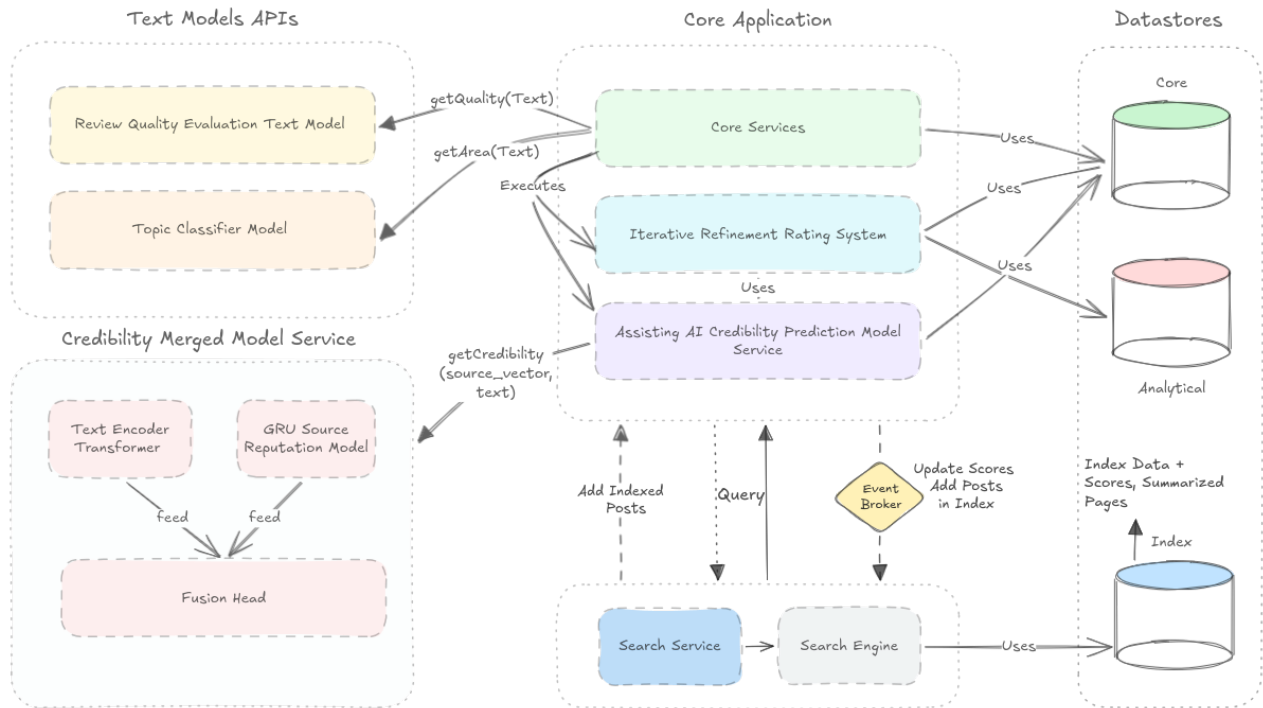


Figure 3.15: Detailed global system design.

oriented approach and a semi distributed design in which specialized modules collaborate to evaluate information credibility, assess review quality, and provide reliability estimates. The core application orchestrates these components, leveraging both machine learning models and supporting services to generate and refine assessments. In parallel, a dedicated search subsystem manages indexing, retrieval and ranking operations, while separate datastores support transactional and analytical workloads.

3.11.1 Data Stores

The core application makes use of two distinct types of data, analytical and the core data which is in part transactional.

Core Database Carries entity tables as well as other types data like cache and taxonomy (topic labeling), use of it is not specific all services interrogate it.

Analytical Database Mainly used by the iterative algorithm services, contains summaries of user and post related data vectors, as well as some general global stats as shown in Figure 3.20.

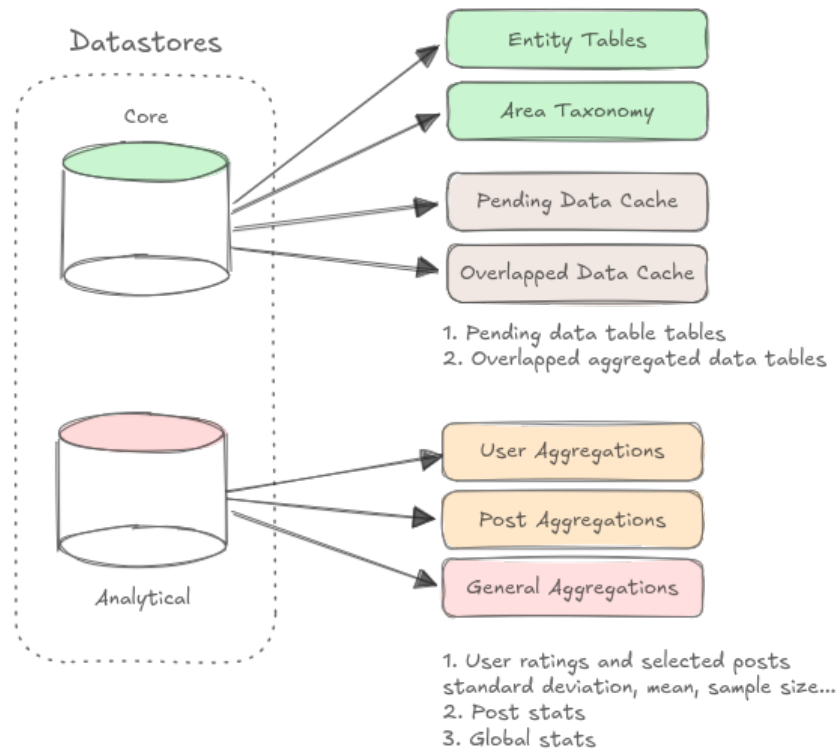


Figure 3.16: Data-stores of the system.

Data Cache The data acts as a sort of an extension to the main tables, it facilitates the work of some processes, and allows some to work as intended. s

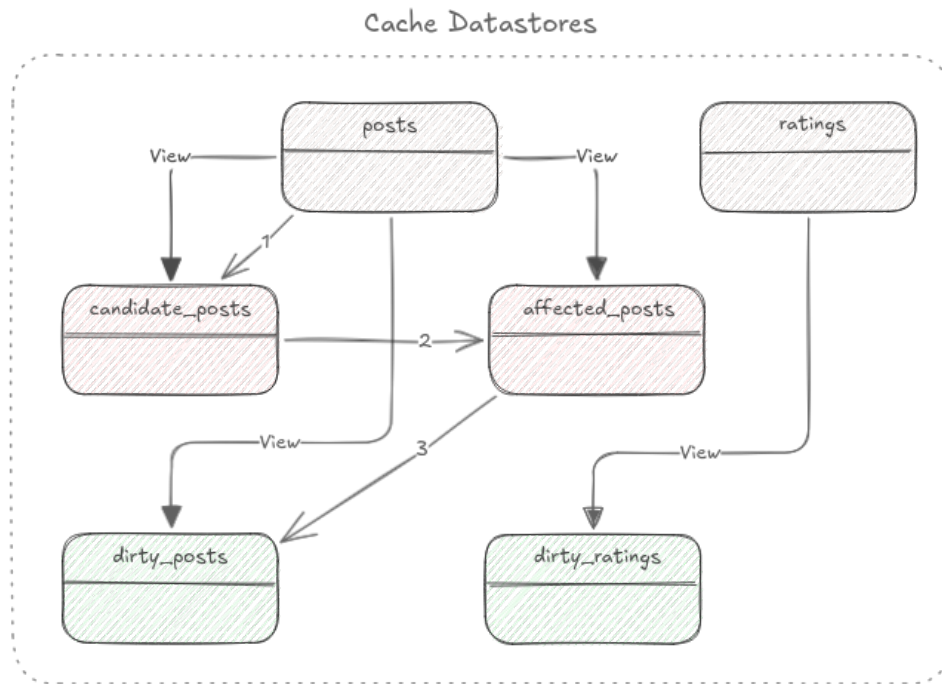


Figure 3.17: Cache structure.

3.11.2 Search Service

Outside of the traditional search engine responsibilities such as crawling, indexing, and document retrieval, the Search Service introduces additional application-specific capabilities. Acting as a bridge between the core application and the search engine, it handles query orchestration, result enrichment with credibility and certainty metrics, index synchronization, de-duplication support, ranking and filtering logic, and inter-service communication. This abstraction layer allows search functionality to be tailored to the platform's misinformation assessment objectives while keeping the underlying search infrastructure decoupled from business logic.

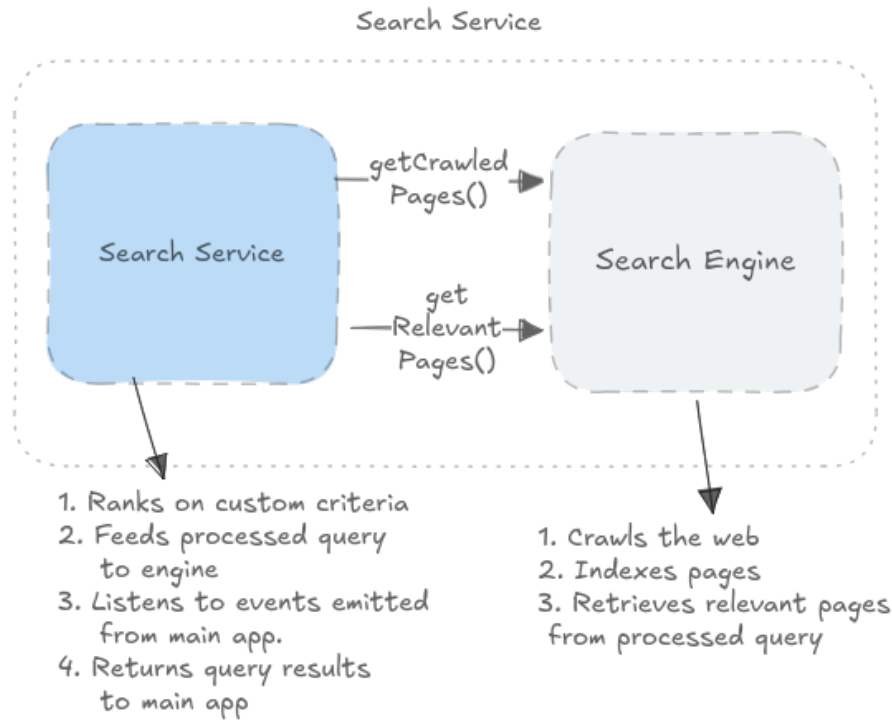


Figure 3.18: Close look at search service.

3.11.3 Iterative Algorithm and Aggregation Execution

Iterative Algorithm Process Following figure shows the iterative algorithm execution process within system bounds

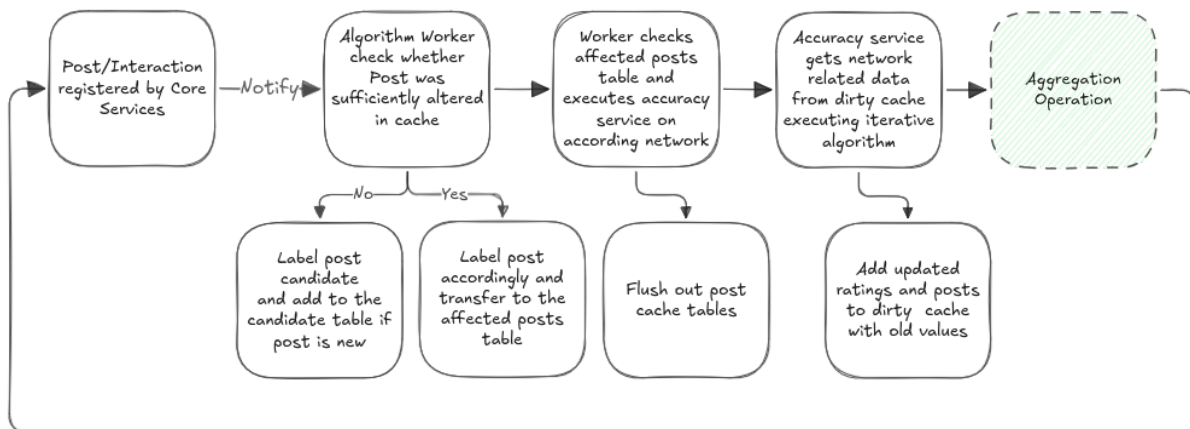


Figure 3.19: Iterative algorithm execution process.

Aggregation Operation Process consequently the implicated aggregation operation process is presented as follows:

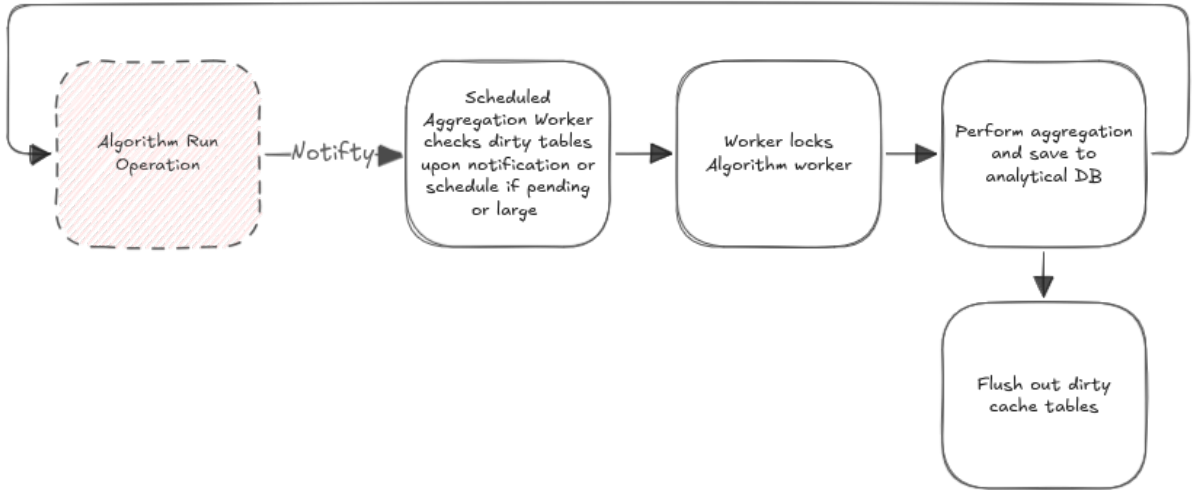


Figure 3.20: Aggregation operation execution process.

3.12 Conclusion

In this chapter, we laid out the theoretical foundations and the overall system architecture that underpin the proposed platform. We examined the key concepts, methodologies, and design choices that guide the development of the solution. Additionally, we introduced the major system components and their interactions within the architecture. These elements establish a coherent framework for implementing the proposed misinformation assessment system. The following chapters will focus on the practical realization and evaluation of these concepts.

Implementation and Testing

4.1 Introduction

This chapter presents the implementation and testing of the proposed rumor verification platform. It describes the development environment, software tools, and technologies used during the realization of the system. The chapter also presents the main interfaces of the application and demonstrates the operation of the platform through a set of testing scenarios. These tests illustrate the interaction between the different components of the system, including the AI analysis module, the community verification mechanism, and the expert validation process.

4.2 Work Environment

The following subsections present the hardware and software environments used during the implementation phase.

4.2.1 Hardware Environment

The application was developed on two machines whose hardware specifications are summarized in Table [4.1](#).

Hardware	Laptop 1	Laptop 2
Processor	AMD Ryzen 7 6800H	Intel Core i5-1135G7
RAM	16 GB DDR5	24 GB DDR4
Operating System	Windows 11 Professional	Ubuntu 26.04 LTS (Resolute Raccoon)

Table 4.1: Hardware Environment

4.2.2 Software and Libraries

The development of the proposed rumor verification platform required the use of various software tools, frameworks, and libraries. These technologies were selected to facilitate the implementation of the web application, database management, artificial intelligence modules, and communication between system components. Table 4.2 summarizes the main tools and libraries used during the development process.

Table 4.2: Libraries and Tools Used

Category	Tool / Library	Purpose
Frontend	HTML, CSS3, JavaScript	Development of the user interface.
Backend	ASP.NET, FastAPI, httpx, pydantic	Implementation of business logic and API services.
Database	Microsoft SQL Server	Storage and management of platform data.
Artificial Intelligence	transformers, torch,	Rumor analysis and credibility assessment.
Development Environment	JetBrains Rider, Visual Studio Code	Application development and debugging.
Visual Presentation	Excalidraw	Diagrams and illustrations.

4.3 Application Presentation

This section presents the main interfaces of the developed rumor verification platform. The application provides dedicated functionalities for different categories of user. The following subsections illustrate the implemented interfaces and describe their primary functionalities.

4.3.1 Home Page

The home page serves as the main entry point to the platform. It provides users with an overview of the system and allows them to access the platform’s main functionalities. Through this interface, users can browse recently verified rumors, search for specific claims, and navigate to the authentication and rumor submission pages. Figure 4.1 illustrates the home page of the application.

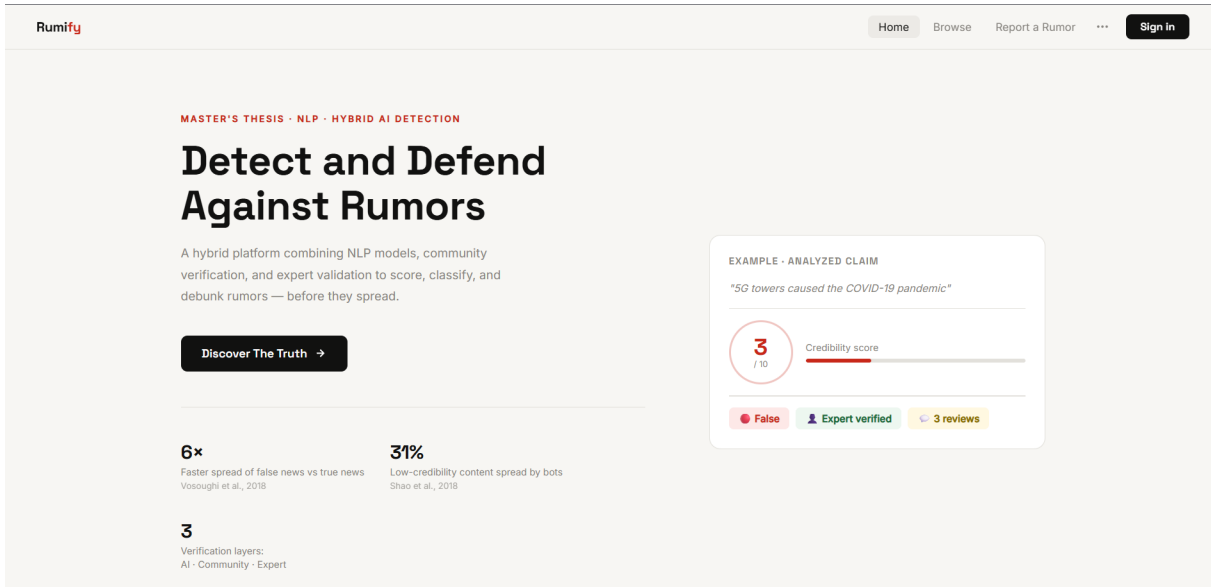


Figure 4.1: Home page of the rumor verification platform.

4.3.2 Browse and Search Interface

The browse and search interface enables users to explore previously submitted rumors and access existing verification reports. Users can search for claims, topics, or keywords through a semantic search mechanism capable of identifying relevant rumors even when different wording is used.

This functionality helps users quickly retrieve verified information and reduces the submission of duplicate rumors. Figures 4.2 and 4.3 illustrates the browse and search interface.

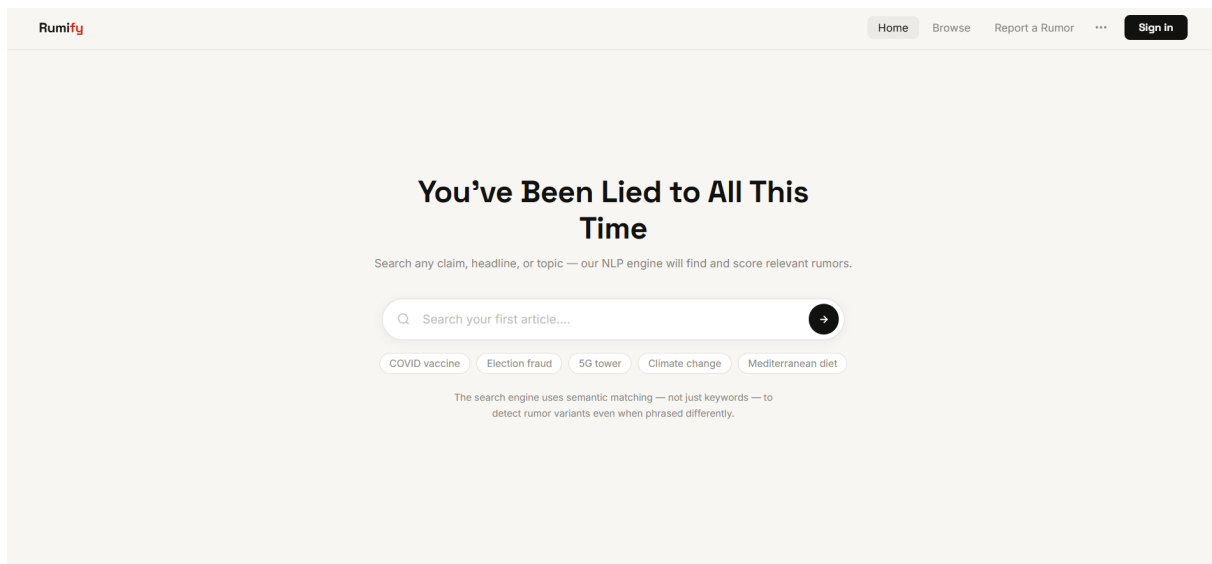


Figure 4.2: illustrates the search interface.

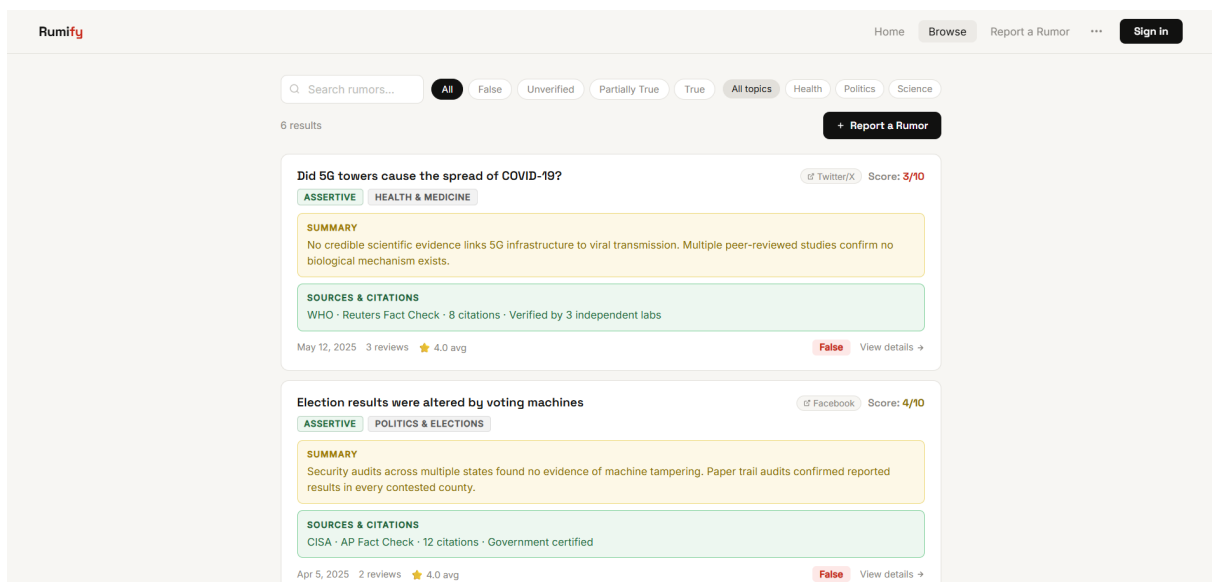


Figure 4.3: illustrates the Browse .

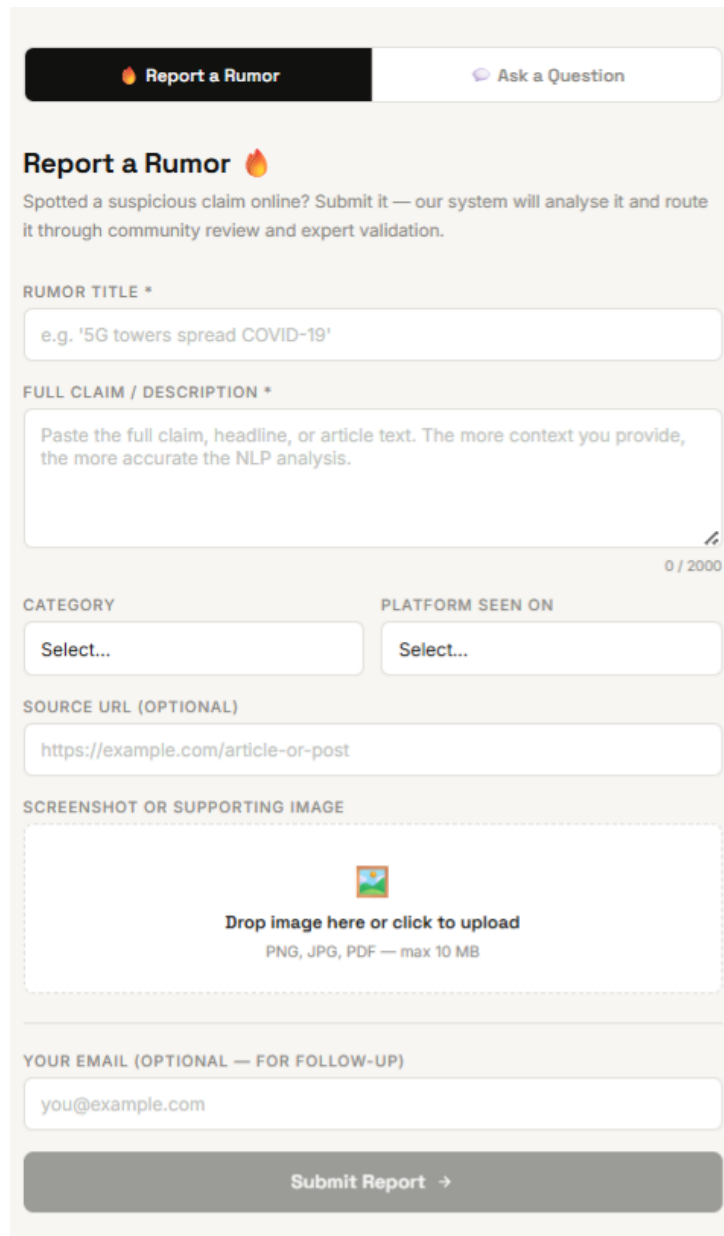
4.3.3 Rumor Submission Interface

The rumor submission interface enables users to report claims that require verification. Users can provide a rumor title, a detailed description of the claim, and additional contextual information such as the category and source platform. The interface also supports the submission of source URLs and multimedia files to facilitate the verification process.

Once submitted, the rumor is automatically published on the platform and becomes accessible to the community and experts for review and verification. Figure 4.4 illustrates

the rumor submission interface.

In addition to rumor reporting, the platform provides an *Ask a Question* feature, shown in Figure 4.5. This functionality allows users to submit questions or uncertain claims and receive evidence-based responses supported by AI analysis and expert review.



The image shows a web form for reporting a rumor. At the top, there are two tabs: "Report a Rumor" (active, with a flame icon) and "Ask a Question" (with a speech bubble icon). Below the tabs, the "Report a Rumor" section has a heading "Report a Rumor" with a flame icon. The text below the heading says: "Spotted a suspicious claim online? Submit it — our system will analyse it and route it through community review and expert validation." The form fields are: "RUMOR TITLE *" with a text input containing "e.g. '5G towers spread COVID-19'"; "FULL CLAIM / DESCRIPTION *" with a large text area containing "Paste the full claim, headline, or article text. The more context you provide, the more accurate the NLP analysis." and a character count "0 / 2000"; "CATEGORY" and "PLATFORM SEEN ON" both with "Select..." dropdown menus; "SOURCE URL (OPTIONAL)" with a text input containing "https://example.com/article-or-post"; "SCREENSHOT OR SUPPORTING IMAGE" with a dashed box containing an image icon, the text "Drop image here or click to upload", and "PNG, JPG, PDF — max 10 MB"; "YOUR EMAIL (OPTIONAL — FOR FOLLOW-UP)" with a text input containing "you@example.com"; and a "Submit Report →" button at the bottom.

Figure 4.4: Rumor submission interface.

The interface for asking a question is displayed within a light gray container. At the top, there are two buttons: 'Report a Rumor' with a flame icon and 'Ask a Question' with a speech bubble icon. Below these, the 'Ask a Question' section is titled with a speech bubble icon. A descriptive text states: 'Not sure if something is a rumor? Ask our analysts — we combine automated NLP scoring with human expert review to give you a reliable, evidence-based answer.'

The form consists of several fields:

- QUESTION TITLE ***: A text input field with a placeholder example: 'e.g. 'Is this vaccine claim true?''
- YOUR QUESTION IN DETAIL ***: A larger text area with a placeholder: 'Describe what you've read or heard, and what specifically you'd like verified or explained.' It includes a character count '0 / 2000' at the bottom right.
- CATEGORY**: A dropdown menu with the text 'Select...'.
- SUPPORTING IMAGE OR DOCUMENT**: A dashed border box containing a trash can icon, the text 'Attach screenshot or document', and the supported formats 'PNG, JPG, PDF — max 10 MB'.
- YOUR EMAIL (OPTIONAL)**: A text input field with the placeholder 'you@example.com'.

At the bottom of the form is a large gray button labeled 'Send Question →'.

Figure 4.5: Ask a Question interface.

4.3.4 Rumor Analysis and Verification Interface

The rumor analysis and verification interface provides a comprehensive view of a submitted claim and its current verification status. As shown in Figure 4.6, the page displays the original claim, its category, publication information, and the associated credibility assessment.

The interface presents a summary of the verification results, supporting sources and citations, and a set of NLP-based indicators, including sentiment, model confidence, toxicity level, and assertiveness. These indicators provide additional context for understanding the characteristics of the analyzed content.

In addition, the platform identifies known variants and mutations of the rumor across

different communication channels. This feature allows users to observe how a claim evolves and changes as it propagates through the information ecosystem.

The lower section of the page, shown in Figure 4.7, provides a community review area where users can submit credibility ratings, share supporting evidence, and discuss the claim. The interface also displays previously submitted reviews alongside the final expert verdict, enabling users to examine different perspectives and supporting arguments before reaching their own conclusions.

The interface shows a claim analysis for the statement: "Mediterranean diet reduces dementia risk by 23%". At the top, there are filters for "TRUE", "ASSERTIVE", "HEALTH & MEDICINE", and "NEWS WEBSITE". The claim is dated Dec 3, 2024, reported by science_digest, and has a community rating of 8.5/10 (2 reviews). The original source is PubMed.

SUMMARY (Score: 9/10): A 10-year longitudinal study found significant correlation between Mediterranean diet adherence and reduced cognitive decline markers, consistent with multiple prior smaller trials.

SOURCES & CITATIONS: Lancet Neurology · PubMed · 11 citations · Peer reviewed

ORIGINAL CLAIM: "A widely shared article reported that adherence to a Mediterranean-style diet was associated with a 23% reduction in dementia risk over a 10-year longitudinal follow-up study."

NLP ANALYSIS:

- Positive** Sentiment
- 92%** Model confidence
- None** Toxicity level
- High** Assertiveness

KNOWN VARIANTS & MUTATIONS (THESIS \$1.5):

- TWITTER/X**: "mediterranean diet reduces dementia risk by 23%"
- WHATSAPP**: Voice note: shortened version of the claim, stripped of context
- MEME**: Image with overlaid text exaggerating the original claim (Sharpening effect)

Figure 4.6: Rumor analysis and verification interface (Part 1).

EXPERT VERDICT — FINAL AUTHORITATIVE DECISION (LAYER 3)

DS
Dr. Sarah K.
Verified Expert · Peer reviewer

True

Study design is robust — prospective longitudinal with 60,000 participants, pre-registered methodology, and results consistent with 3 prior meta-analyses. The 23% relative risk reduction is well-supported.

Validated on Dec 7 2024

Community Reviews (2) Top rated ▾

Your rating: / 10 (0 = false · 10 = true)
Your verdict:

Share your analysis, evidence, or perspective. Well-reasoned comments with sources receive higher quality scores.

Be respectful and evidence-based.

Post review

NM
NutritionMD
EXPERT
TRUE

9/10 Dec 4 2024

The study is well-designed. The effect size is consistent with earlier smaller trials. The 23% figure is a relative risk reduction — important context for lay readers who may interpret it as absolute risk.

👍 38 helpful 🚩 Flag

R1
reader_102
TRUE

8/10 Dec 5 2024

Finally something positive and well sourced! Good to see credible health research being shared rather than panic-inducing claims.

👍 12 helpful 🚩 Flag

Figure 4.7: Rumor analysis and verification interface (Part 2: Community reviews and expert verdict).

4.3.5 Expert Dashboard

The expert dashboard provides a dedicated workspace for domain experts responsible for issuing final verification decisions. As shown in Figure 4.8, the dashboard displays rumors awaiting review together with relevant information, including their category, credibility score, community feedback, and priority level. This interface allows experts to efficiently identify and process claims that require authoritative assessment.

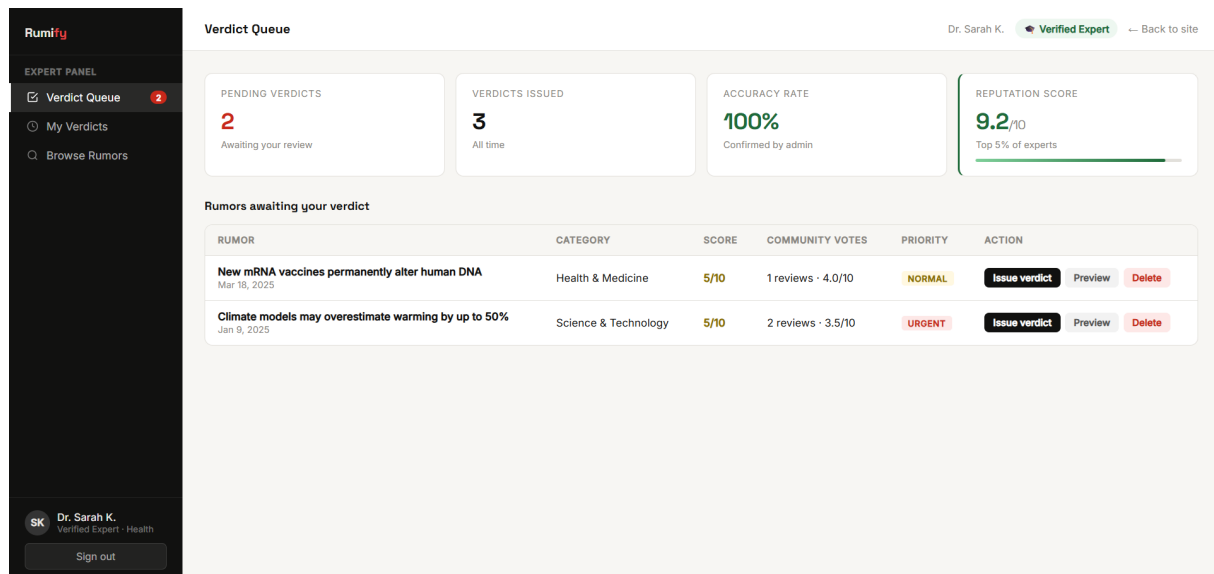


Figure 4.8: Expert dashboard showing pending rumors awaiting verification.

In addition to pending cases, the platform maintains a record of previously issued decisions. Figure 4.9 shows the verdict history interface, where experts can review their past evaluations, access supporting notes, and monitor the status of issued verdicts. The dashboard also presents performance indicators such as the number of completed reviews, accuracy rate, and reputation score.

Together, these features support the final stage of the verification process by providing experts with the information and tools required to issue well-informed and transparent decisions.

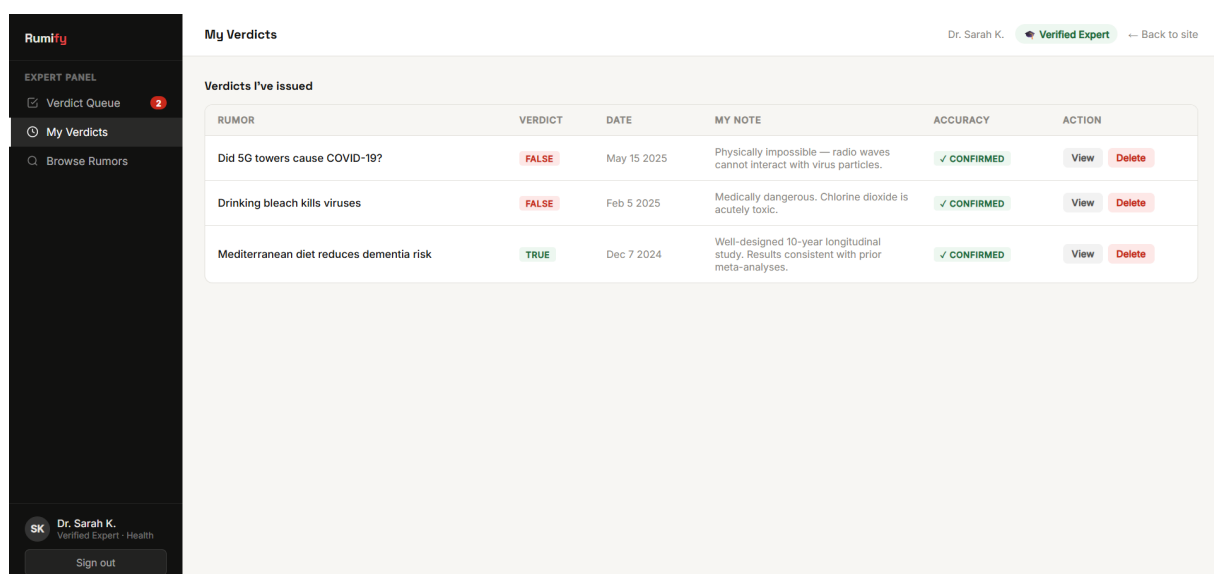


Figure 4.9: Expert dashboard showing the history of issued verdicts.

4.4 Current Limitations

Despite its potential benefits, the proposed rumor verification platform presents several limitations that should be acknowledged.

First, the effectiveness of the verification process depends heavily on the quality and availability of supporting evidence. Rumors for which limited information is available may be more difficult to assess accurately, regardless of the verification method employed.

Second, the platform relies on user participation and expert involvement throughout the verification process. Variations in community engagement, expert availability, or the quality of submitted evidence may affect the speed and consistency of rumor assessment.

Third, rumor verification remains a complex task due to the dynamic nature of online information. Claims may evolve over time, new evidence may emerge after a verdict has been issued, and the interpretation of certain rumors may depend on contextual or cultural factors.

Finally, while the proposed framework combines artificial intelligence, community contributions, and expert validation, no verification system can guarantee perfect accuracy. Consequently, verification results should be interpreted as informed assessments based on the available evidence rather than absolute statements of truth.

4.5 Conclusion

This chapter presented the implementation of the proposed Rumify platform and described its main functionalities through a series of user and expert interfaces. The development environment, software tools, and technologies used during the implementation process were first introduced, followed by a presentation of the platform's core components, including rumor submission, rumor exploration, verification results, community participation, and expert evaluation.

The implemented prototype demonstrates the practical applicability of the proposed hybrid rumor verification framework by integrating AI-assisted analysis, collaborative verification, and expert assessment within a unified platform. The conducted interface tests confirmed the proper functioning of the main features and validated the coherence of the proposed workflow. Overall, the implementation results show that the proposed solution provides a feasible foundation for supporting rumor verification and promoting

more informed information consumption.

General conclusion and perspectives

The proliferation of rumors across digital platforms represents one of the most pressing information integrity challenges of our time. Fueled by the speed of social media, the emotional appeal of unverified claims, and the structural limitations of existing detection tools, rumors today travel faster and farther than corrections can follow, causing measurable harm to public health, political processes, social cohesion, and institutional trust. Addressing this problem requires more than a technical fix. It demands a framework that is scalable, transparent, accountable, and sensitive to the human and social dimensions of information behavior.

Throughout this work, we set out to design and build such a framework. We began by establishing a comprehensive understanding of the rumor phenomenon, examining its definitions, typologies, propagation mechanisms, mutations, sociocultural roles, and consequences. This foundation enabled us to identify key limitations in existing rumor verification approaches, including the fragmentation of solutions, the lack of meaningful integration between artificial intelligence, community participation, and expert knowledge, the limited transparency of automated models, and the insufficient adaptation of existing systems to multilingual and regional contexts.

In response, we designed and implemented Rumify, a hybrid rumor verification platform that combines automated analysis, collaborative verification, and expert oversight within a unified framework. The platform is structured around three complementary verification layers. The first layer consists of an AI-based analysis module that combines a transformer-based text encoder with a GRU-based source credibility model to generate preliminary credibility and certainty estimates, providing an initial assessment that serves as a prior belief for subsequent verification stages. The second layer is a community-based

verification module in which user evaluations are weighted according to a dynamic reputation system that continuously adapts based on user behavior and historical performance. The third layer is an expert validation module responsible for issuing final authoritative verdicts in a transparent and traceable manner while automatically recalibrating community reputation scores according to the alignment of user assessments with the final expert decision.

A major contribution of this work is the design of a trust-weighted ranking framework extending the Iterative Adjustment of Ratings and Reputation (IARR) approach through the incorporation of Bayesian averaging. This extension enables the system to robustly combine machine learning predictions with community credibility signals, particularly in situations where participation is sparse or user credibility information is limited. The proposed framework therefore mitigates the impact of unreliable ratings while preserving the collective intelligence benefits of community participation.

Another contribution is the introduction of a multi-model feedback quality estimation architecture. Rather than relying solely on popularity indicators such as likes or votes, the proposed approach evaluates user commentary according to three intrinsic dimensions: text correctness, thoroughness, and quality of reasoning. By incorporating these criteria into the reputation framework, the platform encourages constructive and evidence-based participation while reducing the influence of low-quality contributions.

Finally, we implemented a functional prototype of the platform, including rumor submission and browsing capabilities, semantic search for duplicate rumor detection, AI-assisted rumor analysis, community review interfaces, and dedicated expert validation tools. The conducted interface tests confirmed the proper functioning of the platform’s core features across all user roles and validated the coherence of the proposed verification workflow.

As for perspectives, this work opens several directions for future development. On the algorithmic side, the ranking and reputation framework would benefit from empirical calibration using real-world datasets, while a rigorous evaluation of convergence behavior under large-scale and adversarial conditions remains to be conducted. The transformer–GRU credibility prediction model should be trained, fine-tuned, and extensively evaluated on rumor-specific datasets to assess its effectiveness across different domains and languages. Particular attention should be given to multilingual and Arabic-dialect content, which

remains both practically relevant and underrepresented in current research. The feedback quality estimation architecture also requires the integration and evaluation of suitable NLP tools capable of reliably assessing correctness, thoroughness, and reasoning quality in user-generated content.

Future versions of the platform could further incorporate graph-based propagation analysis to identify coordinated rumor campaigns and social influence patterns, as well as multimodal verification techniques capable of analyzing images, videos, and other non-textual content. From a governance perspective, establishing clear policies regarding expert accountability, appeal procedures, transparency mechanisms, and bias auditing would further strengthen the trustworthiness and legitimacy of the platform. Finally, a large-scale user study evaluating the platform’s effectiveness, usability, and impact on user behavior would constitute a valuable next step toward validating the proposed approach under real-world conditions and extending its practical applicability.

Bibliography

- [1] X Corp., “X (formerly twitter),” 2023. Accessed: 2025-06-21.
- [2] G. W. Allport and L. Postman, *The Psychology of Rumor*. Henry Holt and Company, 1947.
- [3] N. DiFonzo and P. Bordia, *Rumor Psychology: Social and Organizational Approaches*. American Psychological Association, 2007.
- [4] R. L. Rosnow, “Inside rumor: A personal journey.,” *American psychologist*, vol. 46, no. 5, p. 484, 1991.
- [5] C. R. Sunstein, *On rumors: How falsehoods spread, why we believe them, and what can be done*. Princeton University Press, 2014.
- [6] M. S. Islam, T. Sarkar, S. M. H. R. Khan, *et al.*, “Covid-19–related infodemic and its impact on public health,” *The American Journal of Tropical Medicine and Hygiene*, vol. 103, no. 4, pp. 1621–1629, 2020.
- [7] X. Zhou and R. Zafarani, “A survey of fake news: Fundamental theories, detection methods, and opportunities,” *ACM Computing Surveys*, vol. 53, no. 5, pp. 1–40, 2020.
- [8] O. Oh, M. Agrawal, and H. R. Rao, “Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises,” *MIS Quarterly*, vol. 37, no. 2, pp. 407–426, 2013.

-
- [9] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media: A data mining perspective,” *SIGKDD Explorations*, vol. 19, no. 1, pp. 22–36, 2017.
- [10] V. L. Rubin, Y. Chen, and N. J. Conroy, “Deception detection for news: Three types of fakes,” in *Proceedings of the Association for Information Science and Technology*, vol. 53, pp. 1–4, Wiley Online Library, 2016.
- [11] C. Wardle and H. Derakhshan, “Information disorder: Toward an interdisciplinary framework,” tech. rep., Council of Europe, 2017.
- [12] S. Vosoughi, D. Roy, and S. Aral, “The spread of true and false news online,” *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.
- [13] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, and W. Quattrociocchi, “The spreading of misinformation online,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 3, pp. 554–559, 2016.
- [14] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, “The spread of low-credibility content by social bots,” *Nature Communications*, vol. 9, no. 1, p. 4787, 2018.
- [15] C. Arun, “On whatsapp, rumors, and lynchings,” *The Economic & Political Weekly*, vol. 54, no. 6, pp. 30–35, 2019.
- [16] A. Zubiaga, M. Liakata, R. Procter, G. Wong Sak Hoi, and P. Tolmie, “Detection and resolution of rumors in social media: A survey,” *ACM Computing Surveys (CSUR)*, vol. 51, no. 2, pp. 1–36, 2018.
- [17] A. Friggeri, L. A. Adamic, D. Eckles, and J. Cheng, “Rumor cascades,” in *Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM)*, pp. 101–110, 2014.
- [18] H. J. Larson, L. Z. Cooper, J. Eskola, S. L. Katz, and S. Ratzan, “Addressing the vaccine confidence gap,” *The Lancet*, vol. 378, no. 9790, pp. 526–535, 2011.
- [19] R. Kuo and A. Marwick, “Critical disinformation studies: A syllabus,” *Harvard Kennedy School Misinformation Review*, vol. 2, no. 4, 2021.

-
- [20] G. A. Fine and B. Ellis, *The Global Grapevine: Why Rumors of Terrorism, Immigration, and Trade Matter*. Oxford University Press, 2010.
- [21] N. DiFonzo, *The Watercooler Effect: A Psychologist Explores the Extraordinary Power of Rumors*. Avery Publishing Group, 2008.
- [22] J. A. Tucker, A. Guess, P. Barbera, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, and B. Nyhan, “Social media, political polarization, and political disinformation: A review of the scientific literature,” *SSRN Electronic Journal*, 2018.
- [23] H. Allcott and M. Gentzkow, “Social media and fake news in the 2016 election,” *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, 2017.
- [24] P. C. Tetlock, “Giving content to investor sentiment: The role of media in the stock market,” *The Journal of Finance*, vol. 62, no. 3, pp. 1139–1168, 2007.
- [25] Y. Wang, F. Ma, Z. Jin, and J. Gao, “Fake news detection on social media: A data mining perspective,” *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2020.
- [26] A. Guess, J. Nagler, and J. Tucker, “Exposure to untrustworthy websites in the 2016 us election,” *Nature Human Behaviour*, vol. 4, pp. 472–480, 2020.
- [27] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, *et al.*, “The science of fake news,” *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.
- [28] J. Ma, W. Gao, and K.-F. Wong, “Detect rumors in microblog posts using propagation structure via kernel learning,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (R. Barzilay and M.-Y. Kan, eds.), (Vancouver, Canada), pp. 708–717, Association for Computational Linguistics, July 2017.
- [29] C. Castillo, M. Mendoza, and B. Poblete, “Information credibility on twitter,” in *Proceedings of the 20th international conference on World wide web*, pp. 675–684, 2011.

-
- [30] J. Ma, W. Gao, and K.-F. Wong, “Detecting rumors from microblogs with recurrent neural networks,” in *Proceedings of IJCAI*, 2016.
- [31] R. Anggrainingsih, G. M. Hassan, and A. Datta, “Bert based classification system for detecting rumours on twitter.” <https://arxiv.org/abs/2109.02975>, 2021. Accessed June 2025.
- [32] F. Alam, F. Ofli, and M. Imran, “Crisismmd: Multimodal twitter datasets from natural disasters,” *arXiv preprint arXiv:1805.00713*, 2018.
- [33] R. Française, “Loi n°2020-766 visant à lutter contre les contenus haineux sur internet (“loi avia”).” *Journal Officiel*, June 2020. *Journal Officiel de la République Française*.
- [34] B. Deutschland, “Gesetz zur verbesserung der rechtsdurchsetzung in sozialen netzwerken (netzdg).” *Bundesgesetzblatt*, June 2017. *Bundesgesetzblatt I*, 131–136.
- [35] E. Parliament and Council, “Regulation (eu) 2022/2065 on a single market for digital services (digital services act).” *Official Journal of the European Union*, Nov. 2022. *Official Journal of the European Union*, L 277, 1–149.
- [36] République Algérienne Démocratique et Populaire, “Loi n° 20–06 du 28 avril 2020 modifiant et complétant l’ordonnance n° 66–156 portant code pénal.” *Journal Officiel de la République Algérienne*, Apr. 2020. N° 25, pp. 4–6.
- [37] République Algérienne Démocratique et Populaire, “Loi n°24-06 du 28 avril 2024 modifiant et complétant l’ordonnance n°66-156 du 8 juin 1966 (code pénal).” *Journal Officiel de la République Algérienne*, Apr. 2024. Amendments criminalize speech undermining security or institutions.
- [38] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, “Fake news detection on social media using geometric deep learning,” *arXiv preprint arXiv:1902.06673*, 2019.
- [39] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, “Multimodal fusion with recurrent neural networks for rumor detection on microblogs,” in *Proceedings of the 25th ACM international conference on Multimedia*, pp. 795–816, 2017.

-
- [40] A. Rao, N. Spasojevic, Z. Li, and T. DSouza, “Klout score: Measuring influence across multiple social networks,” in *2015 IEEE International Conference on Big Data (Big Data)*, pp. 1–8, IEEE, 2015.
 - [41] A. Alqwadri, M. Azzeh, and F. Almasalha, “Application of machine learning for online reputation systems,” *arXiv preprint arXiv:2209.04650*, 2022.
 - [42] H. Liao, A. Zeng, R. Xiao, Z.-M. Ren, D.-B. Chen, and Y.-C. Zhang, “Ranking reputation and quality in online rating systems,” *PLoS ONE*, vol. 9, no. 5, 2014.
 - [43] P. Laureti, L. Moret, Y.-C. Zhang, and Y.-K. Yu, “Information filtering via iterative refinement,” *Europhysics Letters*, vol. 75, aug 2006.
 - [44] K. Endo, R. Horio, and D. Okano, “A ranking algorithm for mitigating the influence of contrived ratings on review sites,” *Computer Science and Information Technology*, vol. 3, no. 6, pp. 1–8, 2015.
 - [45] Wikipedia contributors, “Bayesian average — wikipedia, the free encyclopedia,” 2026. Accessed: 2026-05-16.
 - [46] World Health Organization, “Epi-win: Who’s information network for epidemics,” 2020. Accessed June 2025.
 - [47] N. Hassan, C. Li, and M. Tremayne, “Detecting check-worthy factual claims in presidential debates,” in *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM)*, 2015. Accessed June 2025.
 - [48] Meta Platforms Inc., “Partnering with third-party fact-checkers.” <https://www.facebook.com/journalismproject/programs/third-party-fact-checking/selecting-partners>, 2020. Accessed June 2025.
 - [49] American Psychological Association, “Reference examples: Entries beginning with x,” n.d. Accessed: 2025-06-21.