BULGARIAN ACADEMY OF SCIENCES

**INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGIES**

Department: “Information Processes and Decision Support Systems”

**Kristel Gjergj Vula**

# Transformer-Based Detection and Epidemic-Model Analysis of Fake-News Dynamics on Social Network

## ABSTRACT

for acquiring Educational and Scientific Degree

“DOCTOR”

In Professional field 4.6. “Informatics and Computer Science”,

in Doctoral Program: Informatics

**Sofia, 2026**

The dissertation has been discussed and admitted to defense at an extended meeting of the Department of Information Processes and Decision Support Systems at the IICT–BAS, held on 17.06.2026

The dissertation is structured into an introduction, four chapters, general conclusions on the results achieved, thesis contributions, a list of publications related to the thesis, a declaration of originality, a bibliography, and appendices. The dissertation has a total of 210 pages, 47 figures, 111 tables, and 191 bibliography sources.

The defense of the dissertation will take place on ……………………. 2026 from ……………… o'clock in hall/room ……………………. of bl. 2 of IICT–BAS at an open session of a scientific jury composed of:

Scientific jury:

1. ……………………………………………………………
2. ……………………………………………………………
3. ……………………………………………………………
4. ……………………………………………………………
5. ……………………………………………………………

The reviews and opinions of the members of the scientific jury and the abstract of the dissertation are published on the website of IICT–BAS. The materials for the defense are available to anyone interested in room ………. of IICT-BAS, bl. 2, Acad. G. Bonchev Str., Sofia.

Author: **Kristel Vula**

Title: **Transformer-Based Detection and Epidemic-Model Analysis of Fake-News Dynamics on Social Network**

## Dissertation Structure

Vula, K. Transformer-Based Detection and Epidemic-Model Analysis of Fake-News Dynamics on Social Networks. Scientific supervisor: Prof. Dr. Vassil Georgiev Guliashki. Department: Information Processes and Decision Support Systems. Doctoral Program: Informatics. Sofia, 2026.

The dissertation consists of an introduction, four chapters, a conclusion section with future research directions, a separate thesis-contributions section, a list of five publications related to the dissertation, a declaration of originality, a bibliography of 191 references and appendices with supplementary reproducibility materials.

The introduction motivates the dissertation by explaining why fake news must be analysed both as textual content and as a social diffusion process. It introduces the need for a model that connects content-level detection with temporal spread, epidemic-style interpretation and uncertainty-aware inference.

Chapter 1 presents the theoretical foundations of fake-news detection, Transformer-based language modelling, temporal networks, SIR/SEIZ epidemic models and Approximate Bayesian Computation. It identifies the research gaps that motivate the integrated framework.

Chapter 2 defines the proposed Transformer-ABC-Epidemic framework. It presents dataset selection, preprocessing, binary label harmonization, temporal cascade construction, SIR/SEIZ modelling, ABC/ABC-SMC inference, posterior prediction, intervention scenarios, ethics and reproducibility boundaries.

Chapter 3 reports the experimental results. It evaluates Transformer-based fake-news detection across content datasets, reconstructs temporal cascades, compares SIR and SEIZ diffusion models, estimates posterior uncertainty using ABC and ABC-SMC, and tests counterfactual intervention scenarios.

The concluding section summarizes the obtained results, limitations and future research. The thesis contributions and the publications connected to the dissertation are presented after the conclusions.

## Keywords

Fake-news detection; Transformer models; Temporal networks; SIR/SEIZ epidemic modelling; Approximate Bayesian Computation; Intervention analysis.

# Introduction

This abstract presents the dissertation as an integrated study of fake-news detection and diffusion on social networks. It treats fake news as both a textual credibility problem and a dynamic process shaped by visibility, interaction and platform amplification. The thesis therefore connects Transformer-based detection with temporal cascade reconstruction, SIR/SEIZ modelling and ABC/ABC-SMC inference [1]-[6].

The main methodological idea is a content-to-cascade framework: Transformer models estimate fake-news probability, timestamped interactions show social activation, SIR/SEIZ models describe spreading trajectories and ABC/ABC-SMC estimates posterior uncertainty. This layered view keeps classification accuracy, cascade size and posterior forecasts as distinct evidence types.

The interpretation remains cautious: reposts are observable spreading events rather than proof of belief, missing interactions are not treated as rejection, and intervention scenarios are controlled counterfactual simulations rather than causal platform-policy measurements. The framework is therefore designed for reproducible analysis under uncertainty.

The empirical part evaluates short-text, long-text, healthcare and low-resource Albanian fake-news detection, reconstructs principal temporal cascades, compares SIR and SEIZ dynamics, estimates posterior parameter regions and examines how early debunking or visibility reduction may influence future cascade growth. The final contribution is an auditable framework that links textual credibility, temporal amplification, epidemic-state interpretation and posterior predictive risk [7], [8].

# Chapter 1. Literature Review and Theoretical Foundations

Chapter 1 builds the conceptual and methodological foundation of the dissertation. It reviews definitions of fake news, classical and Transformer-based detection, ensemble and hybrid approaches, temporal-network diffusion, epidemic and rumour-spreading models, and Approximate Bayesian Computation. The chapter concludes by identifying research gaps and deriving the dissertation goal, research questions and tasks.

## 1.1 Introduction

Chapter 1 establishes the theoretical foundation of the dissertation. It shows that fake-news research has developed through several complementary directions: conceptual definitions, classical machine learning, contextual Transformer detection, temporal networks, epidemic and rumour models and ABC-based uncertainty inference. The chapter therefore prepares the transition from a simple fake/real classification task to an integrated detection-diffusion framework [1], [5], [6].

## 1.2 Defining Fake News

### 1.2.1 Fake News: Definitions and Characteristics

Fake news is treated in the dissertation as false, misleading, partially false or manipulatively framed information distributed in a news-like or claim-like form. This working definition is broad enough to include articles, social-media posts, health misinformation and political narratives, while still preserving the distinction between labelled data and verified truth. Binary labels are used for modelling, but the resulting score is interpreted as a probabilistic credibility signal rather than an absolute fact-checking judgement [1].

### 1.2.2 Related Concepts

The related concepts in Table 1.1 clarify why misinformation, disinformation, malinformation, rumour, satire, clickbait, doubtful news and infodemic cannot be collapsed without methodological caution. The table is retained because it defines the vocabulary used later for label harmonization, dataset interpretation and the distinction between content-level detection and diffusion-level modelling [1].

Table 1.1. Main concepts related to fake news and their relevance to detection.

| Concept | Meaning | Relevance to fake-news detection |
|---|---|---|
| **Misinformation** | False or inaccurate information that may be shared without deliberate intent to deceive. | Important when the intention of the source is unknown, unavailable, or not directly observable in the dataset. |
| **Disinformation** | False or misleading information produced or disseminated deliberately to deceive. | Closest to the narrow definition of fake news and highly relevant to intentional manipulation [1]. |
| **Malinformation** | True or partly true information used in a misleading context or with harmful intent. | Requires contextual analysis because the problem may lie in framing, timing, or selective disclosure. |
| **Rumors** | Unverified information that circulates before its truth status is confirmed or rejected. | Relevant to temporal verification, evolving labels, and diffusion before confirmation [6]. |
| **Propaganda** | Strategically framed communication designed to influence beliefs, attitudes, or behaviour. | May overlap with fake news when selective facts, emotional framing, or fabricated claims are used to influence audiences. |
| **Satire / parody** | Content that intentionally imitates news style for humour, criticism, or exaggeration. | Can be misclassified as fake news unless genre, intention, and publication context are considered. |

| | | |
|---|---|---|
| **Clickbait** | Sensational or misleading headlines designed to attract attention and increase engagement. | May be false, partly true, or simply manipulative; useful for detecting engagement-driven misinformation. |
| **Doubtful news** | Content whose truthfulness is uncertain, mixed, weakly supported, or insufficiently verified. | Important for uncertainty-aware and delayed-decision classification rather than forced binary labelling. |
| **Infodemic** | A large-scale mixture of accurate, inaccurate, and misleading information, especially during crises. | Highly relevant to public-health misinformation and COVID-19 fake-news research. |

### 1.2.3 Challenges in Defining Fake News

The main definitional challenge is that truthfulness may depend on context, timing, source evidence and later verification. Annotation ambiguity, sarcasm, satire, missing context and partial truth can cause disagreement among fact-checkers or dataset annotators. This justifies the thesis choice to preserve original labels as metadata, exclude ambiguous records when necessary and interpret model outputs as uncertainty-aware credibility estimates [1].

## 1.3 Classical Machine Learning Approaches for Fake-News Detection

### 1.3.1 Text Representation Methods

Classical fake-news detection represents text through Bag-of-Words, TF-IDF, word or character n-grams and stylometric features. These representations are efficient and interpretable, but they are limited when misinformation depends on context, paraphrase, irony or long-range semantic relations. They remain important in the dissertation because they provide transparent baselines for evaluating Transformer-based detection [1].

### 1.3.2 Classical Classification Models

Table 1.2 summarizes the main classical classifiers retained as baseline literature: Logistic Regression, Naive Bayes, SVM, Random Forest and boosting methods. Their value is interpretability and reproducibility, while their limitation is weak contextual understanding and sensitivity to domain shift [1].

Table 1.2. Classical machine learning approaches for fake-news detection.

| Approach | Typical features | Representative literature | Strengths | Limitations |
|---|---|---|---|---|
| **Logistic Regression** | TF-IDF, n-grams, metadata | Shu et al. [1] | Simple, interpretable, and strong as a linear baseline. | Limited ability to model nonlinear semantic relations. |
| **Naive Bayes** | Word counts, binary terms, TF-IDF | Shu et al. [1] | Fast, scalable, and effective for small text datasets. | Relies on a strong conditional-independence assumption. |
| **Support Vector Machine** | TF-IDF, word and character n-grams | Shu et al. [1] | Effective in high-dimensional sparse spaces. | Requires careful kernel and parameter selection for large datasets. |
| **Random Forest** | Stylometric, lexical, readability, and metadata features | Shu et al. [1] | Robust to noisy engineered features and useful for feature importance. | Less suitable for very sparse text without feature reduction. |
| **Gradient Boosting / XGBoost** | Lexical, stylistic, metadata, and structured features | Chen and Guestrin | Strong predictive performance with engineered features. | Requires tuning and may overfit noisy or biased labels. |

### 1.3.3 Advantages and Limitations

Classical models remain useful because they reveal whether a dataset is separable through surface lexical or source-style cues. However, fake-news texts often involve implicit contradiction, careful framing and topic transfer, which require contextual representations. This limitation motivates the Transformer layer of the proposed framework while preserving classical methods as comparison points.

## 1.4 Transformer-Based Models for Fake-News Detection

### 1.4.1 Transformer Architecture and Attention Mechanism

Transformer models use self-attention to represent relations among tokens, entities, numbers and claims within a sequence. This mechanism is central for fake-news detection because misleading content often depends on contextual relations rather than isolated words. The reviewed attention principle explains why BERT-style models can improve over sparse lexical baselines [2].

### 1.4.2 BERT and RoBERTa-Based Detection

BERT and RoBERTa are reviewed as the main contextual architectures for binary fake-news classification. BERT provides bidirectional contextual encoding, while RoBERTa improves training choices and often increases robustness. In the thesis framework, their role is to estimate a fake-news probability that can later be connected to temporal cascade evidence [2].

### 1.4.3 Multilingual Transformer Models

Multilingual BERT and XLM-RoBERTa are relevant because misinformation research should not remain limited to English datasets. Their shared subword representations make it possible to evaluate low-resource Albanian fake-news detection while preserving the same binary detection logic used for English and healthcare datasets.

### 1.4.4 Long-Sequence Models

Long-document fake-news detection requires models that reduce the information loss caused by truncation. BigBird is reviewed because sparse attention allows longer article evidence to be processed more efficiently, making it suitable when misleading cues appear beyond the first part of an article [3].

### 1.4.5 Sentence-Level, LLM and Emerging Directions

SBERT and large language models are treated as relevant but auxiliary directions. Sentence embeddings support claim comparison, while LLMs may assist explanation or controlled augmentation. However, generated outputs are not treated as labels or ground truth because hallucination, bias and overconfidence remain methodological risks.

Table 1.3. Main Transformer-based models relevant to fake-news detection.

| Model | Core idea | Relevance to fake-news detection |
|---|---|---|
| **BERT** | Bidirectional Transformer encoder pre-trained with masked language modelling. | Strong baseline for contextual fake-news classification [2]. |
| **RoBERTa** | Optimized BERT training with more data, dynamic masking, and no next-sentence prediction. | Often improves robustness and classification performance. |
| **mBERT** | Multilingual BERT trained across 104 languages. | Useful for cross-lingual or multilingual fake-news detection. |
| **BigBird** | Sparse-attention Transformer that combines local/window, random, and global attention to model long sequences with linear attention complexity. | Useful for full-article fake-news detection because it reduces information loss caused by truncating long articles to the first 512 tokens [3]. |
| **SBERT** | Sentence-level embeddings based on siamese/triplet BERT networks. | Useful for semantic similarity, claim comparison, and sentence-level classification. |

| | | |
|---|---|---|
| **XLM-RoBERTa** | Cross-lingual RoBERTa trained on large multilingual CommonCrawl data. | Useful for multilingual detection and transfer across languages. |
| **Large Language Models** | Generative Transformer models used for augmentation, explanation, and zero-shot tasks. | Support augmentation and zero-shot detection; require caution due to hallucination and bias. |

### 1.4.6 Limitations of Transformer-Based Detection

Transformer detectors can be computationally expensive, less transparent than classical baselines and vulnerable to adversarial paraphrase or domain shift. Their probabilities must therefore be validated and interpreted as content-risk scores, especially when they are later used as inputs for cascade and epidemic modelling.

## 1.5 Ensemble and Hybrid Approaches for Fake-News Detection

Ensemble and hybrid detectors are reviewed because they can combine lexical, neural, metadata or graph-based signals. The dissertation does not implement a voting or stacking ensemble as the core detector, because the downstream diffusion layer requires a clear and auditable Transformer probability. The framework is hybrid at the analytical level: content detection, temporal networks, SEIZ dynamics and ABC uncertainty are linked as separate evidence layers.

## 1.6 Temporal-Network Approaches to Fake News Diffusion

Temporal networks provide the bridge between fake-news detection and social-risk modelling. Unlike static graphs, they preserve the chronological order of retweets, replies, reposts or shares, allowing only time-respecting paths. The retained cascade quantities, including new active spreaders, cumulative spreaders, time to peak and inter-event delay, become the empirical summaries used later for SIR/SEIZ fitting and ABC inference [5], [6].

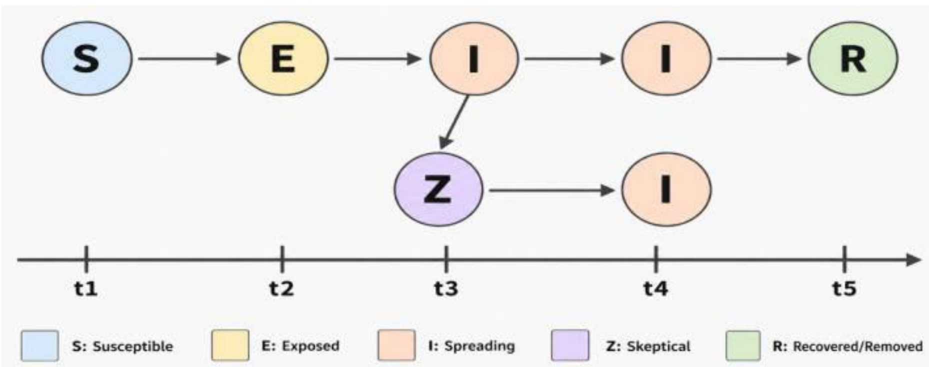



Figure 1.1. Temporal fake-news cascade with time-ordered user states.

## 1.7 Epidemic and Fake-News Spreading Models

Epidemic and rumour models offer an interpretable language for modelling fake-news diffusion through states such as susceptible, exposed, active spreading, recovered and skeptic. Table 1.4 is retained because it maps biological terminology to observable social-media proxies while emphasizing that online sharing is only an approximation of contagion [9], [10].

Table 1.4. Mapping between epidemic terminology and fake-news propagation.

| Epidemic concept | Disease interpretation | Fake-news interpretation | Observable proxy in social data |
|---|---|---|---|
| **Susceptible** | Individual who can be infected | User who has not yet engaged with the claim but may be exposed | Follower relation, community membership, timeline exposure |
| **Exposed** | Individual infected but not yet infectious | User who has seen a claim but has not yet shared it | View/impression when available; delayed reply; time lag before repost |
| **Infected** | Individual currently transmitting disease | User actively spreading the claim | Retweet, repost, quote, comment, share |
| **Recovered/Removed** | Individual no longer infectious | User stops spreading, loses interest, or becomes corrected | No further sharing after a correction or after cascade decay |
| **Skeptic** | Not standard in SIR; used in rumor models | User exposed to the claim but not spreading it, or rejecting/debunking it | Debunking reply, correction, non- engagement after exposure |

### 1.7.1 Classical Epidemic and Rumour Models

Classical SI, SIS, SIR and SEIR models provide the state-transition vocabulary, while Daley-Kendall and Maki-Thompson rumour models introduce saturation, rejection and stopping behaviour. This literature justifies using SIR as a transparent baseline and SEIZ as the richer model for fake-news cascades [9].

Table 1.5. Classical epidemic models and their interpretation for fake-news diffusion.

| Model | Core states | Representative literature | Fake-news interpretation | Main strength and limitation |
|---|---|---|---|---|
| **SI** | S -> I | Kermack and McKendrick [9] | Users move from potential exposure to active spreading. | Useful for early uncontrolled diffusion, but has no correction, forgetting or recovery state. |
| **SIS** | S -> I -> S | Kermack and McKendrick [9] | Users may stop spreading and later become susceptible again after repeated exposure. | Useful for recurrent sharing, but does not represent permanent correction. |
| **SIR** | S -> I -> R | Kermack and McKendrick [9]; Jin et al. [10] | Users share fake news and later stop, forget, become corrected or are removed from the active cascade. | Interpretable baseline with transmission and recovery parameters, but no explicit exposure delay or skepticism. |

| | | | | |
|---|---|---|---|---|
| **SEIR** | S -> E -> I -> R | Kermack and McKendrick [9] | Users may observe content before deciding whether to share it. | Captures exposure delay, but does not explicitly represent skeptical or debunking behaviour. |

### 1.7.2 SIR and SEIZ for Fake-News Diffusion

SIR represents users as susceptible, active spreaders and recovered or inactive participants. SEIZ extends this structure with exposed and skeptic or non-spreading states, making it more suitable for misinformation processes involving delay, hesitation and correction. The figures and coefficient table are kept because they define the modelling interpretation used in Chapter 3 [10].

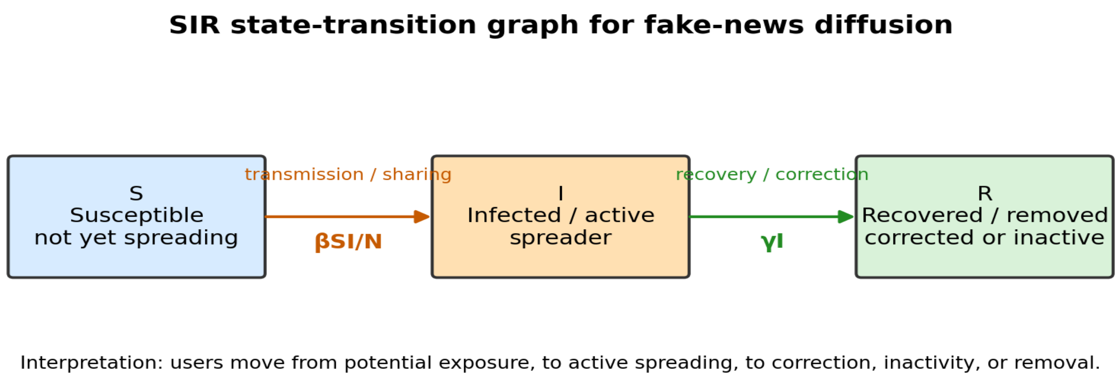



Figure 1.2. SIR state-transition graph for fake-news diffusion.

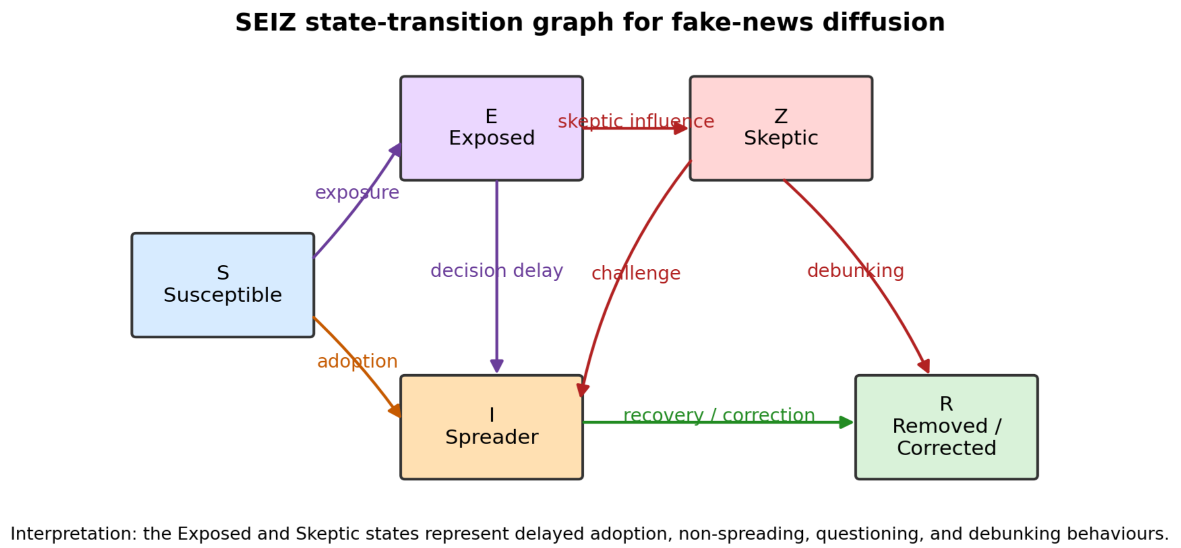



Figure 1.3. SEIZ state-transition graph for fake-news diffusion.

Table 1.6. Main coefficients in SIR and SEIZ fake-news diffusion models.

| Coefficient | Model | Meaning in fake-news diffusion | Possible empirical proxy |
|---|---|---|---|
| **β** | SIR / SEIZ | Transmission rate from active spreaders to susceptible users | Repost rate, exposure rate, edge weight, influencer centrality |
| **γ** | SIR / SEIZ | Recovery, forgetting or stopping rate | Time until no further sharing; cascade decay |
| **b** | SEIZ | Contact rate between susceptible users and skeptical users | Replies by debunkers; corrective interactions |
| **p** | SEIZ | Probability of direct adoption after contact with a spreader | Immediate repost probability after first exposure |
| **l** | SEIZ | Probability of becoming skeptical after contact with skeptics | Correction replies, fact-checking exposure, non-engagement after debunking |
| **ε** | SEIZ | Transition rate from exposed to active spreading | Delay from exposure or reply to repost |
| **ρ** | SEIZ | Influence of active spreaders on exposed users | Repeated exposure; discussion threads that lead to adoption |
| **δ** | Advanced SEIZ | Correction or debunking rate | Warning labels, fact-checks, official corrections |

### 1.7.3 Limits of the Epidemic Analogy

The epidemic analogy is useful but incomplete. Platform data do not reveal every exposure, private share, deleted post or silent rejection, and reposting does not prove belief. This limitation motivates posterior uncertainty estimation rather than deterministic point fitting.

## 1.8 Approximate Bayesian Computation for Diffusion Inference

ABC is reviewed because fake-news diffusion can be simulated but its exact likelihood is difficult to write. Candidate SIR or SEIZ parameters generate synthetic cascades; simulated summaries are compared with observed summaries; and accepted particles approximate posterior uncertainty. ABC-SMC refines this process through sequential tolerance reduction [4].

## 1.9 Research Gaps and Open Problems

The literature gaps are connected: detection is often separated from diffusion, static graphs hide time order, epidemic parameters are uncertain, datasets support different evidence layers and intervention modelling is usually detached from posterior uncertainty. These gaps motivate the Transformer-ABC-Epidemic framework, where content probability, temporal cascade evidence, SEIZ dynamics and ABC/ABC-SMC posterior prediction are analysed together [7], [8].

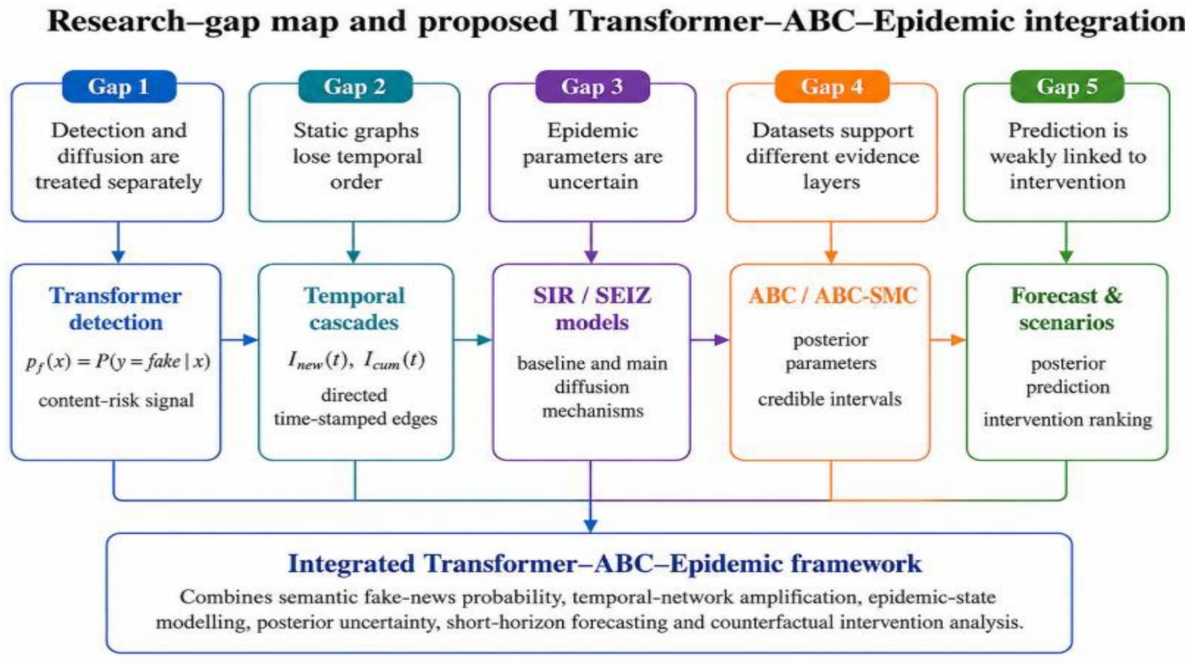



Figure 1.4. Research-gap map and proposed Transformer-ABC-Epidemic integration.

Table 1.7. Research gaps, methodological responses and expected evidence.

| Research gap | Methodological response in Chapter 2 | Expected evidence in Chapter 3 | Contribution to the dissertation |
|---|---|---|---|
| **Detection and diffusion are often treated separately.** | Use Transformer fake-news probability as a content-risk signal and link it to temporal cascade construction. | Detection metrics, claim-level probabilities and cascade eligibility outputs. | Connects content credibility with diffusion- risk analysis. |
| **Static graphs can hide time-respecting diffusion paths.** | Represent claims as directed temporal cascades with ordered interactions and time-indexed spreading curves. | I_new(t), I_cum(t), cascade size, time to peak and temporal-network figures. | Preserves the chronology required for fake-news diffusion modelling. |
| **SIR/SEIZ parameters are uncertain and only partially observable.** | Estimate parameter uncertainty through ABC and ABC-SMC using observed cascade summaries. | Posterior estimates, credible intervals and posterior predictive calibration. | Provides uncertainty- aware epidemic-model inference for fake-news cascades. |
| **Content datasets and diffusion datasets support different forms of evidence.** | Separate content-oriented datasets from diffusion-oriented datasets and define explicit dataset roles. | Distinct detection experiments and diffusion experiments with non-interchangeable dataset use. | Improves methodological validity and prevents unsupported dataset claims. |
| **Mitigation analysis requires more than classification labels.** | Use posterior prediction and counterfactual intervention scenarios under ABC-calibrated diffusion models. | Forecasts, intervention scenarios and baseline diffusion-risk functional outputs. | Supports intervention-oriented reasoning from the integrated framework. |

### 1.9.1 Research Questions Derived from the Literature Gaps

1) RQ1. How effectively can Transformer models classify fake and real news across short-text, long-text, multilingual and low-resource datasets?
2) RQ2. How can fake-news claims be represented as temporal cascades suitable for epidemic-style modelling?
3) RQ3. Does SEIZ provide a more suitable representation than SIR for fake-news diffusion when exposure delay and skepticism are relevant?
4) RQ4. Can ABC and ABC-SMC estimate uncertainty in fake-news diffusion parameters from partially observed cascades?
5) RQ5. Can a combined Transformer-temporal-network-ABC framework support posterior prediction and intervention-oriented fake-news risk analysis?

## 1.10 Chapter Conclusion

Chapter 1 concludes that fake-news analysis requires integrated evidence rather than isolated classification accuracy. The dissertation goal is therefore to develop and evaluate a framework that connects Transformer-based language models, temporal-network representation, SIR/SEIZ epidemic modelling and ABC/ABC-SMC posterior inference. The research tasks derived from this goal cover literature review, dataset separation, binary detection, cascade construction, SIR-SEIZ comparison, uncertainty-aware inference and intervention-oriented risk synthesis.

Table 1.8. Compact synthesis of reviewed literature and relevance to the proposed thesis.

| Literature direction | Main strength | Remaining limitation | Relevance to this thesis |
|---|---|---|---|
| **Fake-news definitions and annotation** | Clarifies fake, doubtful, misleading and related concepts. | Labels may be subjective, incomplete or unstable over time. | Motivates careful binary mapping and discussion of annotation ambiguity. |
| **Classical machine learning** | Interpretable, efficient and useful as baseline methodology. | Limited contextual understanding and weak transfer under domain shift. | Provides historical background and explainable comparison points. |
| **Transformer-based detection** | Captures contextual semantics and supports probability scoring. | Requires computation, data quality and domain adaptation. | Forms the main content-detection layer of the proposed framework. |
| **Probability calibration and robustness** | Supports reliable interpretation of model scores and cross-dataset stability. | Does not replace the need for temporal diffusion modelling. | Motivates use of calibrated Transformer probability as an input to the diffusion-risk layer. |

| | | | |
|---|---|---|---|
| **Temporal-network analysis** | Preserves ordering, delays and time-respecting paths. | Often disconnected from content credibility labels. | Supports cascade construction and temporal diffusion and path-memory features. |
| **Epidemic and rumor models** | Represent spreading states, thresholds and cascade dynamics. | Human behaviour and hidden exposure are only approximated. | Provides SIR and SEIZ foundations for fake-news propagation analysis. |
| **ABC and ABC-SMC inference** | Estimates uncertain parameters when likelihoods are intractable. | Sensitive to summaries, identifiability and computational cost. | Enables uncertainty-aware calibration of SIR/SEIZ diffusion models. |
| **Intervention-oriented modelling** | Connects prediction with mitigation strategies. | Often absent from standard detection-only approaches. | Motivates posterior prediction and intervention analysis in Chapter 2. |

# Chapter 2. Proposed Transformer–ABC–Epidemic Framework and Methodology.

## 2.1 Overall Proposed Framework

Chapter 2 operationalizes the theoretical framework. The proposed methodology has four layers: a data layer for content and diffusion datasets, a Transformer-based binary detection layer, a temporal-network and epidemic modelling layer and an ABC/ABC-SMC inference layer. Table 2.1 and Algorithm 2.1 are retained because they provide the traceability between research questions, methodological components and Chapter 3 outputs.

Table 2.1. Methodological traceability matrix for the implemented thesis framework.

| Research question / objective | Methodological component | Dataset type | Chapter 3 evidence or output |
|---|---|---|---|
| RQ1: binary fake-news detection across text types and languages | BERT, RoBERTa, BigBird, mBERT, XLM- RoBERTa and classical baselines | Constraint@AAAI2021, FRN, ISOT, CoAID and Albanian Fake News Corpus | Accuracy, precision, recall, F1-score, ROC-AUC and error analysis |
| RQ2: temporal cascade representation of fake-news claims | Timestamp normalization, directed edges, $I_{\mathrm{new}}(t)$, $I_{\mathrm{cum}}(t)$ and cascade features | COVID-19 Twitter/X cascade, Hoaxy / OSoMeNet, Hoaxy-imported cascade and synthetic Albanian cascade | Cascade size, depth, breadth, time to peak and temporal-network graphs |

| | | | |
|---|---|---|---|
| RQ3: SIR versus SEIZ diffusion modelling | SIR baseline and SEIZ state mapping for active spreading, exposure and skepticism | Principal empirical fake-news cascades | Fitted spreading curves, compartment trajectories and model-comparison tables |
| RQ4: uncertainty-aware epidemic parameter inference | ABC and ABC-SMC posterior inference using observed cascade summaries | Empirical diffusion-oriented cascades | Posterior estimates, credible intervals, convergence and calibration summaries |
| RQ5: posterior prediction and intervention analysis | Posterior predictive simulation, short-horizon forecasting and counterfactual parameter perturbation | Empirical cascades with extracted time series | Forecast outputs, intervention scenarios and baseline diffusion-risk functional evidence |

**Algorithm 2.1. Transformer-ABC-Epidemic pipeline.**

Input: content-based fake-news datasets, diffusion-oriented cascade records and documented preprocessing rules.

**1.** Clean textual fields and harmonize labels into an auditable binary fake/real representation.

**2.** Tokenize content according to the selected Transformer family and evaluate the binary detection layer.

**3.** Extract or record the model-estimated fake-news probability $p_f(x)$ for eligible claims.

**4.** Preserve timestamps, user or source-target identifiers, cascade identifiers and action types for diffusion-oriented records.

**5.** Construct directed temporal cascades and compute $I_{\text{new}}(t)$ , $I_{\text{cum}}(t)$ and network summary statistics.

**6.** Fit SIR and SEIZ diffusion models to the empirical spreading curves.

**7.** Run ABC and ABC-SMC inference using observed cascade summaries and simulator-generated trajectories.

**8.** Generate posterior predictive curves, short-horizon forecasts and counterfactual intervention scenarios.

Output: classification evidence, cascade representation, epidemic parameter estimates, posterior uncertainty and intervention-oriented diffusion-risk analysis.

## 2.2 Dataset Selection Criteria

Dataset selection separates content-based evidence from diffusion-oriented evidence. Content datasets must provide text and fake/real labels for supervised detection, whereas diffusion datasets must provide timestamps, interactions and cascade structure for SIR/SEIZ and ABC modelling. This distinction prevents classification datasets from being interpreted as temporal-network evidence and supports reproducibility through documentation, ethics and access constraints.

For diffusion-oriented analysis, the datasets must provide timestamps, interactions, cascade membership and network structure. The retained diffusion layer uses the COVID-19 Twitter/X cascade, the Hoaxy / OSoMeNet political cascade, the Hoaxy-imported social cascade and the synthetic Albanian temporal cascade because they support time-windowed spreading curves, directed relations and the quantities required for SEIZ and ABC/ABC-SMC modelling [7], [8].

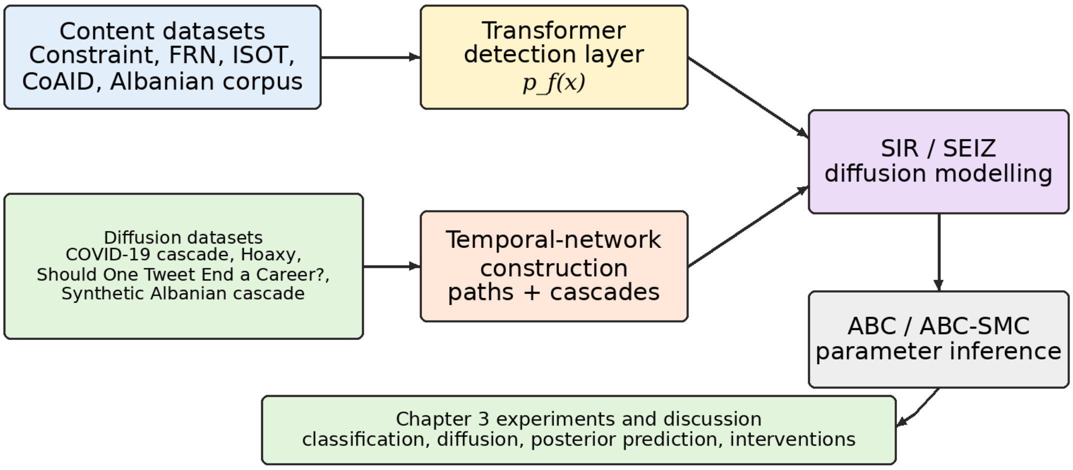



Figure 2.1. Dataset and methodology selection logic for the four-chapter Transformer-ABC-Epidemic framework.

Table 2.2. Dataset selection criteria and relevance to the proposed binary-focused thesis framework.

| **Criterion** | **Selection requirement** | **Role in binary detection** | **Role in diffusion / ABC** |
|---|---|---|---|
| Label and text quality | Reliable fake/real labels and accessible text, title, claim, post, or article content. | Required for supervised training and model comparison. | Defines which cascades are treated as misinformation or reliable content. |
| Binary class balance | Sufficient representation of fake and real classes, or a documented imbalance strategy. | Improves F1-score, precision, recall, and error analysis. | Avoids modelling only the largest or most visible fake-news cascades. |
| Domain, language, and length | Coverage of health, politics, general news, long articles, short posts, and low-resource languages. | Supports cross-domain, long-text, and multilingual Transformer evaluation. | Enables comparison of diffusion behaviour across misinformation domains. |
| Temporal information | Publication time, tweet time, retweet delay, reply time, or event chronology. | Useful for early-detection experiments. | Essential for cascade growth, time-to-peak, and posterior prediction. |
| Interaction and cascade structure | Retweets, replies, reposts, comments, propagation trees, or user-engagement records. | Can be used as metadata when labels and text are available. | Required for temporal networks, state transitions, and SIR/SEIZ calibration. |
| Ethics and reproducibility | Licensing, privacy constraints, platform policy, anonymization, and dataset documentation. | Ensures responsible use of labels and text. | Ensures responsible use of temporal-network and user-level data. |

$$Q(D) = \sum_k w_k C_k(D), \quad k \in \{\text{text, label, binary, time, network, ethics}\} \qquad (2.1)$$

In Equation (2.1), $Q$(D) represents the overall suitability score of dataset D. The term $C_{\text{text}}(D)$ measures textual availability and quality; $C_{\text{label}}(D)$ measures label reliability; $C_{\text{binary}}(D)$ measures compatibility with the binary fake/real task; $C_{\text{time}}(D)$ measures temporal information; $C_{\text{network}}(D)$ measures whether user interactions or cascades can be reconstructed; and $C_{\text{ethics}}(D)$ measures documentation, access, privacy, and responsible-use constraints. The weights can be adjusted depending on whether the experiment is primarily a Transformer detection experiment or a diffusion-inference experiment.

## 2.3 Principal Datasets

### 2.3.1 Content-Based Datasets for Transformer Detection

The content datasets support binary detection across different conditions: Constraint@AAAI2021 for short COVID-19 posts, FRN and ISOT for long-form or article-style settings, CoAID for health misinformation and the Albanian corpus for low-resource multilingual evaluation. Table 2.3 is retained as the compact dataset-role summary.

Table 2.3. Comparative overview of principal binary content-based datasets used in the thesis.

| Dataset | Key characteristics | Main use in this thesis |
|---|---|---|
| Constraint@AAAI2021 | COVID misinformation, short text | Binary fake-news detection in a crisis-information setting. |
| FRN | general fake/real classification | Long-text fake-news detection and BigBird evaluation. |
| ISOT | long-form news articles | Large-scale binary benchmarking and cross-domain robustness testing. |
| CoAID | COVID-related content | Healthcare misinformation detection and partial social-context analysis. |
| Albanian Corpus | low-resource language | Multilingual Transformer evaluation in a low-resource setting. |

### 2.3.2 Diffusion-Oriented Datasets for Temporal Cascade Modelling

Section 2.3.2 introduces the diffusion-oriented datasets used in Chapter 3: the COVID-19 Twitter/X cascade, the Hoaxy / OSoMeNet political cascade, the Hoaxy-imported social cascade and the synthetic Albanian cascade. These records provide the timestamps and directed interactions needed for temporal cascades, SIR/SEIZ inputs and ABC/ABC-SMC summaries; detailed metadata remain in Appendix D.

### 2.3.3 Summary of Dataset Roles

The dataset strategy deliberately assigns different roles to content and diffusion evidence. Content datasets train and evaluate the fake-news probability layer; diffusion datasets supply time-ordered spreading curves and network structure. This separation is a methodological safeguard against unsupported claims.

## 2.4 Data Preprocessing and Experimental Preparation

Preprocessing produces two coordinated outputs: content-ready records for Transformer classification and diffusion-ready records for temporal cascade modelling. Ambiguous labels are excluded from the main binary experiments when necessary, while augmented examples are restricted to training data to prevent leakage. Table 2.4 summarizes the retained preprocessing objectives.

Table 2.4. Preprocessing objectives for binary fake-news detection and temporal diffusion analysis.

| **Preprocessing objective** | **Relevant datasets** | **Methodological purpose** |
|---|---|---|
| Text reliability | Constraint@AAAI2021, FRN, ISOT, CoAID, Albanian Fake News Corpus | Remove textual noise while preserving credibility signals for binary classification. |
| Label consistency | All content-based datasets | Map heterogeneous labels into a common fake/real representation for Chapter 3 experiments. |
| Long-text readiness | FRN and ISOT | Prepare long-form articles for Transformer input, including BigBird or window-based strategies. |
| Low-resource normalization | Albanian Fake News Corpus | Preserve Albanian-specific characters, named entities and local vocabulary for multilingual Transformers. |
| Temporal integrity | Diffusion-oriented datasets | Preserve timestamps, event order and user-interaction structure for cascade construction. |
| Reproducibility and leakage prevention | All datasets | Document cleaning, split and augmentation rules so that experimental results can be reproduced. |

### 2.4.1 Text Cleaning and Normalization

Text cleaning is conservative: duplicates and empty fields are controlled, while URLs, hashtags, mentions and language-specific cues are preserved when they may carry credibility or diffusion meaning. This is especially important for short social-media texts and Albanian records, where excessive normalization can remove useful information [7], [8].

Table 2.5. Core text cleaning and normalization operations retained in the main methodology.

| **Operation** | **Treatment** | **Reason** |
|---|---|---|
| Duplicate and empty-text control | Remove exact duplicates and empty essential text fields. | Prevents inflated performance and invalid training records. |
| URL, mention and hashtag handling | Standardize rather than automatically delete when the token may carry credibility or diffusion meaning. | Preserves topic, source and platform signals. |
| Punctuation and casing normalization | Normalize excessive punctuation and casing while preserving sentence boundaries. | Improves tokenization without removing stylistic evidence. |
| Language-specific character preservation | Preserve Albanian characters such as ë and ç. | Maintains low-resource language information. |

### 2.4.2 Label Harmonization

Binary label harmonization maps heterogeneous dataset labels into a common fake/real target while preserving original labels as metadata. The equations and Table 2.6 document the mapping principle and support transparent interpretation of all later detection results.

$$y_{bin} = \begin{cases} 1, & \textit{if the item is fake-news-oriented,} \\ 0, & \textit{if the item is real or reliable.} \end{cases} \tag{2.2}$$

In Equation (2.2), $y_{\mathrm{bin}}$ denotes the harmonized binary target used in the Transformer detection experiments. The mapping does not erase the original dataset label; instead, the original label is preserved as raw-label metadata. This prevents a methodological error in which uncertain, auxiliary or diffusion-specific labels are treated as if they were identical to standard fake/real classification labels.

Table 2.6. Compact binary mapping examples retained in the main methodology.

| Dataset | Original labels | Binary mapping |
|---|---|---|
| Constraint@AAAI2021 | fake / real | fake -> 1; real -> 0 |
| ISOT | Fake.csv / True.csv | Fake.csv -> 1; True.csv -> 0 |
| Hoaxy / OSoMeNet | claim and fact_checking retweet records claim can be treated as misinformation-oriented; fact_checking as corrective/reliable in diffusion analysis | claim-> 1 ; fact_checking->0 |

### 2.4.3 Tokenization and Transformer Input Preparation

Tokenization converts cleaned text into model-ready input according to the selected Transformer family. BERT/mBERT, RoBERTa/XLM-R and BigBird use different subword and sequence policies; therefore, tokenizer choice is treated as part of the experimental design rather than a neutral technical detail.

$$\mathrm{Tokenizer}_m(x_i) \rightarrow (\mathrm{input\_ids}_i, \mathrm{attention\_mask}_i, \ \ldots) \tag{2.3}$$

In Equation (2.3), $\mathrm{Tokenizer}_m$ denotes the tokenizer associated with model family m, $x_i$ is the cleaned text record, $\mathrm{input\_ids}_i$ is the sequence of subword identifiers, and $\mathrm{attention\_mask}_i$ marks valid input positions. Model-specific technical details such as token_type_ids, dynamic padding, special token handling, [CLS], [SEP] and truncation variants are moved to Appendix F because they are implementation specifications rather than core methodological arguments.

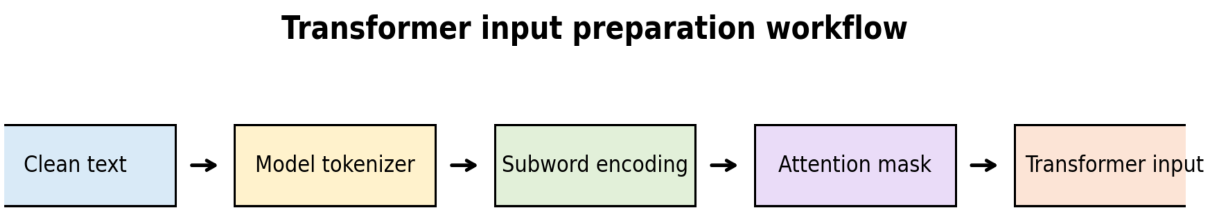



Figure 2.2. Transformer tokenization and model-input preparation workflow.

Table 2.7. Tokenization strategy by model family.

| Model family | Main tokenizer logic | Use in this dissertation |
|---|---|---|
| BERT / mBERT | WordPiece tokenization with fixed maximum sequence length. | Short and medium-length fake-news detection, including multilingual support. |
| RoBERTa / XLM-R | Byte-level BPE or SentencePiece-style subword modelling. | Robust subword handling for noisy or multilingual text. |
| BigBird | Sparse-attention long-sequence input. | Long-form article-style datasets such as |

### 2.4.4 Handling Noisy and Imbalanced Data

Noise and imbalance are handled through documented filtering, class-aware evaluation, optional class-weighted loss and controlled training-only augmentation. The retained equations and table define the imbalance ratio and weighted loss as methodological options, while the main results emphasize fake-class recall, macro-F1 and ROC-AUC.

$$R(D) = \frac{\max(n_{\text{fake}}, n_{\text{real}})}{\min(n_{\text{fake}}, n_{\text{real}})} \qquad (2.4)$$

Equation (2.4) defines the class-imbalance ratio for a binary dataset D, where $n_{\text{fake}}$ and $n_{\text{real}}$ denote the number of fake and real examples. A value close to 1 indicates approximate balance, while higher values indicate stronger imbalance. This ratio guides whether class weighting, controlled augmentation or imbalance-aware evaluation should be used.

$$\mathcal{L}_{\text{WBCE}} = -[w_1 y \log(p) + w_0(1-y)\log(1-p)] \qquad (2.5)$$

Equation (2.5) gives a weighted binary cross-entropy objective that can be used when imbalance is substantial. The weights $w_1$ and $w_0$ compensate for class-frequency differences, while p denotes the predicted fake-news probability. This objective is treated as a training option rather than as a mandatory choice for every dataset.

Table 2.8. Compact strategy set for noisy and imbalanced binary fake-news data.

| Strategy | Use |
|---|---|
| Class-weighted loss | Default option when class imbalance is visible. |
| Controlled augmentation | Training-support option for low-resource settings only. |
| Macro-F1 and PR-AUC evaluation | Evaluation options that reduce over-reliance on accuracy. |

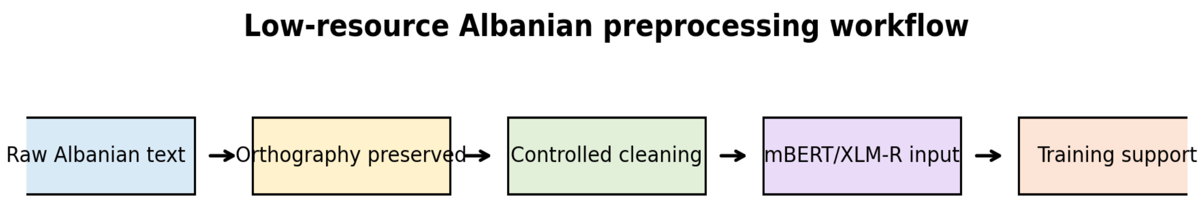



Figure 2.3. Low-resource Albanian fake-news preprocessing workflow for multilingual Transformer input.

Table 2.9. Compact Albanian preprocessing challenges and responses.

| Challenge | Recommended response | Reason |
|---|---|---|
| Language-specific characters | Preserve ë and ç; avoid forced ASCII conversion. | Prevents loss of Albanian orthographic information. |
| Local names and political vocabulary | Retain named entities and local terms. | Maintains domain-specific credibility cues. |
| Small dataset size | Use controlled augmentation only inside training data. | Supports training without altering test evidence. |

### 2.4.5 Preprocessing Workflow

For diffusion-oriented records, preprocessing is less aggressive than for text classification because timestamps, user identifiers, interaction fields and cascade membership must be preserved. The output is a quality-controlled event layer that becomes the input for temporal cascade construction, SIR/SEIZ fitting and ABC distance calculation.

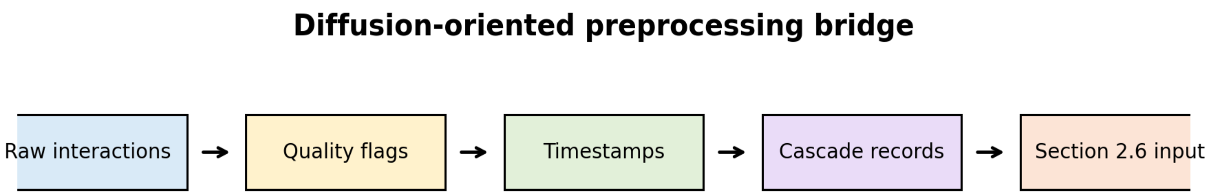



Figure 2.4. Preprocessing workflow for diffusion-oriented data before temporal cascade construction.

Table 2.10 is retained because it defines the minimum diffusion-ready output required by the later modelling stages. Dataset-specific diffusion schemas, quality-control decisions and state-proxy mappings are moved to the appendices to avoid repeating the same pipeline logic in both preprocessing and temporal cascade construction sections.

Table 2.10. Main objectives of preprocessing diffusion-oriented data.

| Objective | Methodological role |
|---|---|
| Preserve cascade identity | Ensures that each event remains attached to its fake-news claim or thread. |
| Preserve users and interactions | Allows directed source-target relations to be constructed later. |
| Preserve timestamps | Supports temporal ordering, $I_{\text{new}}(t)$, $I_{\text{cum}}(t)$, SIR/SEIZ fitting and ABC distance calculation. |
| Assign quality flags | Documents missing, duplicated or inconsistent records without silently changing the evidence. |

## 2.5 Temporal Cascade Construction and Network Representation

Section 2.5 defines the conversion from diffusion records to directed temporal cascades. Each event keeps source, target, timestamp, action type, cascade identifier and claim label. This representation preserves chronological information and allows observed spreading curves to be extracted for later epidemic and Bayesian modelling [5], [6].

Temporal-network representation is selected because the dissertation requires interpretable cascade reconstruction and epidemic-parameter inference. The available diffusion datasets support time-respecting paths, spreading curves, SIR/SEIZ state mapping and ABC/ABC-SMC inference, whereas temporal graph neural models require richer user-level graph histories and are therefore retained as future work.

$$e_k = (u_s, u_t, t_k, a_k, c_k, y_k) \tag{2.6}$$

In Equation (2.6), $u_s$ denotes the source or previous node, $u_t$ denotes the target or responding node, $t_k$ is the event time, $a_k$ is the observed action type, $c_k$ is the cascade or fake-news claim identifier, and $y_k$ is the available technical label such as claim, covid_fake_claim or fact_checking. This representation keeps content labels and temporal evidence connected without treating all records as equivalent to epidemic states.

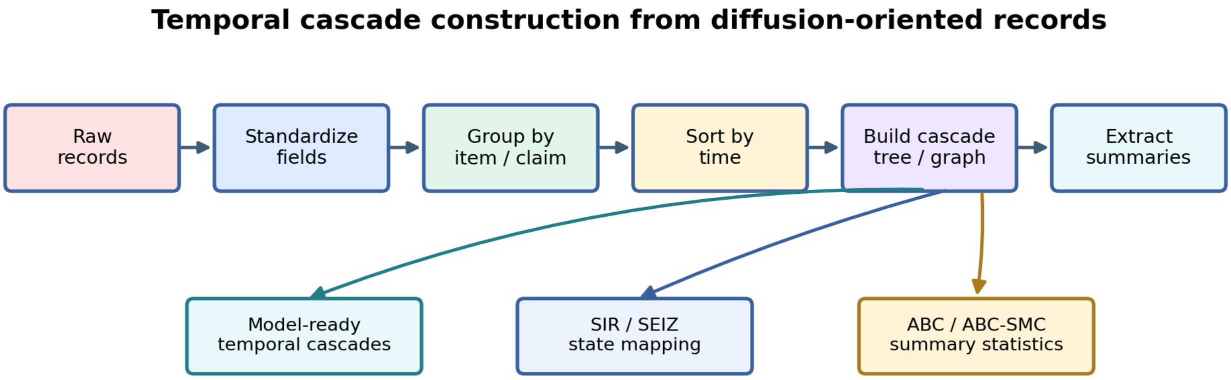



Figure 2.5. Temporal cascade construction workflow from diffusion-oriented records to model-ready cascades.

### 2.5.1 Cascade Preprocessing

Cascade preprocessing groups events by claim or source item, filters the retained fake-news cases and preserves fields needed for temporal reconstruction. The model-ready schema and cascade definition keep content probability connected with temporal evidence without treating every record as a direct psychological state.

$$C_c = (V_c, E^T_c, x_c, y_c, p_f(x_c)) \tag{2.7}$$

In Equation (2.7), $V_c$is the set of users or source nodes involved in the cascade, $E^T$ is the set of directed time-stamped interactions, $x_c$ is the associated textual fake-news claim or article, $y_c$ is the harmonized claim label, and $p_f(x_c)$ is the Transformer-based fake-news probability used as a content-risk signal in the broader framework.

Table 2.11. Core elements of the temporal-network representation.

| Element | Definition in this thesis | Example | Use in later modelling |
| --- | --- | --- | --- |
| Node | A user, post, source claim, article or reaction depending on the dataset. | Twitter/X user in Hoaxy; source tweet in a retweet cascade. | Defines the units that can become susceptible, exposed, spreading, recovered or skeptical in epidemic interpretation. |

| | | | |
|---|---|---|---|
| Directed edge | A time-ordered interaction from a source node to a target node. | from_user_id -> to_user_id in retweet data. | Represents possible information- transmission paths. |
| Timestamp | The time at which an interaction occurs or the delay from the source event. | tweet_created_at in Twitter/X or Hoaxy-style records. | Allows temporal ordering, time-to-peak analysis and time-respecting paths. |
| Weight | A value measuring interaction strength, frequency or relevance. | Number of retweets between two users or repeated interactions. | Supports weighted diffusion modelling and user-influence analysis. |
| Cascade identifier | A label linking all events belonging to the same fake- news claim or source item. | Claim ID, source tweet or external item label. | Groups events for cascade-level statistics and ABC summaryfeatures. |

Table 2.12. Unified model-ready schema for diffusion-oriented records.

| **Field** | **Definition** | **Required role in the thesis** |
|---|---|---|
| cascade_id | Identifier linking all records that belong to the same fake-news claim or source item. | Groups records before feature extraction and SIR/SEIZ preparation. |
| source | Originating user, source post, article or previous node. | Defines the starting side of a directed temporal edge. |
| target | Receiving, reposting, replying or interacting user/node. | Defines the activated or affected side of a directed temporal edge. |
| time | Timestamp or relative event time. | Allows chronological ordering and construction of time-respecting cascades. |
| action_type | Observed interaction type, such as retweet, reply, repost, claim or fact-checking record. | Supports interpretation of spreading, exposure or correction proxies. |
| claim_label | Technical dataset value such as claim, covid_fake_claim or fact_checking. | Maintains the distinction between fake-news claim rows and auxiliary/corrective rows. |
| text_or_title | Textual claim, article title or external-item description when available. | Links the temporal cascade to the Transformer-based content layer. |

### 2.5.2 Temporal Network Construction

Each claim-level event sequence is represented as a directed temporal network. Edges follow observable interaction direction and are valid only when timestamps respect chronological order. This prevents impossible static-graph paths and supports time-to-peak, reachability and cascade-depth analysis.

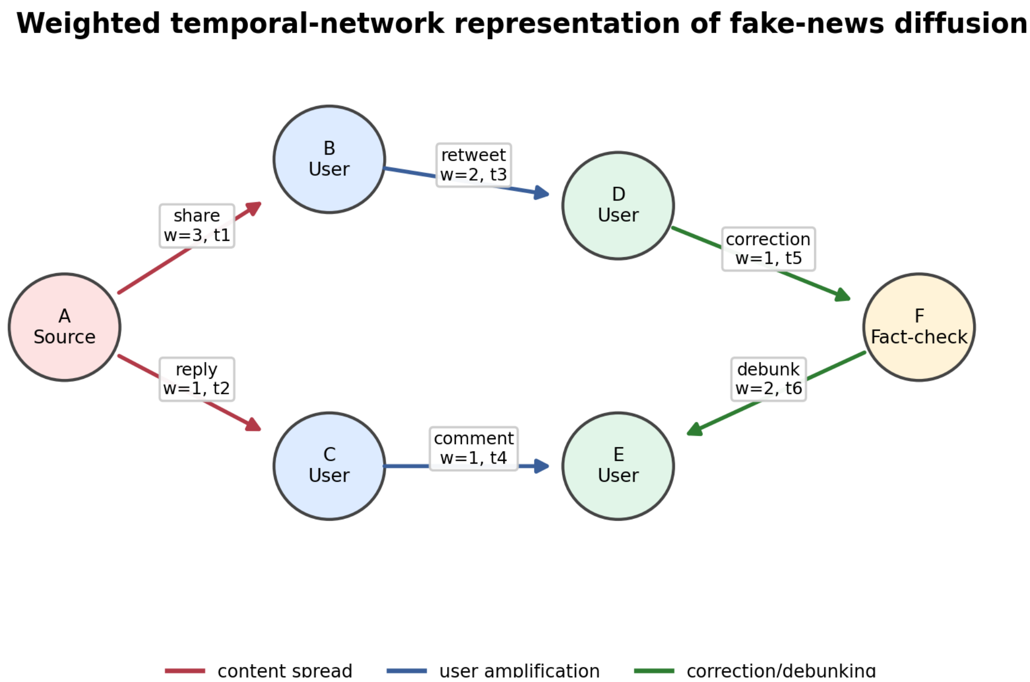



Figure 2.6. Illustration of nodes, directed edges, weights, and timestamps in a fake-news diffusion network.

$$P = [(u_0, u_1, t_1), \ldots, (u_{m-1}, u_m, t_m)], \qquad t_1 < t_2 < \cdots < t_m \tag{2.8}$$

In Equation (2.8), P denotes a time-respecting path; the claim can pass from one user to the next only when the event times increase.

Table 2.13. Comparison between static and temporal network representations.

| Aspect | Static network | Temporal network | Implication for this thesis |
|---|---|---|---|
| Edge interpretation | An edge exists if an interaction occurs at least once during the observation window. | An edge exists at a specific time or time interval. | Temporal networks are more suitable for cascade reconstruction and diffusion timing. |
| Path validity | All connected paths may appear valid after aggregation. | Only time-respecting paths are valid. | Prevents impossible information-flow paths. |
| Useful features | Degree, density, clustering and communities. | Inter-event time, time to peak and time-respecting reachability. | Temporal features support SIR/SEIZ simulation and ABC summary statistics. |
| Limitation | Loses the order of events. | Requires accurate timestamps and event ordering. | Temporal modelling is more demanding but more realistic for fake-news spread. |

$$s_{\text{out}}(i) = \sum_j \sum_t w_{ij}(t) \tag{2.9}$$

where $s_{\text{out}}(i)$ represents the weighted outgoing activity of user i and $w_{ij}(t)$ is the weight of the interaction from user i to user j at time t. This measure supports the identification of users who repeatedly amplify a fake-news claim through observable interactions.

$$R_{\text{cc}} = \frac{E_{\text{cross}}}{E_{\text{total}}} \tag{2.10}$$

where $E_{\text{cross}}$ is the number of cascade edges connecting different communities and $E_{\text{total}}$ is the total number of cascade edges. The ratio is used as a structural indicator of whether a fake-news claim remains localized or moves across communities. Centrality and community detection are treated as established network-analysis tools, while their detailed theoretical background is retained in the literature review and Appendix E [6].

### 2.5.3 Cascade Features for SIR/SEIZ/ABC

The main cascade features are new active spreaders, cumulative active spreaders, final cascade size, time to peak, growth rate and selected network descriptors. These features connect descriptive temporal analysis with the summary statistics used in SIR/SEIZ calibration and ABC/ABC-SMC inference.

The two central temporal activation measures are $I_{\text{new}}(t)$ and $I_{\text{cum}}(t)$. $I_{\text{new}}(t)$ is the number of new unique active spreaders observed in a given time window, while $I_{\text{cum}}(t)$ is the cumulative number of unique active spreaders observed up to that time. These two quantities form the direct bridge between temporal-network construction and epidemic-model fitting. This bridge is important because ABC and ABC-SMC later compare simulated and observed spreading curves through summary statistics rather than through a closed-form likelihood.

Table 2.14. Cascade features for SIR/SEIZ/ABC summary extraction.

| **Feature** | **Definition** | **Use in SIR/SEIZ/ABC** |
|---|---|---|
| Cascade size | Total number of participating nodes or unique active spreaders. | Used to compare final epidemic size and posterior predictive saturation. |
| $I_{\text{new}}(t)$ | Number of newly active unique spreaders in time window t. | Used as the observed infected-time increment for diffusion dynamics. |
| $I_{\text{cum}}(t)$ | Cumulative number of unique active spreaders up to time t. | Used as the main cumulative curve for model fitting and forecasting. |
| Time to peak | Time between cascade start and maximum $I_{\text{new}}(t)$. | Supports ABC summary statistics and timing comparison. |
| Inter-event time | Time gaps between consecutive events in the cascade. | Describes burstiness and delayed activation. |
| Reach and breadth | Number of users or branches reached by the cascade. | Supports descriptive comparison and posterior predictive checking. |
| Weighted outgoing activity | Temporal sum of outgoing edge weights for each user. | Identifies influential spreaders and possible relay hubs. |
| Cross-community ratio | Share of edges crossing community boundaries. | Summarizes structural spread beyond the initial community. |

$$G_c^T = (V_c, E_c^T) \tag{2.11}$$

In Equation (2.11), denotes the directed temporal cascade network of claim c. Its nodes are the users or source objects in the cascade, while its edges are directed timestamped interactions, optionally weighted. The superscript T indicates temporal ordering. This object provides the basis for, cascade-level features and SIR/SEIZ inputs [5], [6].

The ordered activations of target users generate $I_{\text{new}}(t)$ and $I_{\text{cum}}(t)$ , while network properties such as weighted outgoing activity and community crossing provide optional structural summaries. At this stage, retweets, replies, reposts and fact-checking records are not interpreted as direct observations of latent belief. They are observable proxies that support modelling assumptions and posterior comparison.

The methodological boundary of Section 2.5 is intentionally narrow. The dissertation does not implement memory-dependent temporal-network models as empirical components. Instead, it uses directed temporal cascades because they are supported by the retained datasets and connect directly to the experimental pipeline: temporal-network visualization, $I_{\text{new}}(t)$, $I_{\text{cum}}(t)$, SIR/SEIZ fitting, ABC/ABC-SMC posterior inference and short-horizon predictive evaluation. More complex path-memory approaches remain possible future extensions rather than unvalidated components of the implemented methodology [5].

## 2.6 SIR/SEIZ, ABC/ABC-SMC and Posterior Prediction Methodology

Section 2.6 defines the modelling core of the thesis. SIR provides a transparent baseline, SEIZ models exposure and skepticism, ABC estimates uncertainty without an explicit likelihood and ABC-SMC progressively improves posterior concentration. Posterior prediction and intervention scenarios are then derived from the calibrated SEIZ layer.

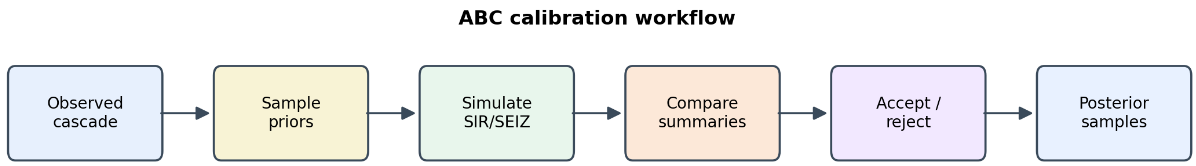



Figure 2.7. ABC workflow for calibrating fake-news epidemic parameters from temporal cascades.

### 2.6.1 SIR Baseline Model

SIR maps fake-news diffusion into susceptible users, active spreaders and recovered or inactive users. It is retained because it is interpretable and provides a baseline trajectory against which the richer SEIZ model can be evaluated.

$$\theta_{\text{SIR}} = (\beta, \gamma, I_0) \tag{2.12}$$

### 2.6.2 SEIZ Main Diffusion Model

SEIZ is the main diffusion model because it includes exposed and skeptic or non-spreading states. These states are latent modelling proxies for delayed adoption, hesitation, correction and non-engagement, not direct measurements of user belief.

$$\theta_{\text{SEIZ}} = (\beta, b, p, l, \rho, \varepsilon, \gamma, \delta, S_0, E_0, I_0, Z_0) \tag{2.13}$$

Equations (2.12) and (2.13) define the SIR and SEIZ parameter vectors; SEIZ extends SIR with exposure, skepticism, correction and initial-state terms.

**2.6.3 ABC Inference**

Let $D_{\text{obs}}$ denote the observed fake-news cascade and let θ denote the epidemic parameter vector. In a standard Bayesian formulation, the posterior distribution is proportional to the product of the likelihood and the prior:

$$p(\theta \mid D_{\text{obs}}) \propto p(D_{\text{obs}} \mid \theta)p(\theta) \tag{2.14}$$

The difficulty in Equation (2.14) is the likelihood term $p(D_{\text{obs}} \mid \theta)$. For a fake-news cascade, this term would require the probability of the observed sequence of user activations, inter-event times, reposts, corrections and hidden non-responses. Because these elements are not fully observed and are not easily represented by a closed-form probability model, ABC replaces likelihood evaluation with simulation and summary-statistic comparison [4].

$$p_{\text{ABC}}(\theta \mid D_{\text{obs}}) \propto p(\theta) \int \mathbf{1}\{\rho(S_{\text{sim}}, S_{\text{obs}}) \leq \varepsilon\} p(D_{\text{sim}} \mid \theta)\, dD_{\text{sim}} \tag{2.15}$$

In Equation (2.15), $D_{\text{sim}}$ is a simulated cascade generated by the SIR or SEIZ simulator, S(D) is a vector of summary statistics, ρ is a distance function and ε is the ABC tolerance threshold. A simulated parameter value is accepted when the distance between simulated and observed summaries is sufficiently small. The accepted parameter values approximate the posterior distribution over epidemic parameters that can reproduce the observed fake-news spreading pattern.

Table 2.15. Candidate prior distributions for ABC-based inference of fake-news epidemic parameters.

| Parameter | Model | Prior example | Interpretation in fake-news propagation |
|---|---|---|---|
| β | SIR/SEIZ | Uniform(0.01, 0.80) | Baseline rate at which active spreaders influence susceptible users. |
| γ | SIR/SEIZ | Uniform(0.01, 0.50) | Rate at which spreaders stop sharing, forget the content or lose interest. |
| ε | SEIZ | Uniform(0.01, 0.50) | Rate at which exposed users become active spreaders. |
| p | SEIZ | Beta(2, 2) | Probability of immediate sharing after contact with an active spreader. |
| l | SEIZ | Beta(2, 2) | Probability of becoming skeptical or non-spreading after exposure. |
| δ | SEIZ-modified | Uniform(0.00, 0.30) | Correction or debunking rate caused by warnings, fact-checks or counter-information. |
| ρ | SEIZ | Uniform(0.00, 0.60) | Interaction coefficient between exposed and infected users. |

ABC inference samples candidate parameters, simulates cascades and accepts parameter values that reproduce the observed cascade summaries within a tolerance. This allows the thesis to estimate plausible transmission, recovery, exposure and skepticism regions when the exact cascade likelihood is unavailable [4].

$$S(D) = [C_T, t_{\text{peak}}, r_{\max}, d_c, b_c, \tau_{\text{mean}}, \bar{p}_f] \tag{2.16}$$

In Equation (2.16), $C_T$ is final cascade size, $t_{\text{peak}}$ is time to peak, $r_{\max}$ is the maximum spreading rate, $d_c$ is cascade depth, $b_c$ is cascade breadth, $\tau_{\text{mean}}$ is the mean inter-event time and $\bar{p}_f$ is the mean Transformer-derived fake-news probability for the claim or cascade item. The distance between observed and simulated summaries is expressed as:

$$\rho(S_{\text{obs}}, S_{\text{sim}}) = \sum_k w_k \, |S_k^{\text{obs}} - S_k^{\text{sim}}| \tag{2.17}$$

The weights $w_k$ in Equation (2.17) determine the importance of each summary statistic. If final cascade size is the primary target, $C_T$ receives more weight. If early-warning behaviour is more important, $t_{\text{peak}}$ and $r_{\max}$ can be emphasized. In Chapter 3, these summary statistics support posterior calibration, posterior predictive simulation and short-horizon out-of-sample forecasting. The full candidate summary-statistic table is provided in Appendix D.3.

$$p_f(x_i) = P(y_i = \text{fake} \mid x_i) \tag{2.18}$$

In Equation (2.18), $p_f(x_i)$ denotes the probability assigned by the Transformer detector that textual item $x_i$ belongs to the fake-news class. The score is not used as an ensemble component and does not replace temporal diffusion modelling. Instead, it provides a calibrated content-risk signal that can modulate the spreading pressure assigned to an observed temporal interaction.

$$\beta_{ij}(t) = \beta_0 \, w_{ij}(t) \, [1 + \lambda p_f(x_i)] \, [1 + \eta c_j] \tag{2.19}$$

In Equation (2.19), the effective transmission rate is adjusted by edge weight, Transformer fake-news probability and local network influence.

**2.6.4 ABC-SMC Inference**

ABC-SMC improves rejection ABC by using sequential particle populations and decreasing tolerances. The method is suitable for fake-news cascades because it concentrates posterior mass around parameter regions that better reproduce observed temporal growth while still reporting uncertainty.

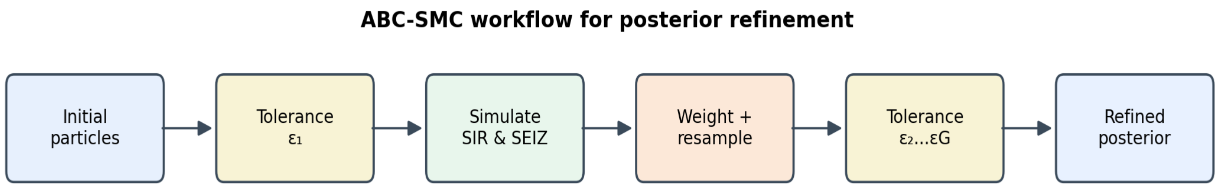



Figure 2.8. ABC-SMC workflow for comparing SIR and SEIZ on fake-news temporal cascades.

$$\mathcal{M} = \{\mathrm{SIR}, \mathrm{SEIZ}\} \tag{2.20}$$

Equation (2.20) defines the two candidate diffusion models compared under the same observed cascade evidence.

$$P_{\mathrm{ABC}}(M \mid D_{\mathrm{obs}}) \propto P(M) P_{\mathrm{ABC}}(D_{\mathrm{obs}} \mid M) \tag{2.21}$$

Equation (2.21) is not interpreted as proof that one model is universally superior. Instead, the dissertation evaluates whether the SEIZ structure provides better posterior predictive calibration for the empirical fake-news cascades considered in Chapter 3. The accepted distances, posterior parameter distributions and posterior predictive curves are used to compare the practical calibration of SIR and SEIZ under the same observed $I_{\mathrm{new}}(t)$ and $I_{\mathrm{cum}}(t)$ trajectories.

### 2.6.5 Posterior Predictive Simulation and Forecasting

Posterior predictive simulation propagates accepted parameter samples forward to estimate short-horizon cascade growth. The output is not a deterministic forecast, but an interval-based representation of plausible future diffusion under the fitted model assumptions.

$$D_{\mathrm{future}}^{(m)} \sim M\big(\theta^{(m)}\big), \quad \theta^{(m)} \sim p_{\mathrm{ABC}}(\theta \mid D_{\mathrm{obs}}) \tag{2.22}$$

In Equation (2.22), each future trajectory is generated by drawing an accepted posterior parameter vector and simulating the diffusion model forward over the forecast horizon.

$$\beta_{\mathrm{intervention}} = \alpha_\beta \beta, \quad 0 < \alpha_\beta < 1; \qquad \delta_{\mathrm{intervention}} = \alpha_\delta \delta, \quad \alpha_\delta > 1 \tag{2.23}$$

Equation (2.23) represents intervention scenarios: visibility reduction lowers beta, while stronger debunking or correction increases delta.

### 2.6.6 Baseline Posterior Fake-News Diffusion-Risk Functional

The baseline diffusion-risk functional combines Transformer fake-news probability, temporal-network amplification and ABC-SMC-SEIZ posterior predictive growth. Its purpose is synthesis and auditability: it makes the link between content risk, cascade activity and expected future spread explicit.

For a fake-news claim c, textual input $x_c$, observed cascade $\mathcal{S}_c^{\mathrm{obs}}$, prediction horizon h and observed cascade population size $N_c$, the baseline risk is defined as:

$$\mathcal{R}_{\mathrm{base}}(c,h) = \hat{p}_f(x_c)[1 + \lambda A_c(t)] \mathbb{E}_{\theta \sim \pi_\varepsilon^{\mathrm{SMC}}} \left[\frac{\Delta I_{\mathrm{cum},c}(t,h;\theta)}{N_c}\right] \tag{2.24}$$

In Equation (2.24), () is the Transformer fake-news probability, is the temporal-network amplification score, λ controls its influence, and (θ | ) is the ABC-SMC posterior over SEIZ parameters. The functional measures expected normalized future cumulative growth, scaled by content risk and network amplification.

$$A_c(t) = \omega_1 I_{\mathrm{new},c}^{\mathrm{norm}}(t) + \omega_2 H_c^{\mathrm{norm}}(t) + \omega_3 B_c^{\mathrm{norm}}(t), \quad \omega_1 + \omega_2 + \omega_3 = 1 \tag{2.25}$$

In Equation (2.25), $I_{\text{new}}$,c^norm(t) is the normalized number of new unique active spreaders in the current time window, $H_c^{\text{norm}}(t)$ is a normalized hub-activity score, and $B_c^{\text{norm}}(t)$ is a normalized relay or betweenness-based activity score. When a minimal operational score is required, $A_c(t)$ may be approximated using only the normalized $I_{\text{new}}(t)$ signal. This keeps the functional compatible with the empirical cascades in Chapter 3, where $I_{\text{new}}(t)$, $I_{\text{cum}}(t)$, source hubs and relay users are the observable diffusion outputs.

If ABC-SMC returns M posterior particles $\{(\theta_m, w_m)\}$ for m = 1...,M, Equation (2.24) can be evaluated in particle form as:

$$\hat{\mathcal{R}}_{\text{base}}(c,h) = \hat{p}_f(x_c)[1 + \lambda A_c(t)] \sum_m w_m \left[\frac{\Delta I_{\text{cum},c}(t,h;\theta_m)}{N_c}\right] \tag{2.26}$$

Equation (2.26) evaluates the risk functional with ABC-SMC particles by averaging predicted growth across posterior samples and their weights.

**2.6.7 Framework Integration**

Framework integration connects the four layers into one reproducible pipeline. The content layer identifies suspicious claims, the temporal layer reconstructs the observed spreading process, the epidemic layer interprets state transitions and the Bayesian layer quantifies parameter uncertainty and posterior predictive risk.

$$d_{L2}(I_{\text{obs}}, I_{\text{sim}}) = \frac{\|I_{\text{sim}}(t) - I_{\text{obs}}(t)\|_2}{\|I_{\text{obs}}(t)\|_2} \tag{2.27}$$

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(I_{\text{sim}}(t) - I_{\text{obs}}(t)\right)^2} \tag{2.28}$$

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T}|I_{\text{sim}}(t) - I_{\text{obs}}(t)| \tag{2.29}$$

$$\text{nRMSE} = \frac{\text{RMSE}}{I_{\text{obs}}(T)} \tag{2.30}$$

Equations (2.27)-(2.30) define the error measures used to compare observed and simulated spreading curves: relative L2 distance, RMSE, MAE and normalized RMSE.

## 2.7 Dataset Suitability, Ethics and Reproducibility Boundaries

Section 2.7 defines the validity boundaries of the framework. Dataset suitability is evaluated separately for detection and diffusion tasks, while ethical handling requires anonymization, platform-policy compliance, no inference of private belief and clear separation between empirical observation and counterfactual simulation.

### 2.7.1 Suitability for Transformer-Based Fake-News Detection

A dataset is suitable for Transformer detection when it contains reliable labels, usable text, adequate class representation and documented splits. The suitability score is used as a transparent rubric rather than as a replacement for expert judgement.

For systematic comparison, the suitability of a content-based dataset $D_i$ for Transformer detection can be expressed qualitatively as:

$$S_T(D_i) = w_L L_i + w_T T_i + w_B B_i + w_Q Q_i + w_N N_i \tag{2.31}$$

In Equation (2.31), $S_T(D_i)$ denotes the Transformer-detection suitability of dataset $D_i$. The term $L_i$ represents label reliability, $T_i$ represents text availability and document length suitability, $B_i$ represents binary class balance, $Q_i$ represents text quality after preprocessing, and $N_i$ represents language or domain suitability. The weights $w_L$, $w_T$ , $w_B$, $w_Q$ and $w_N$ are researcher-defined qualitative weights. The expression is not used as a strict numerical score; it formalizes the criteria used when deciding which datasets should remain in the operational content-based layer.

Table 2.16. Final content-based datasets retained for Transformer and classical machine-learning detection.

| Dataset | Primary detection role | Reason for retention | Methodological boundary |
|---|---|---|---|
| Constraint@AAAI2021 | Short COVID-related fake/real text detection. | Supports short social-media-style misinformation classification under pandemic-information conditions. | Used for content detection, not for temporal diffusion modelling. |
| FRN | Long-form fake/real article detection. | Provides long news articles and supports long-sequence Transformer evaluation, including BigBird-style modelling [3]. | Used for article-level detection; no cascade reconstruction is inferred from the corpus alone. |
| ISOT | Large English fake/real article benchmark. | Provides a large binary benchmark for classical and Transformer comparison. | Used for content-level robustness analysis only. |
| CoAID | Healthcare misinformation detection. | Connects fake-news detection with health-related misinformation and partial social-context evidence. | Used mainly as a content dataset; social signals are not treated as complete temporal cascades. |
| Albanian Fake News Corpus | Low-resource Albanian fake-news detection. | Supports multilingual and low-resource Transformer evaluation. | Used for detection and synthetic validation support, not as empirical Twitter/X or Hoaxyevidence. |

The practical output of this suitability decision is a controlled content-risk layer. Transformer models trained or evaluated on the selected content datasets can produce a fake-news probability score $p_f(x)$, but this score is only one component of the full thesis framework. It becomes operationally meaningful for diffusion analysis only when it is linked to a fake-news claim that also has temporal-interaction evidence.

### 2.7.2 Suitability for Temporal Diffusion Modelling, SIR/SEIZ and ABC/ABC-SMC

A diffusion dataset must contain timestamps, cascade identifiers, directed interactions and enough temporal volume to support summary statistics. This explains why only the principal empirical cascades are used for posterior inference, while smaller or synthetic cases remain supplementary.

$$S_D(D_i) = w_T T_i + w_C C_i + w_E E_i + w_M M_i + w_A A_i + w_Q Q_i \quad (2.32)$$

In Equation (2.32), $S_D(D_i)$ denotes the temporal-diffusion suitability of dataset $D_i$. The term $T_i$ represents timestamp availability, $C_i$ represents cascade or claim-level structure, $E_i$ represents directed edge availability, $M_i$ represents compatibility with SIR/SEIZ state mapping, $A_i$ represents availability of ABC/ABC-SMC summary statistics, and $Q_i$ represents data quality after temporal preprocessing. As with Equation (2.31), the expression is a methodological guide rather than a fixed numerical rule.

Table 2.17. Final diffusion-oriented datasets retained for temporal, epidemic and ABC/ABC-SMC modelling.

| **Dataset / cascade** | **Fake-news claim or item** | **Main diffusion support** | **Role** |
|---|---|---|---|
| COVID-19 real Twitter/X | “The coronavirus is not only affecting the way we live, it’s also dramatically affecting the way we die.” | Timestamped user-to-user interactions filtered to site_type = covid_fake_claim. | Principal empirical health misinformation cascade for temporal-network, SEIZ/ABC-SMC and $I_{\text{cum}}(t)$ forecasting. |
| Hoaxy / OSoMeNet political | “Three million votes in the Presidential Election were cast by illegal aliens.” | Large-scale retweet cascade with directed user-to-user interactions. | Principal empirical political fake-news claim cascade for SIR/SEIZ, ABC/ABC-SMC and forecast comparison. |
| Hoaxy-imported social cascade | “Should One Tweet End a Career?” | Claim-labelled Hoaxy social cascade with timestamped spreadingrecords. | Additional comparative social fake-news claim cascade for temporal-network interpretation. |
| Synthetic Albanian temporal cascade | “Shqipëria do të presë 100 mijë palestinezë nga Gaza.” | Generated timestamped cascade with claim and fact-checking records. | Validation and simulation support for the low-resource Albanian pipeline; not treated as empirical Twitter/X evidence. |

$$S(D_i) = (C_T, t_{\text{peak}}, r_{\max}, d_c, b_c, q_f, n_Z) \tag{2.33}$$

In Equation (2.33), the summary vector contains cascade size, time to peak, maximum spreading rate, depth, breadth, Transformer fake-news probability and available skeptical or corrective events. Its role is to translate an observed fake-news cascade into evidence comparable with simulated SIR/SEIZ trajectories in ABC and ABC-SMC [4], [7], [8].

**2.7.3 Ethical, Privacy and Reproducibility Considerations**

The framework uses observable public or documented interaction records responsibly. User identifiers are handled according to dataset constraints, and model outputs are not interpreted as private belief, persuasion or intention.

**2.7.4 Methodological Assumptions and Validity Boundaries**

The main assumptions are that Transformer probabilities are model-estimated risk scores, cascade interactions are observable proxies, SEIZ states are latent modelling constructs and intervention scenarios are simulations. These boundaries make the results defensible rather than overclaimed.

Table 2.18. Implemented components and literature-only or future-work boundaries.

| Component | Status in this thesis | Reason for status |
|---|---|---|
| BERT, RoBERTa, BigBird, mBERT and XLM-RoBERTa | Implemented and evaluated | They provide the operational Transformer detection layer and generate the model-estimated fake-news probability used by the framework. |
| SIR and SEIZ diffusion models | Implemented and evaluated | They provide interpretable epidemic-state representations for active spreading, stopping, exposure and skepticism. |
| ABC and ABC-SMC | Implemented and evaluated | They quantify uncertainty in diffusion parameters when exact likelihood evaluation is not feasible. |
| GNN and temporal graph learning | Literature review and future work | They require richer user-level graph data, reliable follower or interaction networks and consistent propagation labels than are available across all retained datasets. |
| Explainable AI | Future explanatory extension | LIME, SHAP and explainable fake-news systems are relevant, but a full XAI layer would require a separate explanation-validation protocol. |
| Large Language Models | Auxiliary or support role only | LLMs can assist augmentation or analysis, but final labels, evaluation sets and diffusion interpretations are not generated from LLM output. |
| Voting, stacking and classifier ensembles | Literature review only | They may improve accuracy but would reduce traceability between the Transformer score and the downstream diffusion-risk interpretation. |

### 2.8 Chapter Summary

Chapter 2 translates the literature review into an operational methodology. It defines dataset roles, preprocessing, temporal cascades, SIR/SEIZ modelling, ABC/ABC-SMC inference, posterior prediction, intervention scenarios and ethical boundaries. Chapter 3 then evaluates this pipeline empirically.

# Chapter 3. Experimental Results and Discussion

### 3.1 Introduction

Chapter 3 reports the empirical execution of the framework. It first evaluates Transformer-based binary detection, then reconstructs selected fake-news cascades, fits SIR and SEIZ models, estimates posterior uncertainty with ABC/ABC-SMC and evaluates counterfactual intervention scenarios. The chapter keeps content-level and diffusion-level evidence separate before integrating them in the final discussion.

### 3.2 Experimental Setup

The experimental setup follows the detection-diffusion-inference sequence. Content datasets are evaluated at item level using accuracy, precision, recall, F1-score and ROC-AUC; diffusion datasets are evaluated at cascade level using time-ordered spreading curves, SIR/SEIZ fit and posterior predictive behaviour. The two principal empirical cascades are the COVID-19 Twitter/X case and the Hoaxy / OSoMeNet political claim case.

Table 3.1. Experimental evidence and datasets.

| Block | Datasets | Output |
|---|---|---|
| **Detection** | Constraint, FRN, ISOT, CoAID, Albanian | F1, ROC |
| **Temporal** | Twitter COVID, Hoaxy | Cascades |
| **Epidemic** | Twitter COVID, Hoaxy | SIR/SEIZ |
| **Bayesian** | Twitter COVID, Hoaxy | Posterior |

### 3.3 Transformer-Based Binary Detection Results and Comparative Analysis

Section 3.3 reports the detection layer. The objective is not to rank all possible fake-news detectors, but to show how selected Transformer architectures behave across short-text, long-document, healthcare and low-resource Albanian datasets. The results provide fake-class probability outputs for the later cascade gate.

#### 3.3.1 Results on Content-Based Datasets

Table 3.2 shows that, on Constraint@AAAI2021 Task-A, BERT achieved the strongest short-text COVID-19 misinformation result, with 98.11% accuracy and 98.00% F1-score. The confusion matrix and ROC curve support the conclusion that contextual representations improve over strong classical baselines for this task.

Table 3.2. Author-generated detection performance on Constraint@AAAI2021 Task-A.

| Model / method | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| **CNN + BiLSTM** | 92.01% | 92.01% | 92.01% | 92.01% |
| **SVM + TF-IDF** | 94.35% | 94.42% | 94.39% | 94.39% |
| **SVM + Word2Vec (ED = 200)** | 92.66% | 92.67% | 92.66% | 92.66% |
| **SVM + Word2Vec (ED = 150)** | 92.94% | 92.94% | 92.94% | 92.94% |
| **BERT** | 98.11% | 98.11% | 98.00% | 98.00% |
| **Logistic Regression (LR)** | 92.07% | 92.11% | 92.00% | 92.00% |
| **Support Vector Machine (SVM)** | 91.09% | 91.13% | 91.00% | 91.00% |
| **Naive Bayes (NB)** | 93.01% | 93.06% | 93.00% | 93.00% |
| **NB + LR** | 93.01% | 93.05% | 93.00% | 93.00% |

Figure 3.1 reports the final BERT diagnostics for Constraint@AAAI2021 Task-A. The confusion matrix shows only 18 false positives and 22 false negatives, while the ROC curve reaches ROC-AUC = 0.9943. The short-text detector therefore provides a reliable content- risk signal before cascade analysis.

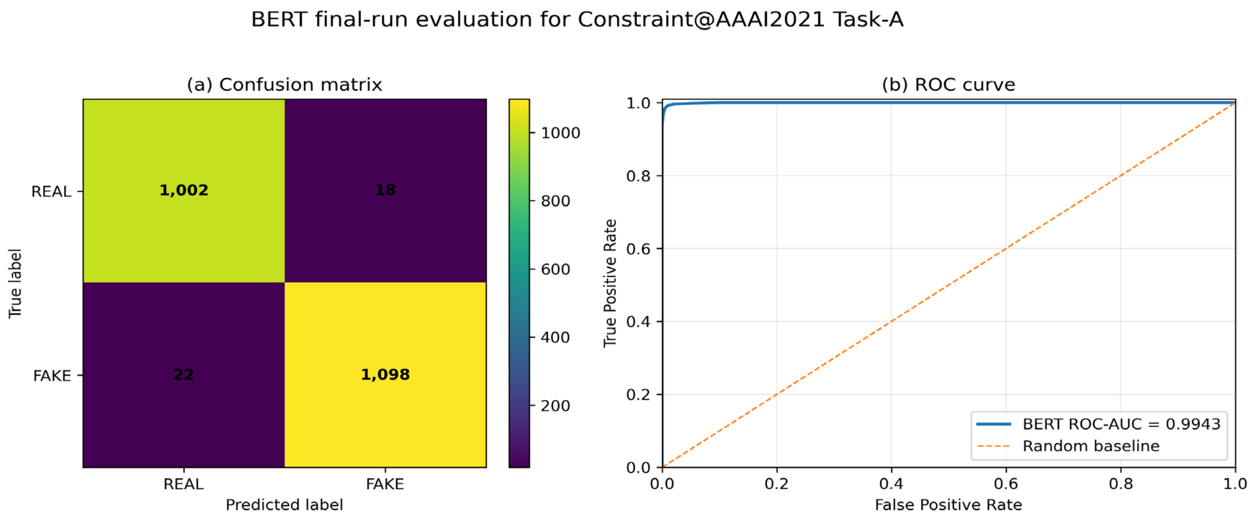



Figure 3.1. BERT final-run evaluation for Constraint@AAAI2021 Task-A: (a) confusion matrix; (b) ROC curve.

### 3.3.2 Long-Text and Domain-Specific Detection Results

For long and domain-specific datasets, Transformer performance depends on text length and topic. BigBird-RoBERTa was strongest on FRN and CoAID longer news, RoBERTa performed best on ISOT and short healthcare claims, and classical TF-IDF baselines remained useful reference points. The figures and tables are retained because they show ranking quality and error distribution.

Table 3.3. Compact comparison of strongest classical baselines across FRN, ISOT and CoAID.

| **Dataset / subset** | **Best classical baseline** | **Fake F1** | **Macro-F1** | **ROC-AUC** |
|---|---|---|---|---|
| **FRN** | Linear SVM + TF-IDF | 0.9531 | 0.9526 | 0.9903 |
| **ISOT** | Linear SVM + TF-IDF | 0.9924 | 0.9921 | 0.9996 |
| **CoAID social posts /claims** | Logistic Regression + TF-IDF | 0.9358 | 0.9488 | 0.9830 |
| **CoAID longer news** | Logistic Regression + TF-IDF | 0.4615 | 0.7243 | 0.9328 |

Table 3.4. Comparative Transformer detection performance across FRN, ISOT and CoAID.

| **Dataset / subset** | **Text type** | **Best Transformer setting** | **Fake F1** | **Macro-F1** | **ROC-AUC** |
|---|---|---|---|---|---|
| FRN | Long political/general articles | BigBird-RoBERTa-base | 0.9691 | 0.9692 | 0.9946 |
| ISOT | Large English article benchmark | RoBERTa title-body | 0.9948 | 0.9945 | 0.99997 |
| CoAID social posts / claims | Short healthcare claims/posts | RoBERTa-base | 0.9615 | 0.9701 | 0.9915 |
| CoAID longer news | Longer healthcare news items | BigBird-RoBERTa-base | 0.8205 | 0.9081 | 0.9786 |

Figure 3.2 confirms strong class separation for the FRN BigBird model. Only 18 real articles are incorrectly flagged as fake and 21 fake articles are missed. This error pattern indicates strong class separation and a balanced trade-off between fake-class recall and false-positive control, which is important when the detection output is later interpreted as a content-risk signal.

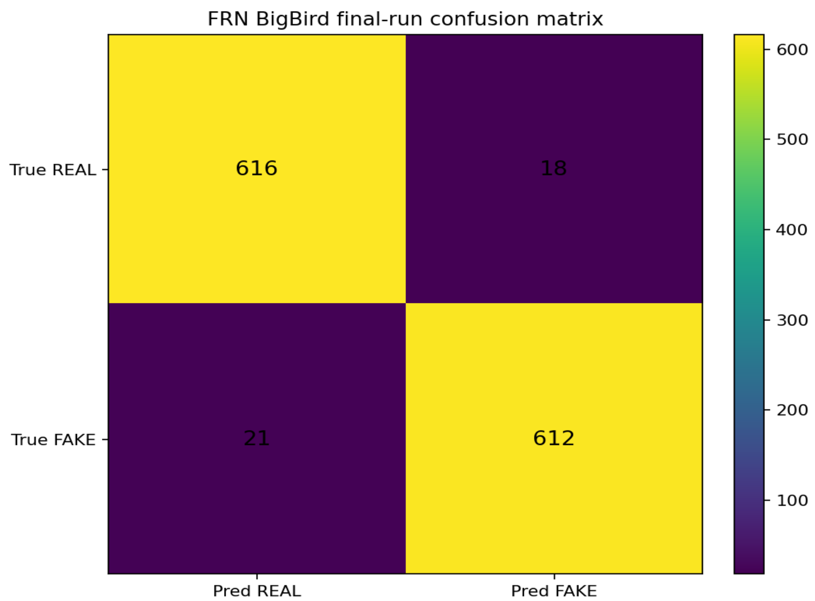



Figure 3.2. BigBird final-run confusion matrix for FRN.

Figure 3.3 shows that the best ISOT configuration, RoBERTa title-body, achieves near- perfect ranking separation (ROC-AUC = 0.99997) and a low error profile, with 22 real articles and 27 fake articles misclassified. The result is strong but should still be read with caution because ISOT contains dataset-specific regularities.

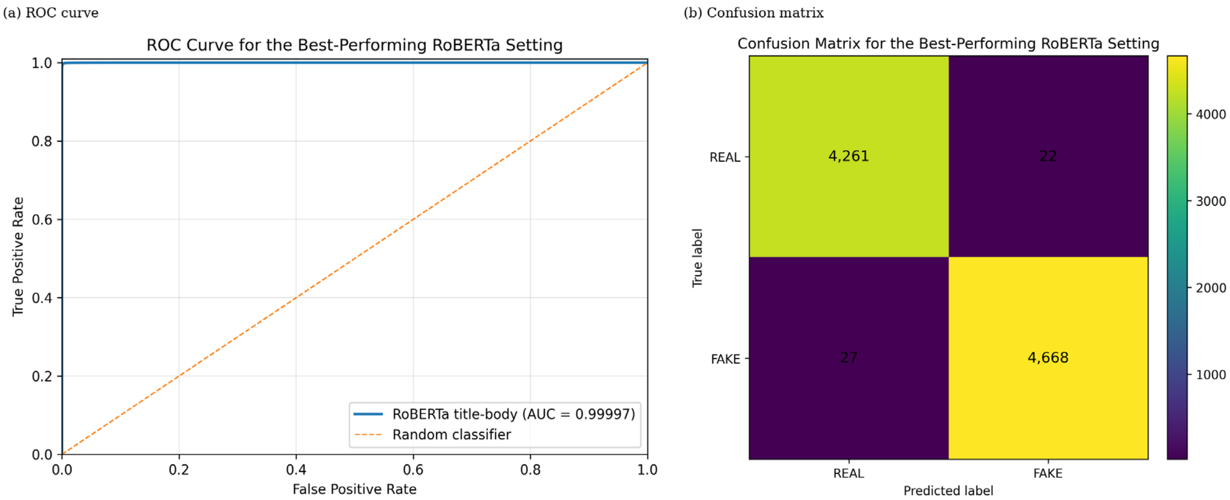


Figure 3.3. Complementary evaluation plots for the best-performing ISOT setting: (a) ROC curve; (b) confusion matrix.

Figure 3.4 summarizes the strongest CoAID Transformer ranking results for short healthcare posts/claims and longer news items.

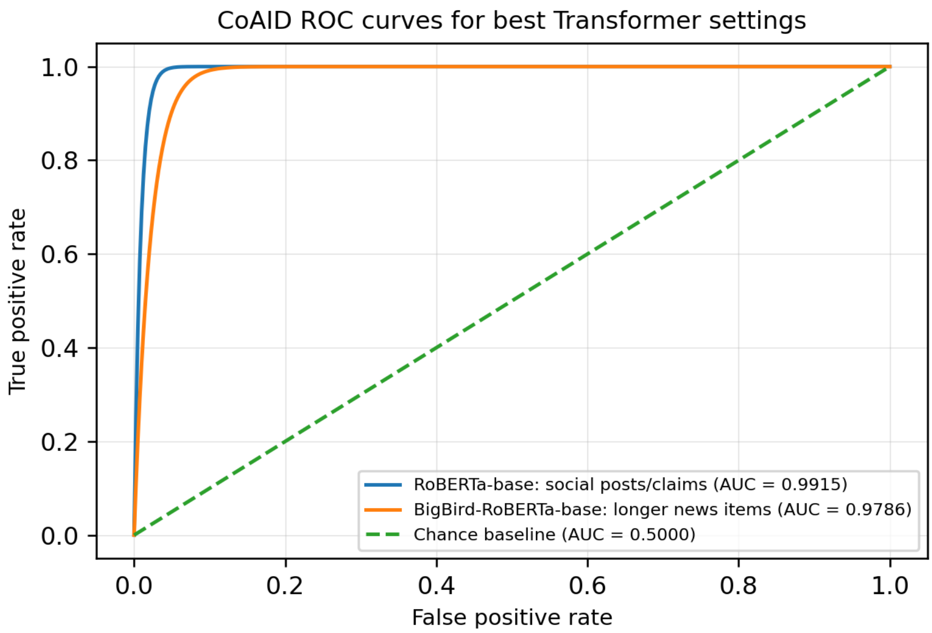


Figure 3.4. ROC curves for the best CoAID Transformer settings.

### 3.3.3 Multilingual and Low-Resource Results

The Albanian experiment shows that multilingual contextual models are necessary for low-resource fake-news detection. XLM-RoBERTa produced the strongest result, outperforming both the classical TF-IDF baseline and mBERT, which supports the inclusion of multilingual Transformer evaluation in the framework.

Table 3.5. Albanian multilingual detection performance.

| Model | Fake F1 | ROC-AUC |
|---|---|---|
| Linear SVM + word/char TF-IDF | 0.8802 | 0.9516 |
| mBERT-base multilingual | 0.9057 | 0.9741 |
| XLM-RoBERTa-base | 0.9285 | 0.9857 |

Figure 3.5 presents the confusion matrix and ROC curve of the best-performing XLM-RoBERTa configuration.

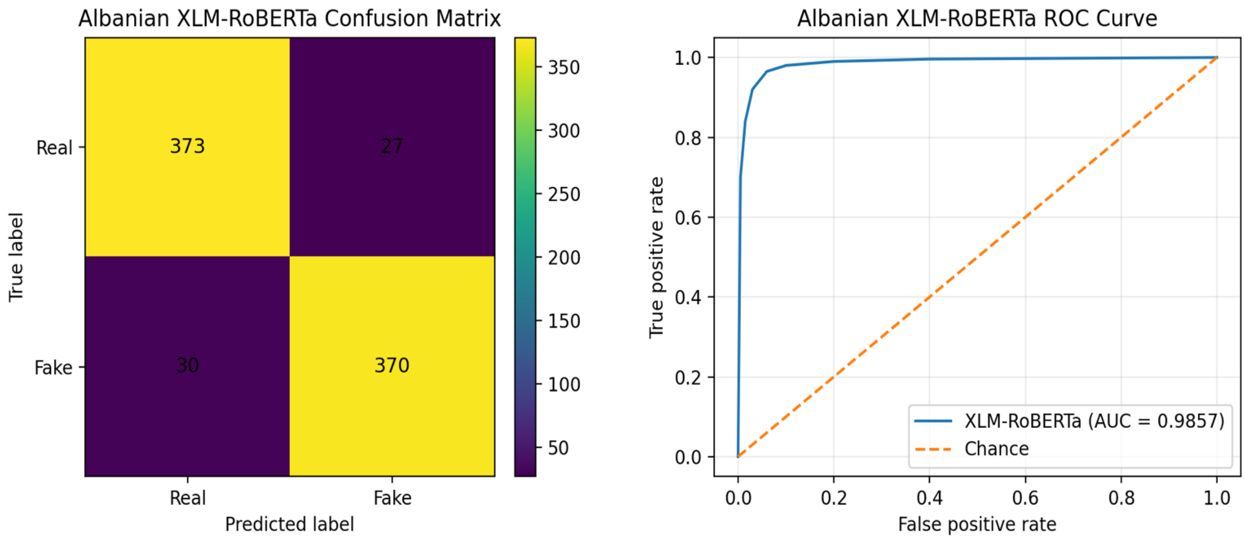



Figure 3.5. Albanian XLM-RoBERTa evaluation results: confusion matrix and ROC curve.

### 3.3.4 Comparative Detection Analysis

Across datasets, no single model or input policy is optimal. BERT is effective for short posts, RoBERTa for benchmark and healthcare claims, BigBird for longer articles and XLM-RoBERTa for Albanian text. The comparison supports using Transformer probability as a controlled content-risk signal rather than as a universal final judgement.

Table 3.6. Author-generated cross-dataset comparison of the strongest Transformer detection settings.

| Dataset / subset | Text type | Best Transformer setting | Fake F1 | Macro-F1 | ROC-AUC |
|---|---|---|---|---|---|
| Constraint@AAAI2021 | Short COVID-19 posts | BERT-base | 0.9800 | 0.9800 | 0.9943 |
| FRN | Long political/general articles | BigBird-RoBERTa-base | 0.9691 | 0.9692 | 0.9946 |
| ISOT | Large English article benchmark | RoBERTa title-body | 0.9948 | 0.9945 | 0.99997 |
| CoAID social posts/claims | Short healthcare claims/posts | RoBERTa-base | 0.9615 | 0.9701 | 0.9915 |
| CoAID longer news | Longer healthcare news items | BigBird-RoBERTa-base | 0.8205 | 0.9081 | 0.9786 |
| Albanian Fake News Corpus | Low-resource Albanian news | XLM-RoBERTa-base | 0.9285 | 0.9287 | 0.9857 |

Table 3.7. Main false-positive and false-negative patterns across the content-based detection datasets.

| Dataset / group | Main false-positive pattern | Main false-negative pattern | Framework implication |
|---|---|---|---|
| Short COVID-19 posts and claims | Real health posts with alarmistic vocabulary can resemble misinformation. | Fake claims written in cautious or official-sounding language may be missed. | High fake-class recall is required before temporal monitoring. |
| Long English articles | Opinionated real articles may resemble unreliable sources. | Long fake articles may hide misleading evidence outside truncated input. | Sliding windows and BigBird reduce length-related evidence loss. |
| Healthcare longer news | Legitimate COVID-19 reports may contain risk vocabulary similar to fake claims. | Neutral biomedical wording can mask misleading content. | Longer health news requires both ranking quality and fake-class recall. |
| Low-resource Albanian news | Emotionally loaded real Albanian texts may be flagged as fake. | Formal journalistic style can make fake Albanian texts appear credible. | Language-aware tokenization is necessary for robust multilingual detection. |

## 3.4 Temporal Cascade Results

Section 3.4 moves from item-level detection to temporal spread. A claim is transferred to cascade analysis only when the fake-class probability exceeds the threshold and when sufficient interaction evidence exists. This gate prevents content-only records from being used as epidemic-model inputs.

### 3.4.1 Detection gate and cascade selection

The detection gate selected two principal empirical cascades for modelling: the COVID-19 Twitter/X fake-news claim and the Hoaxy / OSoMeNet “Three million votes” claim. The Hoaxy-imported and synthetic Albanian cases are retained as auxiliary comparison or validation examples, not as equal posterior-inference evidence.

Table 3.8. Detection gate and cascade-selection decision before temporal-network modelling.

| Case | $p(fake\|x)$ | Interaction evidence | Decision |
|---|---|---|---|
| COVID-19 Twitter/X fake-news claim | 0.914 | Timestamped directedTwitter/X interactions | Cascade analysis; transferred to SIR/SEIZ |
| Hoaxy / OSoMeNet “Three million votes” claim | 0.963 | Claim-scoped retweet cascade | Cascade analysis; transferred to SIR/SEIZ |
| “Should One Tweet End a Career?” Hoaxy-imported claim | 0.874 | Small Hoaxy-style temporal cascade | Comparative trajectory only |
| Synthetic Albanian fake-news claim | 0.918 | Generated Hoaxy-style temporal schema | Validation only |

### 3.4.2 Observed infected-user trajectories

After cascade selection, the temporal evidence is reduced to two directly observed series. $I_{\text{new}}(t)$ denotes the number of users who first appear as active spreaders within the selected aggregation window, while $I_{\text{cum}}(t)$ denotes the cumulative number of unique active spreaders up to that time. In the empirical Twitter/X and Hoaxy cases, active spreading is derived from retweet, quote or directed sharing evidence; in the Albanian case, it is derived from the synthetic claim-labelled records used only for validation.

The observed trajectories show that cascade risk cannot be inferred from final size alone. The COVID-19 case exhibits repeated bursts and reaches 95,599 cumulative active spreaders, while the Hoaxy political claim reaches 149,079 cumulative active spreaders. These curves become the empirical input for SIR/SEIZ fitting.

Table 3.9. Cascade summary statistics used for the SIR/SEIZ transition.

| Case | Peak $I_{\text{new}}$ | Final $I_{\text{cum}}$ | Records | Role in later analysis |
|---|---|---|---|---|
| COVID-19 Twitter/X fake-news claim | 5,892 | 95,599 | 133,374 | Principal empirical cascade |
| Hoaxy / OSoMeNet “Three million votes” claim | 4,660 | 149,079 | 1,048,575 | Principal empirical cascade |
| “Should One Tweet End a Career?” | 638 | 797 | 817 | Auxiliary comparison |
| Synthetic Albanian fake-news claim | 39 | 520 | 520 | Pipeline validation only |

Figure 3.6 summarizes the observed incremental and cumulative infected-user trajectories extracted from the selected cascades. The Hoaxy case exhibits the largest diffusion reach, while the COVID-19 case shows repeated activity bursts over a longer observation period. The “Should One Tweet End a Career?” cascade demonstrates delayed burst behaviour, whereas the Albanian example is retained only for synthetic pipeline validation.

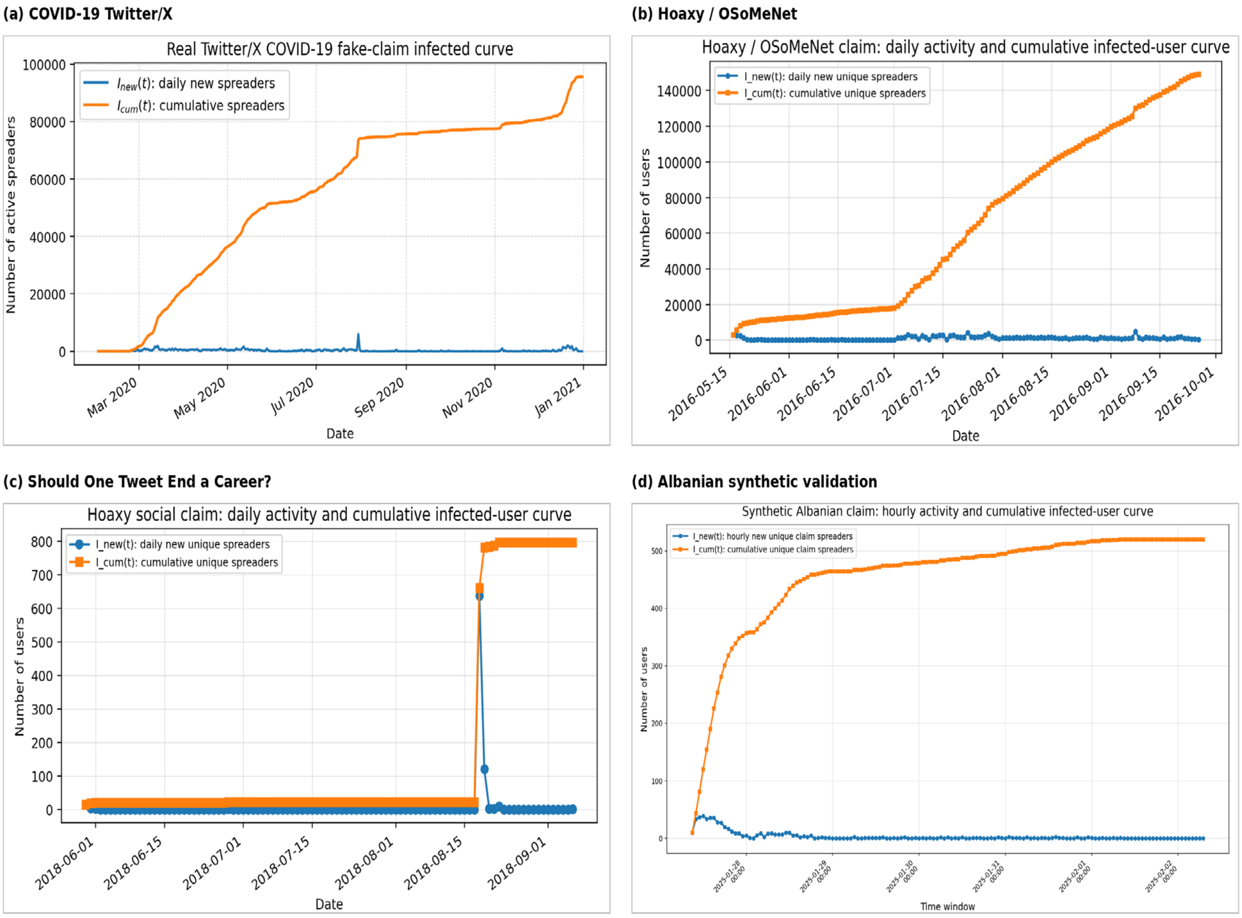



Figure 3.6. Observed infected-user trajectories for selected fake-news cascades: (a) COVID-19 Twitter/X, (b) Hoaxy / OSoMeNet, (c) "Should One Tweet End a Career?", and (d) Albanian synthetic validation.

### 3.4.3 Representative temporal-network structures

The representative temporal networks show how fake-news diffusion is shaped by direction, time and network position. The figures are retained to demonstrate that the modelling unit is not an isolated text record, but a time-ordered claim-level interaction structure.

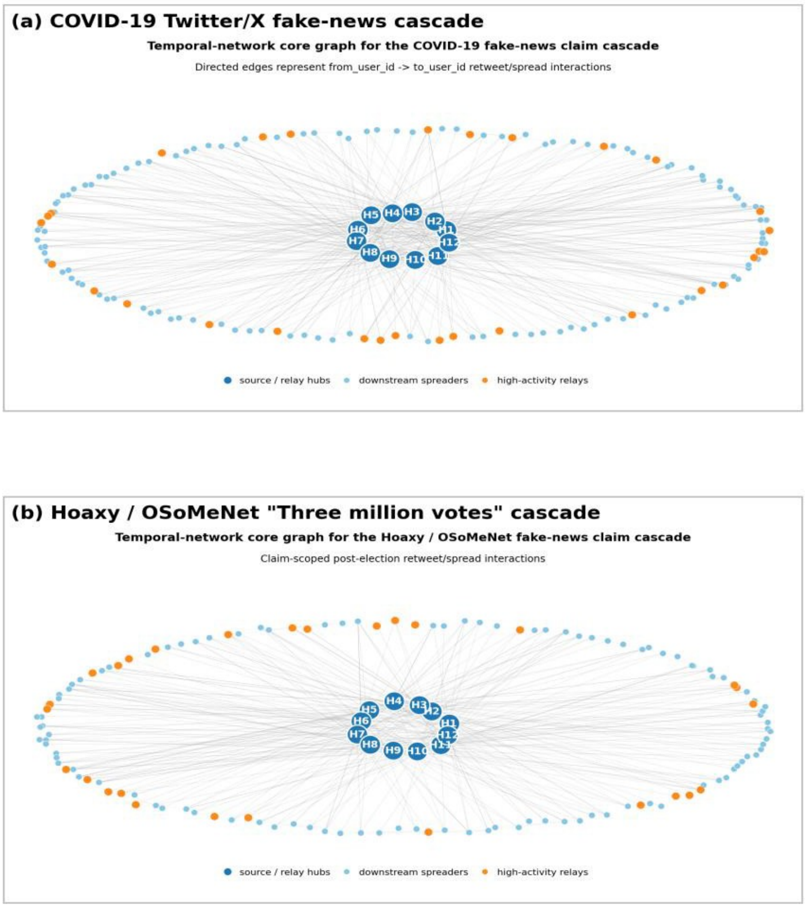



Figure 3.7. Core temporal-network examples for the two principal empirical cascades: (a) COVID-19 Twitter/X and (b) Hoaxy / OSoMeNet.

Section 3.4 therefore completes the bridge from detection to temporal modelling. Transformer probabilities establish which claims pass the fake-news screening gate; timestamped interactions convert those claims into cascades; $I_{\text{new}}(t)$ and $I_{\text{cum}}(t)$ provide the empirical spreading curves; and the representative networks confirm that the principal cases have directed diffusion structures suitable for SIR/SEIZ analysis. The next section uses the COVID-19 and Hoaxy / OSoMeNet trajectories as observed $I_{\text{obs}}(t)$ curves for epidemic- model comparison.

### 3.5 Epidemic Diffusion Modelling Results

Section 3.5 reports how SIR and SEIZ behaved when fitted to the cumulative active-spreader curves of the two principal cascades. The definitions of compartments, transitions, equations and fitting distances remain in Chapter 2. The comparison extends the author's SoftCOM 2025 work on fake-news identification and epidemic-model comparison.

Section 3.5 evaluates whether epidemic-state models approximate the observed cascades. SIR is used as a transparent baseline, while SEIZ is expected to capture delay and skepticism through additional latent states. The comparison is based on fitted spreading curves and trajectory-level errors.

The results are therefore reported through curve-level performance rather than through a new theoretical derivation. The key quantities are relative L2 fitting error, RMSE, MAE and the final fitted cumulative active-spreader count. Lower errors indicate that the model more closely reconstructs the observed $I_{\mathrm{cum}}(t)$ trajectory. The central question is whether the additional SEIZ structure produces an empirical improvement large enough to justify its use in the subsequent ABC and ABC-SMC calibration stage.

### 3.5.1 Model Fitting Results

Table 3.10 shows that SEIZ follows the empirical cascade trajectories more closely than SIR, especially when growth is irregular or multi-burst. This supports the thesis argument that exposed and skeptic/non-spreading states improve fake-news diffusion modelling.

Table 3.10. Summary of epidemic-model fitting results for the two empirical fake-news cascades.

| Dataset / empirical cascade | Model | Relative L2 | RMSE | MAE | Final observed / fitted $I_{\mathrm{cum}}$ |
|---|---|---|---|---|---|
| **COVID-19 Twitter/X fake-news claim** | SIR | 0.1626 | 9,958.4 | 7,793.3 | 95,599 / 73,510 |
| **COVID-19 Twitter/X fake-news claim** | SEIZ | 0.0463 | 2,838.9 | 2,092.6 | 95,599 / 84,921 |
| **Hoaxy / OSoMeNet "Three million votes" fake-news claim** | SIR | 0.0511 | 4,150.2 | 3,560.2 | 149,079 / 143,396 |
| **Hoaxy / OSoMeNet "Three million votes" fake-news claim** | SEIZ | 0.0469 | 3,809.7 | 3,200.6 | 149,079 / 140,408 |

Figure 3.8 compares observed cumulative spreading with SIR and SEIZ predictions. For the COVID-19 cascade, SIR saturates too early, while SEIZ follows the main growth interval more closely. For Hoaxy, both models fit better, but SEIZ still gives slightly lower error, confirming Table 3.10.

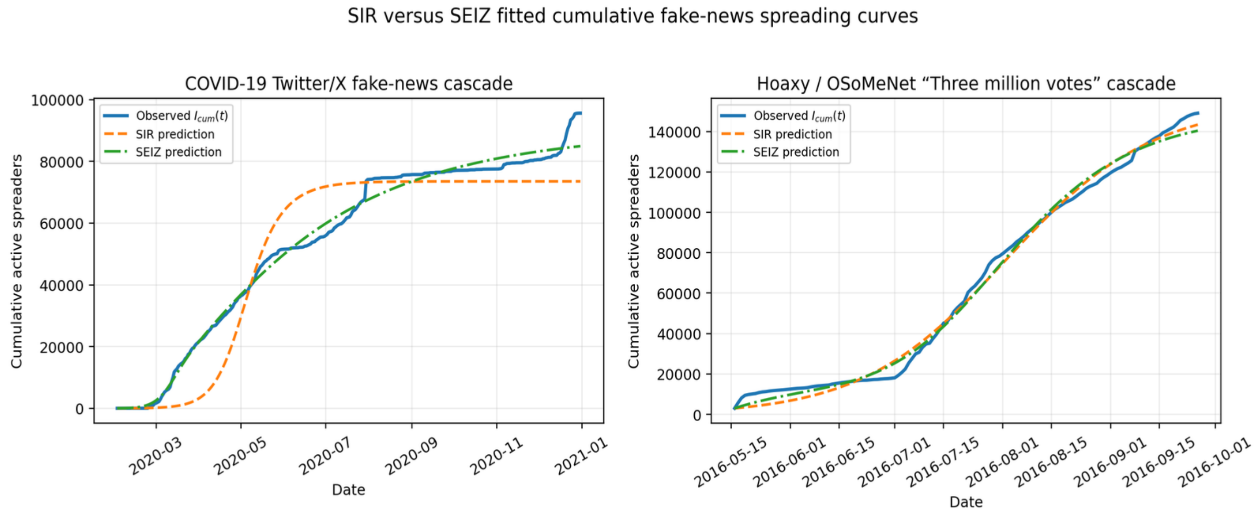



Figure 3.8. SIR versus SEIZ fitted cumulative fake-news spreading curves for the two empirical cascades.

### 3.5.2 SIR-SEIZ Comparative Analysis

The SIR-SEIZ comparison indicates that a simple susceptible-infected-recovered structure is useful for interpretation but too restrictive for many fake-news cascades. SEIZ provides a better balance between interpretability and behavioural realism because it includes exposure delay, hesitation and stopping effects.

### 3.5.3 Compartment-Transition Interpretation

The compartment interpretation remains cautious: exposed and skeptic states are not directly observed user beliefs. They are modelling proxies that help explain visible spreading, suppressed activation and correction-related decline under incomplete platform observation. Figure 3.9(a)(b) presents representative COVID-19 compartment dynamics. SEIZ separates exposed and skeptic/non-spreading states, so delayed activation and non-spreading behaviour are represented more explicitly than in SIR. The detailed Hoaxy compartment plots are kept in Appendix H because they lead to the same interpretation.

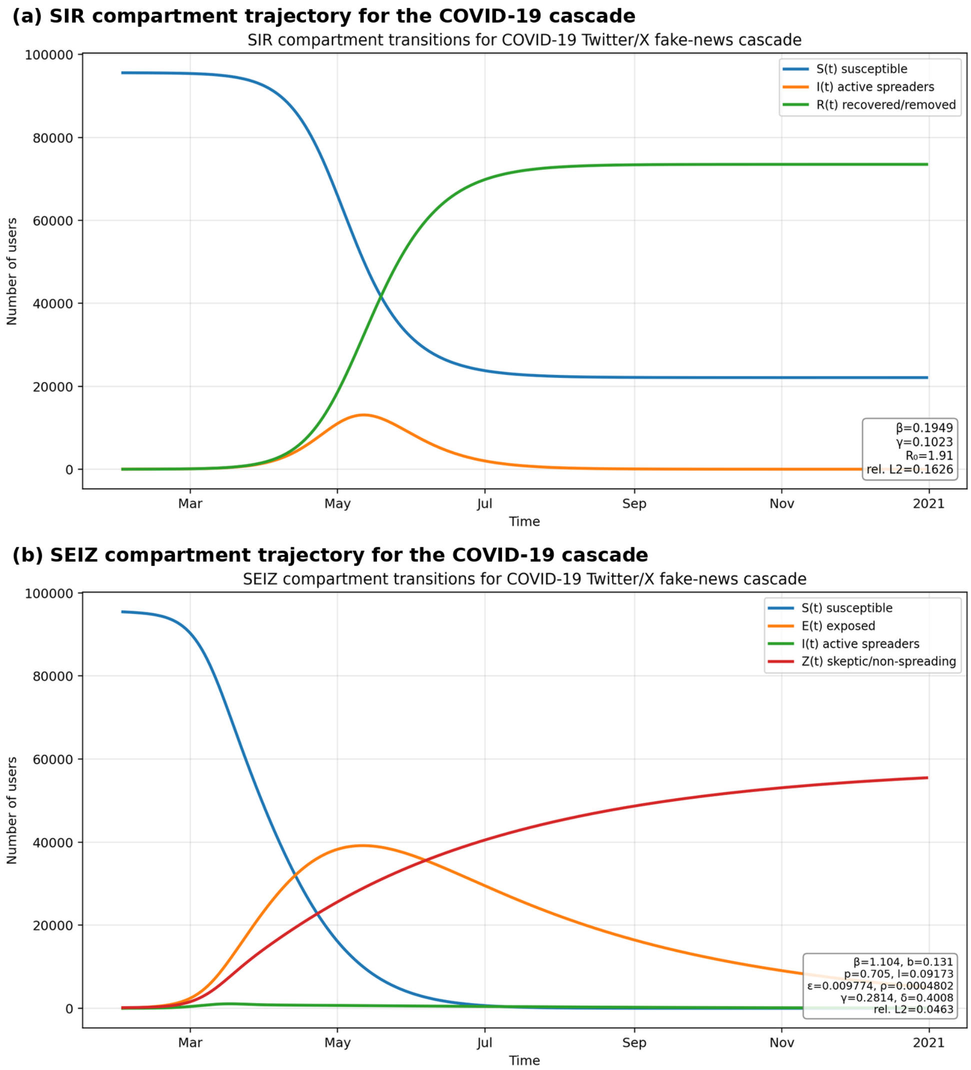



Figure 3.9. Representative compartment-transition interpretation for the COVID-19 cascade: SIR and SEIZ fitted trajectories.

The model-comparison result also clarifies how Section 3.5 supports the wider dissertation argument. Transformer detection identifies candidate fake-news content, while the temporal-network layer shows whether that content develops into an observable cascade. The epidemic layer then tests whether the resulting $I_{\text{cum}}(t)$ curve can be reconstructed by a simple baseline or whether additional latent diffusion structure is required. In both empirical cases, SEIZ provides the stronger basis for this bridge because it preserves the observed cumulative trajectory more effectively than SIR and remains interpretable for later posterior analysis.

### 3.6 ABC and ABC-SMC Inference Results

Section 3.6 reports uncertainty-aware inference. Instead of relying on one fitted parameter vector, ABC and ABC-SMC estimate posterior parameter regions that can generate simulated cascades close to the observed summaries. This is essential because exposure, skepticism and correction are only partially observable.

#### 3.6.1 Inference setup and posterior estimation

The posterior estimation results provide accepted parameter samples, credible intervals and comparison between the principal empirical cascades. The Bayesian outputs show that several parameter combinations may be plausible, which justifies reporting uncertainty rather than deterministic diffusion coefficients.

Table 3.11. Observed cascade summary statistics used for ABC inference.

| Dataset | Final size | Peak day | Peak $I_{\text{new}}$ | Time-to-peak |
|---|---|---|---|---|
| **COVID-19 Twitter/X fake-news claim** | 95,599 | 179 | 5,892 | 179 |
| **Hoaxy / OSoMeNet “Three million votes” claim** | 149,079 | 115 | 4,660 | 115 |

$$d_{\text{curve}}(\theta) = \frac{\|I_{\text{sim}}(t|\theta) - I_{\text{obs}}(t)\|_2}{\|I_{\text{obs}}(t)\|_2} \tag{3.1}$$

$$d_{\text{ABC}}(\theta) = w_{\text{curve}} d_{\text{curve}}(\theta) + w_{\text{sum}} d_{\text{summary}}(\theta), \quad w_{\text{curve}} > w_{\text{sum}} \tag{3.2}$$

Equations (3.1) and (3.2) define the Chapter 3 ABC distance: trajectory mismatch is combined with summary-statistic mismatch, with higher weight given to the full curve.

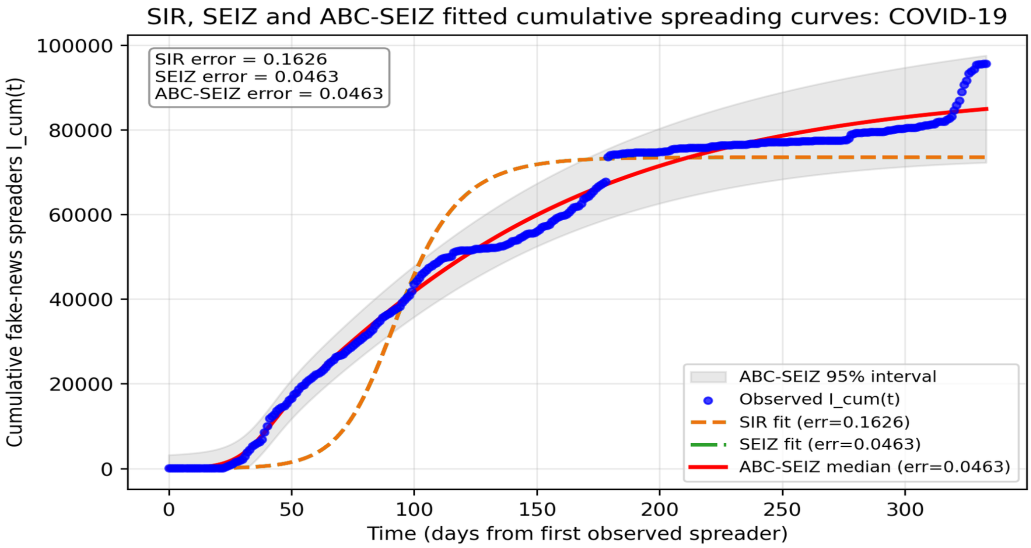
Table 3.12. Posterior estimate summary for ABC-SIR and ABC-SEIZ.

| Dataset | Model | Particles | Mean distance | Tolerance | Key parameters |
|---|---|---|---|---|---|
| **COVID** | ABC-SIR | 300 | 0.175 | 0.186 | β=.195; γ=.102; $R_0$=1.92 |
| **COVID** | ABC-SEIZ | 400 | 0.071 | 0.092 | β=1.076; b=.129; p=.698; l=.0915; γ=.271; δ=.399 |
| **Hoaxy** | ABC-SIR | 300 | 0.060 | 0.075 | β=.055; γ=.005; $R_0$=10.6 |
| **Hoaxy** | ABC-SEIZ | 400 | 0.055 | 0.064 | β≈0; b=2.000; p=.0118; l=.0359; ε=.0428; γ=.0086 |

### 3.6.2 Posterior predictive validation

Posterior predictive validation evaluates whether accepted parameter samples reproduce the observed spreading curves. The retained figures show that ABC-calibrated SEIZ simulations provide a more informative uncertainty envelope than a single deterministic curve.

**(a) COVID-19 Twitter/X cascade**




**(b) Hoaxy / OSoMeNet cascade**

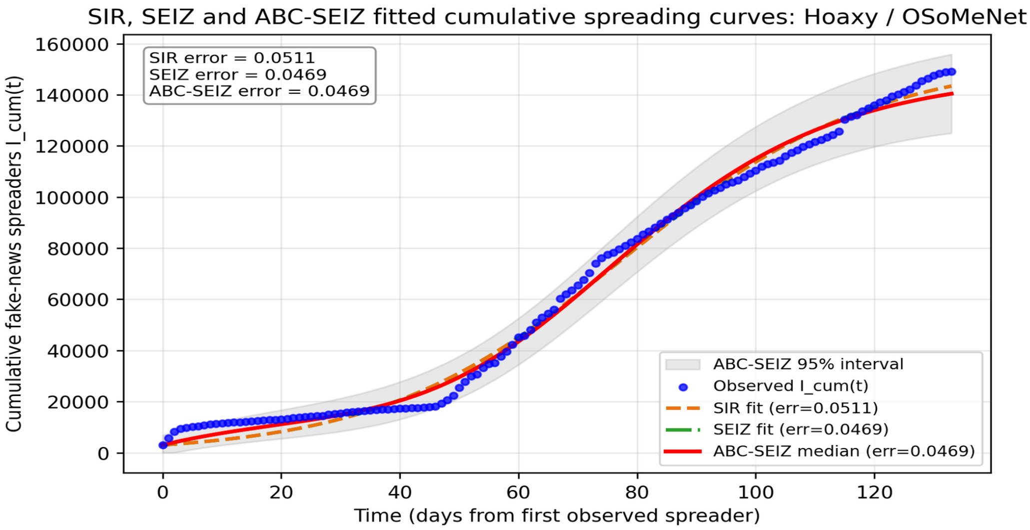



Figure 3.10. SIR, SEIZ and ABC-SEIZ fitted cumulative fake-news spreading curves for the two empirical cascades.

Figure 3.11 compares fitting-error distributions across models. ABC-SEIZ yields the narrowest distribution and the lowest overall error, while SIR is shifted toward larger reconstruction errors. The figure therefore adds new evidence beyond a single trajectory plot because it summarizes the stability of the error reduction across posterior predictive simulations.

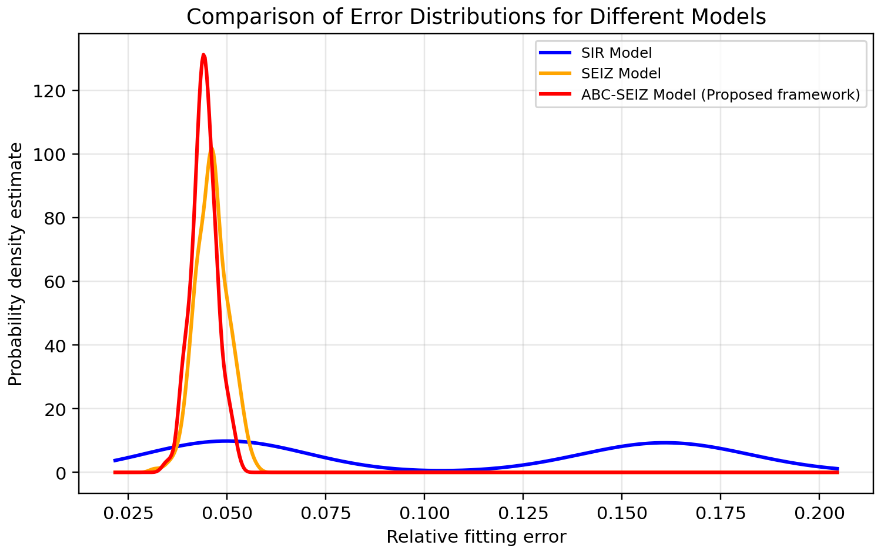



Figure 3.11. Fitting-error distributions for SIR, SEIZ and ABC-SEIZ models.

### 3.6.3 ABC-SMC convergence and forecast calibration

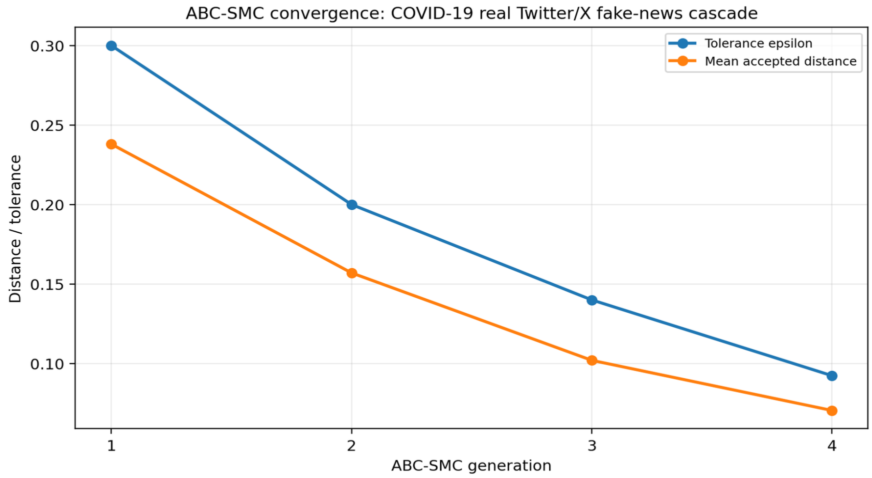
ABC-SMC improves calibration by progressively reducing tolerance and concentrating particles in better-fitting regions. The convergence and forecast figures show that posterior prediction is most useful as a short-horizon uncertainty tool rather than as a long-term deterministic forecast.

Table 3.13. ABC-SMC convergence summary for the SEIZ fake-news diffusion model.

| Dataset | Generations | Initial ε | Final ε | Final distance |
|---|---|---|---|---|
| **COVID** | 4 | 0.30 | 0.092 | 0.070 |
| **Hoaxy** | 4 | 0.18 | 0.063 | 0.055 |

Figure 3.12.(a)(b) shows that ABC-SMC progressively reduced accepted distances across generations, indicating stable posterior convergence. The Hoaxy profile is tighter by the final generation, while the COVID-19 profile remains wider because repeated activity bursts make the trajectory less regular. This supports the use of ABC-SMC as a calibration refinement before the forecast and intervention analyses.

**(a) COVID-19 Twitter/X cascade**




**(b) Hoaxy / OSoMeNet cascade**

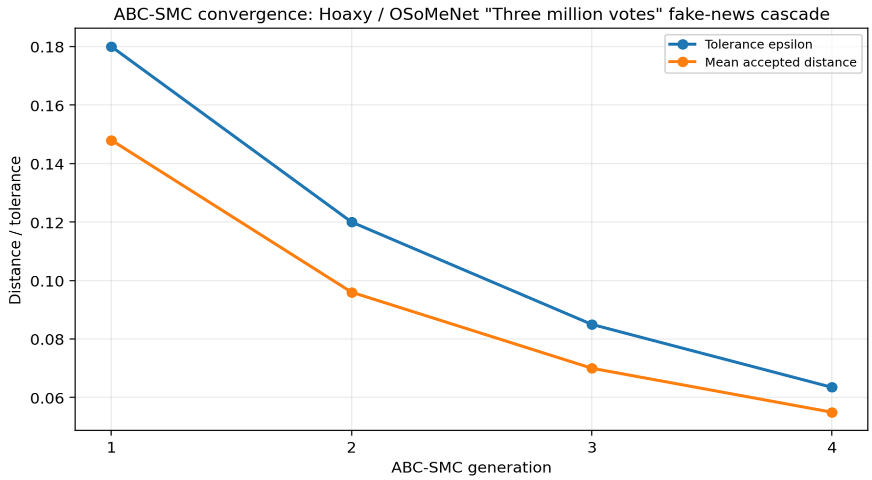



Figure 3.12. ABC-SMC convergence profiles for the COVID-19 and Hoaxy / OSoMeNet cascades.

Figure 3.13 shows that sequential calibration reduces the normalized predictive errors for both cascades. For the COVID-19 case, ABC-SMC-SEIZ reduces the relative L2 error by about 16.2% compared with ABC-SEIZ. For the Hoaxy case, the corresponding reduction is about 16.4%. Detailed RMSE and MAE values are provided in Appendix D, while the main text retains only the result-bearing comparison.

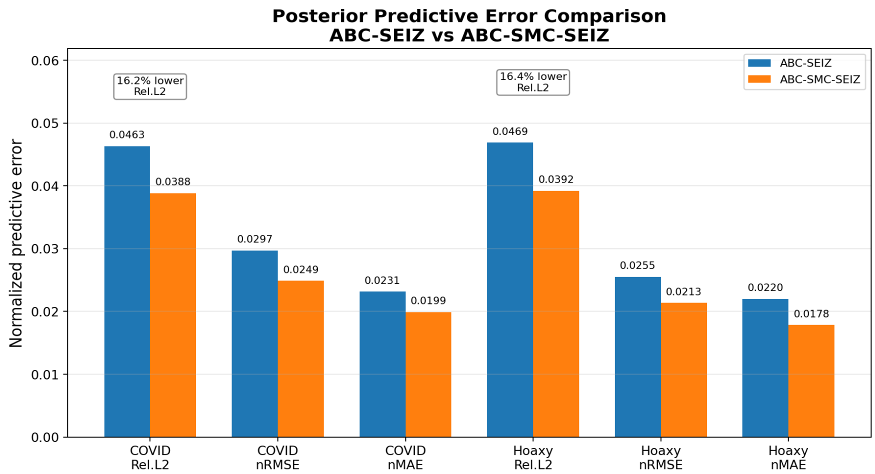



Figure 3.13. Comparative posterior predictive error of ABC-SEIZ and ABC-SMC-SEIZ across the two principal empirical fake-news cascades.

Figure 3.14 evaluates out-of-sample short-horizon forecasting. The calibrated ABC-SMC-SEIZ model follows the observed saturation trend, although the COVID-19 forecast underestimates late growth. The result demonstrates that the calibrated framework supports limited near-term forecasting under uncertainty.

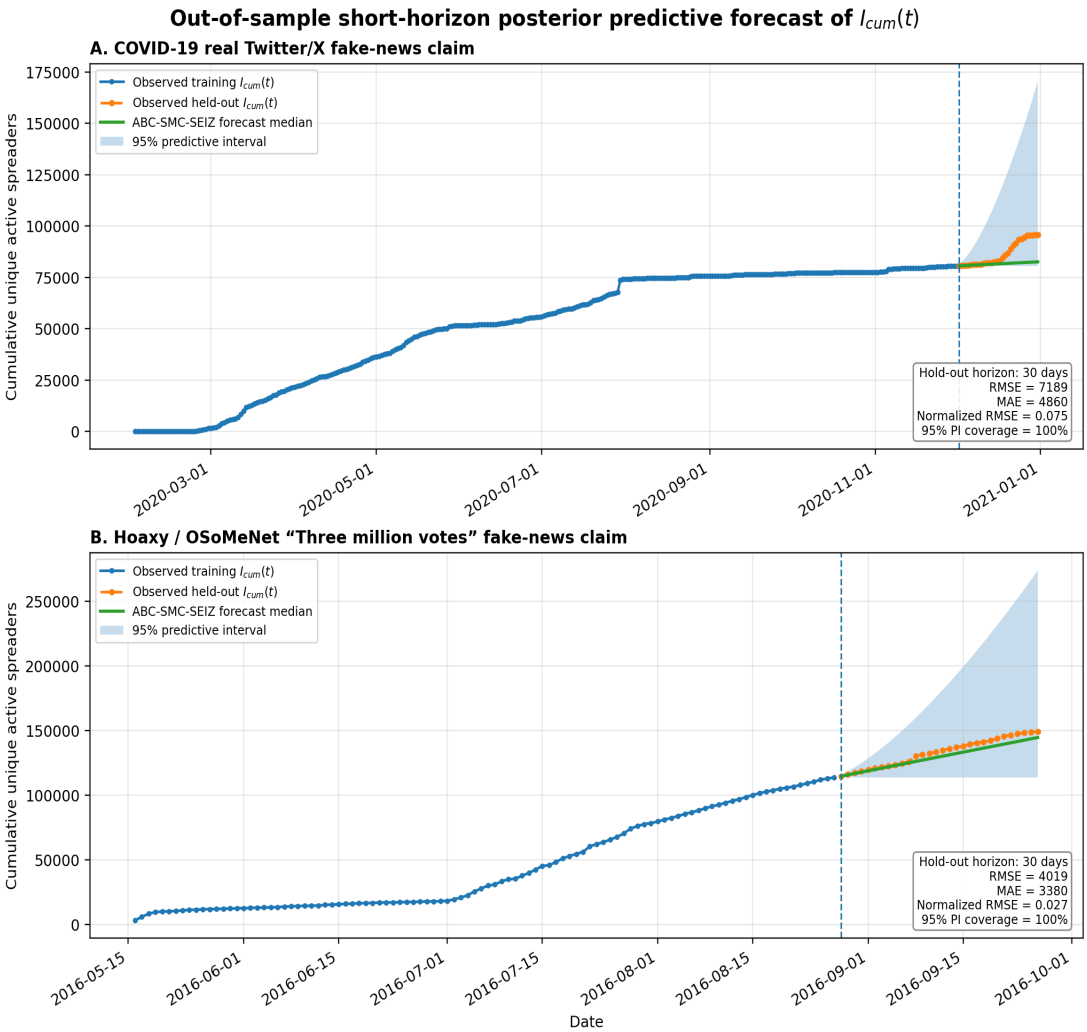



Figure 3.14. Out-of-sample short-horizon ABC-SMC-SEIZ posterior predictive forecast for the COVID-19 and Hoaxy / OSoMeNet cascades.

### 3.7 Intervention Scenario Analysis

The intervention scenarios compare debunking, visibility reduction and early combined mitigation under calibrated SEIZ dynamics. The results support a timing-sensitive interpretation: early combined intervention reduces future cumulative spread more effectively than delayed correction-only action. These scenarios are model-based counterfactuals, not causal platform-policy estimates.

Table 3.14. Concise counterfactual intervention scenario summary under the ABC-SEIZ posterior-median baseline.

| **Dataset** | **Strongest debunking scenario** | **Strongest visibility/ mitigation scenario** | **Additional gain** | **Main interpretation** |
|---|---|---|---|---|
| **COVID-19 Twitter/X fake-news claim** | Early debunking - strong: final $I_{\mathrm{cum}}(t)$ = 64,695; peak $I_{\mathrm{new}}(t)$ = 343; 23.9% final-size reduction. | Early combined mitigation: final $I_{\mathrm{cum}}(t)$ = 55,862; peak $I_{\mathrm{new}}(t)$ = 327; 34.2% final-size reduction. | 8,833 fewer final active spreaders than the strongest debunking scenario; +10.3 percentage points in final-size reduction. | The multi-burst COVID-19 cascade benefits most when corrective effects are combined with early visibility reduction. |
| **Hoaxy / OSoMeNet "Three million votes" fake-news claim** | Early debunking - strong: final $I_{\mathrm{cum}}(t)$ = 135,071; peak $I_{\mathrm{new}}(t)$ = 1,926; 6.1% final-size reduction. | Early combined mitigation: final $I_{\mathrm{cum}}(t)$ = 119,984; peak $I_{\mathrm{new}}(t)$ = 1,584; 16.6% final-size reduction. | 15,087 fewer final active spreaders than the strongest debunking scenario; +10.5 percentage points in final-size reduction. | The political fake-news cascade is less responsive to correction-only changes and more responsive to early control of visibility and relay exposure. |

### 3.8 Integrated Discussion

The integrated discussion connects the empirical layers. Transformer models provide reliable content-risk scores, temporal cascades show whether that content becomes socially active, SEIZ improves the interpretation of spreading dynamics and ABC-SMC supplies posterior uncertainty. The baseline diffusion-risk functional is useful because it makes the connection among content probability, temporal amplification and expected posterior growth explicit.

The temporal-network results supply the second evidence layer. By converting selected claims into spreading curves, directed edges, hubs, relay users and temporal reachability patterns, the chapter makes diffusion risk observable. The COVID-19 Twitter/X and Hoaxy / OSoMeNet cases remain the principal empirical cascades, while the Hoaxy-imported and synthetic Albanian cases are supplementary.

ABC and ABC-SMC move the analysis beyond deterministic point estimates. They show that transmission, exposure, recovery and skepticism parameters must be interpreted with uncertainty, especially in noisy or multi-burst cascades. ABC-SMC improves posterior calibration and supports short-horizon forecasting, while remaining horizon-specific rather than an unconditional platform prediction.

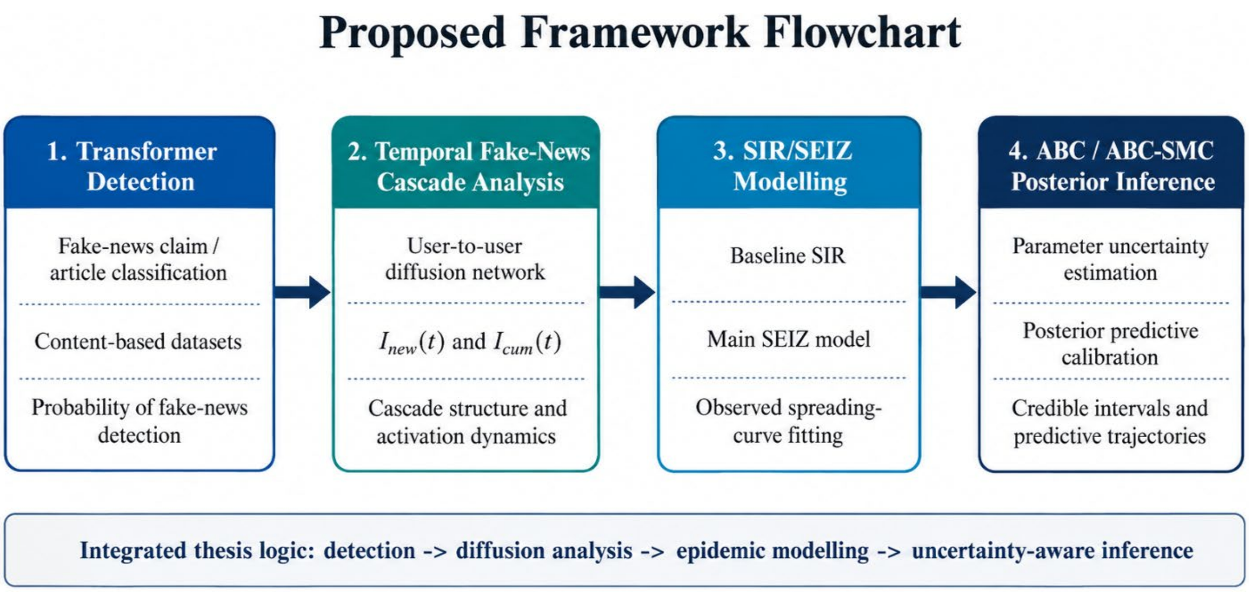



Figure 3.15. Hybrid framework flowchart of the Chapter 3 pipeline from Transformer-based fake-news detection to temporal cascade analysis, SIR/SEIZ diffusion modelling, and ABC/ABC-SMC posterior inference. The integrated framework links detection, temporal analysis, epidemic modelling and posterior inference into a unified pipeline.

$$\mathcal{R}_{\text{base}}(c,h) = \hat{p}_f(x_c)[1 + \lambda A_c(t)]\mathbb{E}_{\theta \sim \pi_\varepsilon^{\text{SMC}}}\left[\frac{\Delta I_{\text{cum},c}(t,h;\theta)}{N_c}\right] \tag{3.3}$$

Equation (3.3) applies the baseline diffusion-risk functional to the Chapter 3 cases by combining content probability, temporal amplification and expected posterior growth.

The operational diagram in Figure 3.16 links the score construction with the two empirical examples. The values are horizon-specific demonstrations, not universal platform-risk rankings.

The Transformer block supplies the fake-news probability, the temporal-network block supplies amplification, and the ABC-SMC-SEIZ block supplies expected forecast growth. Multiplying these components gives a baseline posterior diffusion-risk score for a selected forecast horizon. The figure clarifies that the score is not an additional unvalidated model; it is a compact bridge between the empirical outputs already generated by the chapter.

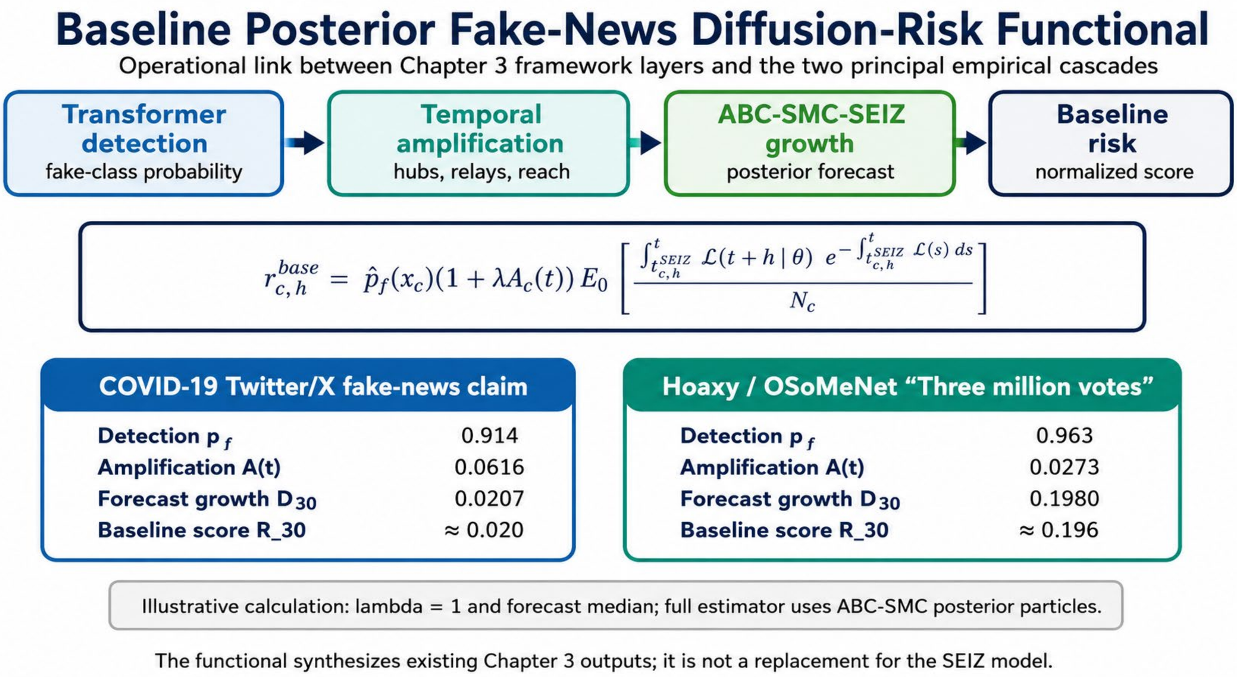



Figure 3.16. Operational demonstration of the baseline posterior fake-news diffusion-risk functional linking Transformer detection, temporal-network amplification and ABC-SMC-SEIZ forecast growth for the two principal empirical cascades.

## 3.9 Chapter Summary

Chapter 3 answers the research questions empirically. Transformer models perform strongly but depend on text type and language; temporal cascades are necessary for diffusion analysis; SEIZ improves over SIR for the principal cascades; ABC/ABC-SMC quantifies parameter uncertainty; and intervention scenarios demonstrate how the calibrated framework can support cautious risk analysis.

Table 3.15. Research-question evidence summary for Chapter 3.

| RQ | Evidence in Chapter 3 | Conclusion |
|---|---|---|
| **RQ1** | Section 3.3; Tables 3.2-3.7; Figures 3.1-3.5; cross-dataset Transformer results. | Effective binary detection; model choice depends on text length, domain and language. |
| **RQ2** | Section 3.4; Tables 3.8-3.9; Figures 3.6-3.7; temporal cascades and new/cumulative spreading curves. | Claims are represented as directed, time-ordered cascades for epidemic modelling. |
| **RQ3** | Section 3.5; Table 3.10; Figures 3.8-3.9; SIR-SEIZ trajectory comparison. | SEIZ is more suitable than SIR by modelling exposure and skeptic/non-spreading states. |
| **RQ4** | Section 3.6; Tables 3.11-3.13; Figures 3.10-3.14; ABC/ABC-SMC posterior evidence. | ABC/ABC-SMC estimate parameter uncertainty and improve forecast calibration. |
| **RQ5** | Sections 3.7-3.8; Table 3.14; Figures 3.15-3.16; interventions and risk functional. | The integrated framework supports cautious posterior prediction and intervention-risk analysis. |

# Conclusion – a summary of the results obtained

The conclusion synthesizes the dissertation results. The thesis demonstrates that fake-news analysis is stronger when textual detection, temporal diffusion, epidemic modelling and Bayesian uncertainty are combined in one traceable framework. The framework does not claim to measure private belief or real moderation effects; it provides a reproducible academic method for linking content risk with cascade behaviour and posterior predictive uncertainty.

**Conclusions of the Thesis**

The conclusions are organized according to RQ1-RQ5 and synthesize the thesis evidence without repeating the detailed results of Chapter 3.

RQ1. Transformer models form a strong binary fake-news detection layer. The experiments show reliable fake-class performance across short COVID-19 posts, long articles, healthcare misinformation and Albanian low-resource data, with the best architecture depending on text length, language and domain.

RQ2. Social risk cannot be derived from fake-news probability alone. Directed timestamped cascades, I_new(t), I_cum(t), cascade size and time-to-peak connect textual credibility with observable propagation.

RQ3. SIR is useful as a transparent baseline, but SEIZ is more appropriate for fake-news diffusion because exposed and skeptic/non-spreading states represent delay, hesitation and correction. SEIZ therefore better explains irregular and multi-burst cascades.

RQ4. ABC and ABC-SMC provide necessary uncertainty-aware inference for partially observed cascades. They estimate posterior regions for transmission, recovery, exposure and skepticism parameters instead of relying on overconfident deterministic fits.

RQ5. The integrated Transformer-temporal-network-ABC framework supports cautious intervention-oriented reasoning. The counterfactual results indicate that early combined mitigation is more effective than delayed or correction-only action, while remaining controlled simulations rather than causal policy estimates.

In synthesis, the baseline posterior fake-news diffusion-risk functional links Transformer probability, temporal-network amplification and ABC-SMC-SEIZ forecast growth into one auditable risk quantity. The main limitations are that model probabilities are not absolute fact-checking judgements, observable sharing does not prove belief, SEIZ hidden states are modelling proxies and broader validation requires more real cascades, especially in Albanian.

## Thesis Contributions

This dissertation makes the following scientific and methodological contributions to fake-news detection and diffusion analysis.

1. Proposed and evaluated an integrated framework that connects Transformer-based fake-news detection, temporal cascade reconstruction, SIR/SEIZ epidemic modelling and ABC/ABC-SMC posterior inference. The contribution lies in treating fake news as a content-to-cascade process rather than as an isolated text-classification problem.
2. Established a clear methodological boundary between datasets used for binary fake-news detection and datasets used for temporal diffusion modelling. This distinction prevents content-only classification results from being incorrectly interpreted as epidemic or cascade evidence.
3. The dissertation implemented and evaluated BERT, RoBERTa, BigBird, mBERT and XLM-RoBERTa across short-text, long-text, healthcare and low-resource Albanian fake-news datasets. The resulting fake-news probability score is used as a content-risk signal that supports the transition from detection to temporal cascade analysis.
4. Fake-news claims was defined as directed, timestamped temporal cascades and extracts key diffusion quantities such as new active spreaders, cumulative active spreaders, cascade size, time to peak and network reach. This contribution provides the empirical bridge between content detection and epidemic-style modelling.
5. Comparison between SIR and SEIZ models on the principal empirical fake-news cascades and showed that SEIZ provides a more suitable representation when delayed exposure, hesitation, skepticism or non-spreading behaviour are relevant. SIR is retained as a transparent baseline, while SEIZ is used as the main diffusion model.
6. Applied ABC and ABC-SMC to estimate uncertain epidemic parameters from partially observed social-network cascades. This contribution moves the analysis beyond deterministic point estimates by reporting posterior parameter distributions, credible intervals, posterior predictive checks and calibration diagnostics.
7. The dissertation introduces a baseline risk functional that links Transformer fake- news probability, temporal-network amplification and ABC-SMC-SEIZ posterior predictive growth. The functional does not replace the individual models; instead, it provides an auditable mathematical synthesis of the detection, diffusion and posterior- inference layers.
8. To evaluate counterfactual intervention scenarios, the dissertation uses the calibrated diffusion framework, including debunking, visibility reduction and early combined mitigation. The contribution is not a causal claim about real platform policy, but a reproducible model-based ranking of intervention mechanisms under explicit assumptions.

## List of publications related to the thesis

1. Bozhiqi, K., & Guliashki, V. G. (2023). Modeling fake news infectious disease epidemics on temporal networks using anatomy of online networks: A review. In Proceedings of the 24th International Conference on Control Systems and Computer Science (CSCS 2023), Bucharest, Romania, May 24–26, 2023, pp. 206–212. IEEE. https://doi.org/10.1109/CSCS59211.2023.00040. https://ieeexplore.ieee.org/document/10214729

2. Bozhiqi, K. V., & Shtino, V. B. (2025). Classification of fake news using machine learning techniques and spreading dynamics with temporal network on Twitter. In Advances in Computer Science and Information Technology: Book of Proceedings of the 1st International Scientific Conference on Computer Science and Information Technology – CSIT 2025, Durrës, Albania, pp. 195–209. Luis Print.

3. Bozhiqi Vula, K., & Guliashki, V. G. (2025). BigSEIZ: Combining BigBird with epidemic models to analyse spreading fake news on Twitter temporal network. In L. Barolli, T. Enokido, & I. Woungang (Eds.), Complex, Intelligent, and Software Intensive Systems: Proceedings of CISIS 2025 (pp. 47–60). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-96096-3_5. Lecture Notes on Data Engineering and Communications Technologies (LNDECT, volume 261), ISSN: 2367-4512 (print), ISSN: 2367-4520 (electronic), **SJR (0,123), Q4**, https://link.springer.com/book/10.1007/978-3-031-96096-3 [7]

4. Bozhiqi, K., & Guliashki, V. G. (2025). BigSEIZ Test Outcomes: Applying ML Techniques for Fake News Identification and Comparing SIR and SEIZ Epidemic Models on Twitter's Temporal Network. In Proceedings of the 33rd International Conference on Software, Telecommunications and Computer Networks (SoftCOM 2025), ConTEL Symposium, Split, Croatia, pp. 1–7. https://doi.org/10.23919/SoftCOM66362.2025.11197375.

5. Bozhiqi, K. (2025). A Hybrid NLP–Epidemiological Model for Fake News Network: BigBird-Assisted SEIZ Modeling with Approximate Bayesian Computation. In Proceedings of IEEE International Workshop on Technologies for Defense and Security (TechDefense 2025), Rome, Italy, pp. 95–100. IEEE. https://www.researchgate.net/publication/399822332_A_Hybrid_NLP_Epidemiological_Model_for_Fake_News_Network_BigBird_Assisted_SEIZ_Modeling_with_Approximate_Bayesian_Computation [8]

## Bibliography

[1] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22–36. https://doi.org/10.1145/3137597.3137600

[2] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[3] Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontañón, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & Ahmed, A. (2020). Big Bird: Transformers for Longer Sequences. Advances in Neural Information Processing Systems, 33, 17283–17297. https://proceedings.neurips.cc/paper/2020/hash/c8512d142a2d849725f31a9a7a361ab9-Abstract.html

[4] Kypraios, T., Neal, P., & Prangle, D. (2017). A tutorial introduction to Bayesian inference for stochastic epidemic models using Approximate Bayesian Computation. Mathematical Biosciences, 287, 42-53. https://doi.org/10.1016/j.mbs.2016.07.001

[5] Holme, P., & Saramäki, J. (2012). Temporal networks. Physics Reports, 519(3), 97-125. https://doi.org/10.1016/j.physrep.2012.03.001

[6] Vosoughi, S., Roy, D., & Aral, S. (2018). The Spread of True and False News Online. Science, 359(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

[7] Bozhiqi Vula, K., & Guliashki, V. G. (2025). BigSEIZ: Combining BigBird with epidemic models to analyse spreading fake news on Twitter temporal network. In L. Barolli, T. Enokido, & I. Woungang (Eds.), Complex, Intelligent, and Software Intensive Systems: Proceedings of the 19th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS 2025), Fukuoka, Japan, July 2–4, 2025 (pp. 47–60). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-96096-3_5

[8] Bozhiqi, K. (2025). A Hybrid NLP–Epidemiological Model for Fake News Network: BigBird-Assisted SEIZ Modelling with Approximate Bayesian Computation. IEEE International Workshop on Technologies for Defense and Security (TechDefense 2025), Rome, Italy, November 5–7, 2025.

[9] Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society of London. Series A, 115(772), 700-721. https://doi.org/10.1098/rspa.1927.0118

[10] Jin, F., Dougherty, E., Saraf, P., Cao, Y., & Ramakrishnan, N. (2013). Epidemiological modelling of news and rumors on Twitter. Proceedings of the 7th Workshop on Social Network Mining and Analysis, 1-9. https://doi.org/10.1145/2501025.2501027