

Abstracts of Dissertations

**Institute of Information and
Communication Technologies**

**BULGARIAN ACADEMY OF
SCIENCES**



10/ 2026



**Transformer-Based
Detection and Epidemic-
Model Analysis of Fake-
News Dynamics on
Social Networks**

Kristel Vula

**Базирано на
трансформатори
откриване и анализ на
епидемичния модел на
динамиката на фалшиви
новини в социалните
мрежи**

Кристел Вула

Автореферати на дисертации

**Институт по информационни и
комуникационни технологии**

БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ

ISSN: 1314-6351

Поредицата „Автореферати на дисертации на Института по информационни и комуникационни технологии при Българската академия на науките“ представя в електронен формат автореферати на дисертации за получаване на научната степен „Доктор на науките“ или на образователната и научната степен „Доктор“, защитени в Института по информационни и комуникационни технологии при Българската академия на науките. Представените трудове отразяват нови научни и научно-приложни приноси в редица области на информационните и комуникационните технологии като Компютърни мрежи и архитектури, Паралелни алгоритми, Научни пресмятания, Лингвистично моделиране, Математически методи за обработка на сензорна информация, Информационни технологии в сигурността, Технологии за управление и обработка на знания, Грид-технологии и приложения, Оптимизация и вземане на решения, Обработка на сигнали и разпознаване на образи, Интелигентни системи, Информационни процеси и системи, Вградени интелигентни технологии, Йерархични системи, Комуникационни системи и услуги и др.

Редактори

Геннадий Агре

Институт по информационни и комуникационни технологии, Българска академия на науките
E-mail: agre@iinf.bas.bg

Райна Георгиева

Институт по информационни и комуникационни технологии, Българска академия на науките
E-mail: rayna@parallel.bas.bg

Даниела Борисова

Институт по информационни и комуникационни технологии, Българска академия на науките
E-mail: dborissova@iit.bas.bg

Настоящото издание е обект на авторско право. Всички права са запазени при превод, разпечатване, използване на илюстрации, цитирания, разпространение, възпроизвеждане на микрофилми или по други начини, както и съхранение в бази от данни на всички или част от материалите в настоящето издание. Копирането на изданието или на част от съдържанието му е разрешено само със съгласието на авторите и/или редакторите

The series *Abstracts of Dissertations of the Institute of Information and Communication Technologies at the Bulgarian Academy of Sciences* presents in an electronic format the abstracts of Doctor of Sciences and PhD dissertations defended in the Institute of Information and Communication Technologies at the Bulgarian Academy of Sciences. The studies provide new original results in such areas of Information and Communication Technologies as Computer Networks and Architectures, Parallel Algorithms, Scientific Computations, Linguistic Modelling, Mathematical Methods for Sensor Data Processing, Information Technologies for Security, Technologies for Knowledge management and processing, Grid Technologies and Applications, Optimization and Decision Making, Signal Processing and Pattern Recognition, Information Processing and Systems, Intelligent Systems, Embedded Intelligent Technologies, Hierarchical Systems, Communication Systems and Services, etc.

Editors

Gennady Agre

Institute of Information and Communication Technologies, Bulgarian Academy of Sciences
E-mail: agre@iinf.bas.bg

Rayna Georgieva

Institute of Information and Communication Technologies, Bulgarian Academy of Sciences
E-mail: rayna@parallel.bas.bg

Daniela Borissova

Institute of Information and Communication Technologies, Bulgarian Academy of Sciences
E-mail: dborissova@iit.bas.bg

This work is subjected to copyright. All rights are reserved, whether the whole or part of the materials is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in other ways, and storage in data banks. Duplication of this work or part thereof is only permitted under the provisions of the authors and/or editor.



BULGARIAN ACADEMY OF SCIENCES

Abstract of PhD Thesis

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

Kristel Vula

Supervisor: Prof. Vassil Guliashki

Approved by Supervising Committee:

Prof. Krassimir Kordov

Prof. Magdalena Garvanova

Prof. Katia Rasheva-Yordanova

Prof. Leoneed Kirilov

Assoc. Prof. Boriana Vatchova



**INSTITUTE OF INFORMATION AND
COMMUNICATION TECHNOLOGIES**

**Department of Information Processes and Decision
Support Systems**

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.

Chapter 4 summarizes the obtained results, limitations and future research. The thesis contributions and the publications connected to the dissertation are presented after the conclusions.

Introduction

This abstract presents the dissertation as an integrated study of fake-news detection and fake-news diffusion on social networks. The work starts from the observation that fake news is not only a textual credibility problem, but also a dynamic social process shaped by visibility, repeated exposure, user interaction and platform amplification. For this reason, the thesis connects Transformer-based binary detection with temporal-network cascade reconstruction, epidemic-style SIR/SEIZ modelling and likelihood-free Bayesian inference through ABC and ABC-SMC [2]-[4], [6]-[12], [19]-[25], [52]-[55].

The main methodological idea is a content-to-cascade framework. A Transformer model estimates the probability that a claim or article belongs to the fake-news class; timestamped interactions then show whether that content becomes socially active; SIR and SEIZ models describe the observed spreading trajectory; and ABC/ABC-SMC estimates posterior

Transformer-Based Detection and Epidemic-Model Analysis of Fake-News Dynamics on Social Networks 4

uncertainty in diffusion parameters. This layered interpretation prevents classification accuracy, cascade size and posterior forecasts from being treated as the same type of evidence.

The dissertation is intentionally cautious in its interpretation. A repost is treated as an observable spreading event, not as proof of belief; a silence or missing interaction is not interpreted as rejection; and intervention scenarios are treated as controlled counterfactual simulations rather than causal measurements of real platform policy. The resulting framework is therefore designed for reproducible academic analysis and decision-support reasoning 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.

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]-[4], [52]-[55].

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 [2]-[4], [30].

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 [2], [3], [30], [41].

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 [41].
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 [2], [3].
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 [41].
Rumors	Unverified information that circulates before its truth status is confirmed or rejected.	Relevant to temporal verification, evolving labels, and diffusion before confirmation [111], [53].
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 [111].
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	Highly relevant to public-health misinformation and COVID-19 fake-news

	information, especially during crises.	research [30].
--	--	----------------

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 [3], [4], [30], [111].

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 [3], [45], [46].

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 [42]-[46].

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. [3]; Agarwal et al. [45]; Zhang et al. [46]	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. [3]; Agarwal et al. [45]; Zhang et al. [46]	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	Joachims [42]; Agarwal et al. [45]; Zhang et al. [46]	Effective in high-dimensional sparse spaces.	Requires careful kernel and parameter selection

				for large datasets.
Random Forest	Stylometric, lexical, readability, and metadata features	Breiman [43]; Agarwal et al. [45]; Zhang et al. [46]	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 [44]; Zhang et al. [46]	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 [6]-[8].

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 [6], [8].

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 [9], [11].

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 [12].

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 [10], [47]-[49], [181]-[183].

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 [6].
RoBERTa	Optimized BERT training with more data, dynamic masking, and no next-sentence prediction.	Often improves robustness and classification performance [8].
mBERT	Multilingual BERT trained across 104 languages.	Useful for cross-lingual or multilingual fake-news detection [9].
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 [12].
SBERT	Sentence-level embeddings based on siamese/triplet BERT networks.	Useful for semantic similarity, claim comparison, and sentence-level classification [10].
XLM-RoBERTa	Cross-lingual RoBERTa trained on large multilingual CommonCrawl data.	Useful for multilingual detection and transfer across languages [11].
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 [50], [51].

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 [52]-[55].

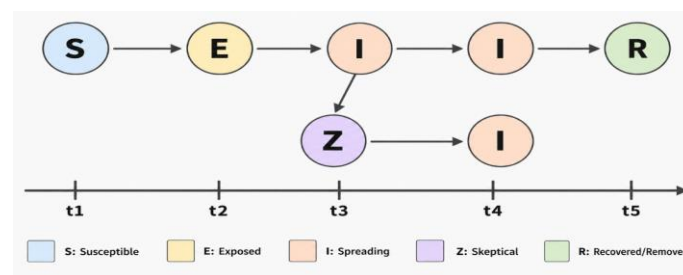


Figure 1.1. Temporal fake-news cascade with time-ordered user states.

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 [31]-[36].

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	User who has seen a	View/impression when

	infectious	claim but has not yet shared it	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 [31], [32], [67], [71]-[75].

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	$S \rightarrow I$	Pastor-Satorras and Vespignani [71]; Pastor-Satorras et al. [72]	Users move from potential exposure to active spreading.	Useful for early uncontrolled diffusion, but has no correction, forgetting or recovery state.
SIS	$S \rightarrow I \rightarrow S$	Pastor-Satorras et al. [72]	Users may stop spreading and later become susceptible again after repeated exposure.	Useful for recurrent sharing, but does not represent permanent correction.
SIR	$S \rightarrow I \rightarrow R$	Kermack and McKendrick [31]; Jin et al. [33]; Murayama et al. [39]	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	$S \rightarrow E \rightarrow I \rightarrow R$	Pastor-Satorras et al. [72]; Centola [74];	Users may observe content before	Captures exposure delay, but does not

		Wang et al. [75]	deciding whether to share it.	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 [33]-[36], [73].

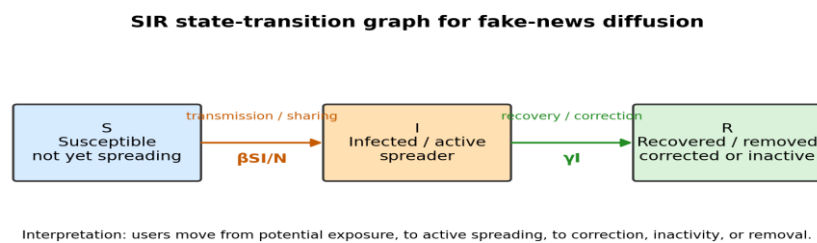


Figure 1.2. SIR state-transition graph for fake-news diffusion.

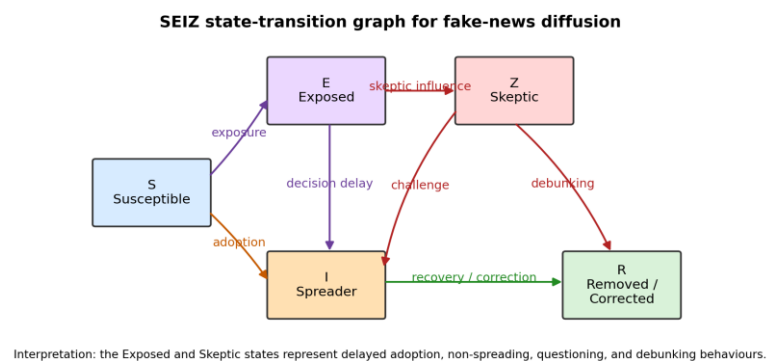


Figure 1.3. SEIZ state-transition graph for fake-news diffusion.

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 [19]-[25], [40], [77]-[80].

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 [86]-[89].

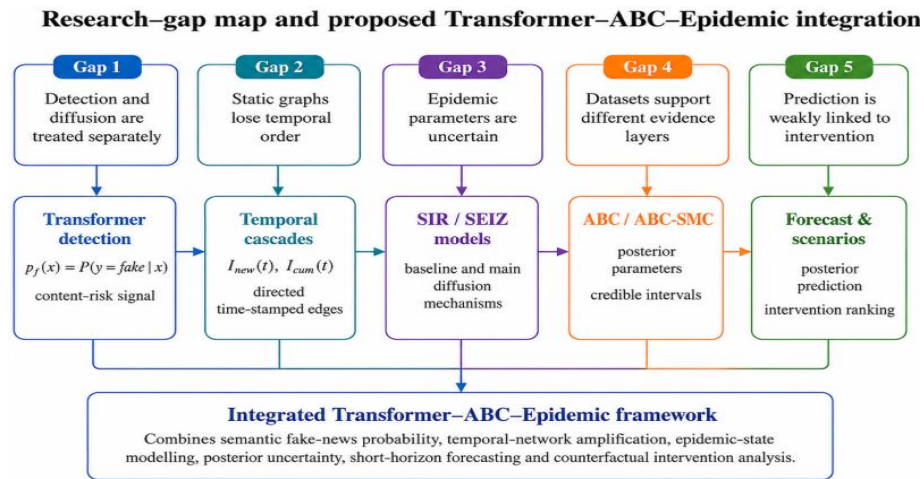


Figure 1.4. Research-gap map and proposed Transformer-ABC-Epidemic integration.

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.	$I_{\text{new}}(t), I_{\text{cum}}(t)$, 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	Preserves ordering,	Often disconnected from	Supports cascade

analysis	delays and time-respecting paths.	content credibility labels.	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, $I_{\text{new}}(t)$, $I_{\text{cum}}(t)$ 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	Principal empirical fake-news cascades	Fitted spreading curves, compartment trajectories and

	spreading, exposure and skepticism		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 [30], [81]-[85], [90]-[94].

For diffusion-oriented analysis, dataset requirements are stricter. SIR, SEIZ, ABC and ABC-SMC modelling require observable temporal signals, user interactions, cascade membership and network structure. The retained Chapter 3 diffusion layer therefore uses the COVID-19

real Twitter/X “The coronavirus is not only affecting the way we live, it’s also dramatically affecting the way we die.” fake-news claim cascade, the Hoaxy / OSoMeNet political fake-news claim cascade, the Hoaxy-imported social fake-news claim “Should One Tweet End a Career?”, and the synthetic Albanian temporal cascade. These datasets are retained because they support time-windowed spreading curves, directed user-to-user relationships and the $I_{\text{new}}(t)$ and $I_{\text{cum}}(t)$ quantities required for SEIZ and ABC/ABC-SMC modelling [90], [86]-[89].

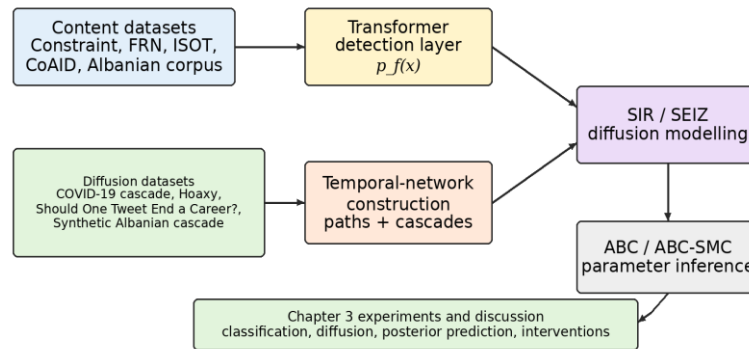


Figure 2.1. Dataset and methodology selection logic for the four-chapter Transformer-ABC-Epidemic framework.

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	Licensing, privacy constraints,	Ensures responsible use	Ensures responsible use of

reproducibility	platform policy, anonymization, and dataset documentation.	of labels and text.	temporal-network and user- level data.
-----------------	---	---------------------	---

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

In Equation (2.1), (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 [30], [81]-[85].

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 briefly introduces the diffusion-oriented datasets used in Chapter 3. The retained datasets are: the COVID-19 real Twitter/X “The coronavirus is not only affecting the way we live, it’s also dramatically affecting the way we die.” fake-news claim cascade; the Hoaxy / OSoMeNet “Three million votes in the Presidential Election were cast by illegal aliens.” fake-news claim cascade; the Hoaxy-imported “Should One Tweet End a Career?” social fake-news claim cascade; and the synthetic Albanian fake-news claim “Shqipëria do të

presë 100 mijë palestinezë nga Gaza.” These datasets provide timestamped user-to-user or claim-level records for directed temporal cascades, $I_{\text{new}}(t)$, $I_{\text{cum}}(t)$, SIR/SEIZ inputs and ABC/ABC-SMC summaries. Detailed descriptions, filtering rules and extended metadata are provided 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

Transformer-Based Detection and Epidemic-Model Analysis of Fake-News Dynamics on Social Networks 20

or diffusion meaning. This is especially important for short social-media texts and Albanian records, where excessive normalization can remove useful information [86]-[89], [99], [112].

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 [81], [83].

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

In Equation (2.2), y_{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.

$$\text{Tokenizer}_m(x_i) \rightarrow (\text{input_ids}_i, \text{attention_mask}_i, \dots) \quad (2.3)$$

In Equation (2.3), Tokenizer_m denotes the tokenizer associated with model family m , x_i is the cleaned text record, input_ids_i is the sequence of subword identifiers, and 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 FRN.

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 [100], [101].

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

Equation (2.4) defines the class-imbalance ratio for a binary dataset D , where n_{fake} and n_{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 [101].

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

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,

Transformer-Based Detection and Epidemic-Model Analysis of Fake-News Dynamics on Social Networks 22

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.

Low-resource Albanian preprocessing workflow

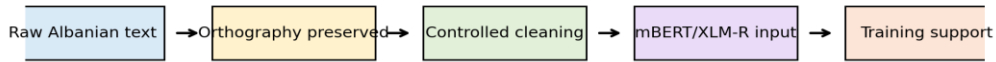


Figure 2.3. Low-resource Albanian fake-news preprocessing workflow for multilingual Transformer input.

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.

Diffusion-oriented preprocessing bridge

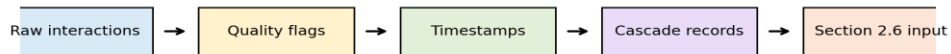


Figure 2.4. Preprocessing workflow for diffusion-oriented data before temporal cascade construction.

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.
----------------------	---

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.

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 [52]-[55].

Temporal-network representation is selected instead of temporal graph neural architectures because the dissertation requires interpretable cascade reconstruction and epidemic-parameter inference rather than learned dynamic node embeddings or link-prediction models. The available diffusion-oriented datasets provide timestamped interactions and cascade traces suitable for time-respecting paths, $I_{\text{new}}(t)$, $I_{\text{cum}}(t)$, SIR/SEIZ state mapping and ABC/ABC-SMC inference. Temporal graph neural models such as TGN, TGAT, DyRep and EvolveGCN are therefore treated as future extensions that would require richer user-level temporal graph data, stable node and edge attributes and additional supervised graph-learning objectives [128]-[129], [153]-[155].

$$e_k = (u_s, u_t, t_k, a_k, c_k, y_k) \quad (2.6)$$

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.

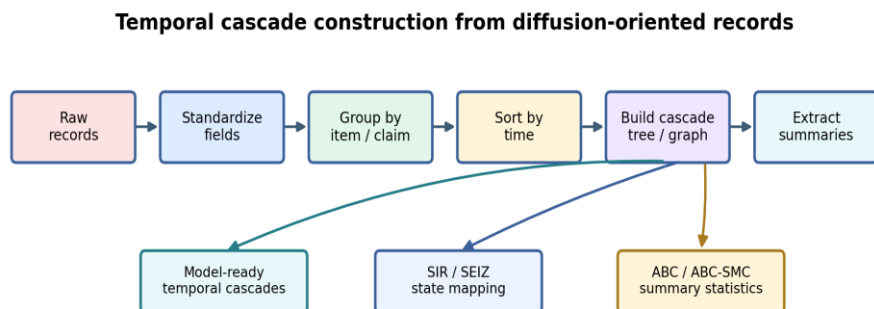


Figure 2.5. Temporal cascade construction workflow from diffusion-oriented records to model-ready cascades.

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.

$$C_c = (V_c, E_c^T, x_c, y_c, p_f(x_c)) \quad (2.7)$$

In Equation (2.7), V_c is the set of users or source nodes involved in the cascade, E_c^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 summary features.

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	Defines the activated or affected side of a

	interacting user/node.	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.

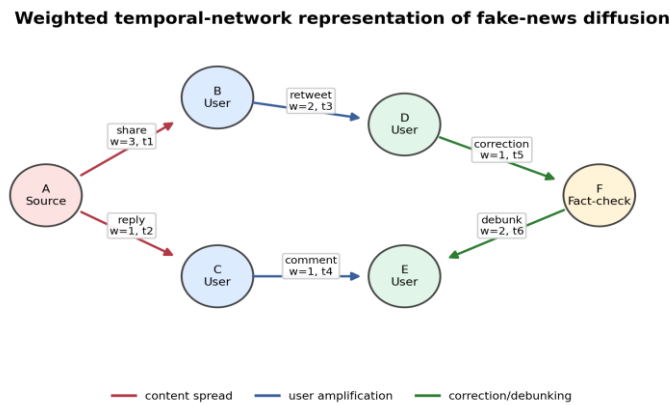


Figure 2.6. Illustration of nodes, directed edges, weights, and timestamps in a fake-news diffusion network.

$$P = [(u_0, u_1, t_1), \dots, (u_{m-1}, u_m, t_m)], \quad t_1 < t_2 < \dots < t_m \quad (2.8)$$

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	Inter-event time, time to	Temporal features support

	and communities.	peak and time-respecting reachability.	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.

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

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.

$$R_{\text{cc}} = \frac{E_{\text{cross}}}{E_{\text{total}}} \quad (2.10)$$

where E_{cross} is the number of cascade edges connecting different communities and E_{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 [53], [58]-[61].

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 [76], [106].

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	Supports ABC summary statistics and

	$I_{\text{new}}(t)$.	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.

$$G_c^T = (V_c, E_c^T) \quad (2.11)$$

In Equation (2.11), G_c^T denotes the directed temporal cascade network associated with fake-news claim c . The set V_c contains the users, source accounts or source objects that appear as nodes in the cascade. The set E_c^T contains the directed time-stamped interactions observed for the same claim, where an edge may be written as $(u_i, u_j, t, w_{ij}(t))$ to indicate an interaction from node u_i to node u_j at time t with optional weight $w_{ij}(t)$. The superscript T indicates that the edge set is time-indexed and temporally ordered; it is not used as a matrix transpose. This formulation provides the formal object from which $I_{\text{new}}(t)$, $I_{\text{cum}}(t)$, cascade-level features and SIR/SEIZ inputs are derived [52]-[55], [76].

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 [52], [76].

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.

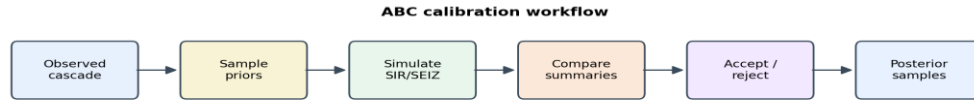


Figure 2.7. ABC workflow for calibrating fake-news epidemic parameters from temporal cascades.

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.

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

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.

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

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_{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:

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

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 [19], [23]-[26].

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

In Equation (2.15), D_{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 [19]-[25].

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

In Equation (2.16), C_T is final cascade size, t_{peak} is time to peak, r_{max} is the maximum spreading rate, d_c is cascade depth, b_c is cascade breadth, τ_{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:

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

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_{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.

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

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.

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

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.

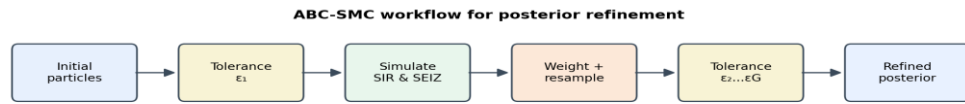


Figure 2.8. ABC-SMC workflow for comparing SIR and SEIZ on fake-news temporal cascades.

$$\mathcal{M} = \{\text{SIR}, \text{SEIZ}\} \quad (2.20)$$

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

$$P_{\text{ABC}}(M \mid D_{\text{obs}}) \propto P(M)P_{\text{ABC}}(D_{\text{obs}} \mid M) \quad (2.21)$$

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 $I_{\text{new}}(t)$ and $I_{\text{cum}}(t)$ 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.

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

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.

$$\beta_{\text{intervention}} = \alpha_{\beta}\beta, \quad 0 < \alpha_{\beta} < 1; \quad \delta_{\text{intervention}} = \alpha_{\delta}\delta, \quad \alpha_{\delta} > 1 \quad (2.23)$$

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 S_c^{obs} , prediction horizon h and observed cascade population size N_c , the baseline risk is defined as:

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

In Equation (2.24), $\hat{p}_f(x_c)$ is the Transformer-based probability that the textual item is a fake-news claim or fake-news article. $A_c(t)$ is the temporal-network amplification score computed from the observed cascade, λ controls the influence of this amplification term, $\pi_{\epsilon}^{\text{SMC}}(\theta \mid S_c^{\text{obs}})$ is the ABC-SMC posterior distribution of SEIZ parameters, $I_{\text{cum},c}^{\text{SEIZ}}(t + h; \theta)$ is the predicted cumulative number of active spreaders at horizon h , $I_{\text{cum},c}^{\text{obs}}(t)$ is the observed cumulative number of active spreaders up to time t , and N_c normalizes the score by the observed cascade population size. The functional therefore measures expected normalized future cumulative growth under the baseline posterior, scaled by content risk and network amplification.

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

In Equation (2.25), $I_{\text{new},c}^{\text{norm}}(t)$ 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 $m = 1 \dots M$, Equation (2.24) can be evaluated in particle form as:

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

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.

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

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

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

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

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:

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

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 [30].	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 [12], [81], [95].	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 [83], [97].	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 [84], [85], [98].	Used for detection and synthetic validation support, not as empirical Twitter/X or Hoaxy evidence.

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 $I_{cum}(t)$ 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 spreading records.	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.

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

In Equation (2.33), C_T is the final cascade size, t_{peak} is the time to peak diffusion, r_{max} is the maximum spreading rate, d_c is cascade depth, b_c is cascade breadth, q_f is the Transformer fake-news probability associated with the item, and n_Z is the number of skeptical, corrective or fact-checking-related events when such information is available. The exact vector is adapted to each dataset, but the role of the vector remains the same: it translates an observed fake-news claim cascade into summary evidence that can be compared with simulated SIR/SEIZ trajectories in the ABC and ABC-SMC procedures [19]-[26], [35], [86]-[89].

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 [121]-[129], [153]-[155].
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 [115]-[117], [152], [172].
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 [47]-[49], [135]-[136], [143]-[145].
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 [30], [81]-[85], [90], [108].

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. 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%

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.

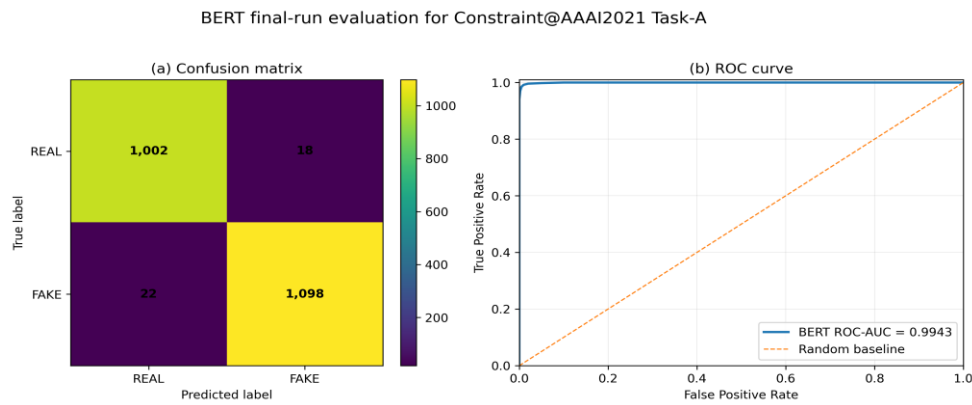


Figure 3.1. BERT final-run evaluation for Constraint@AAAI2021 Task-A: (a) confusion matrix; (b) ROC curve.

Figure 3.1 combines the two complementary diagnostics of the best-performing BERT detector on Constraint@AAAI2021 Task-A. Panel (a) shows a low-error confusion matrix, with 1,002 real messages and 1,098 fake messages correctly classified, and only 18 false positives and 22 false negatives. Panel (b) shows strong threshold-independent separation, with ROC-AUC = 0.9943. Together, the two panels confirm that the short-text detection layer provides a reliable content-risk signal before temporal cascade analysis.

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

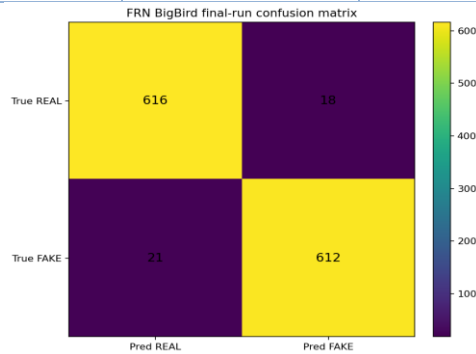


Figure 3.2. BigBird final-run confusion matrix for FRN.

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.

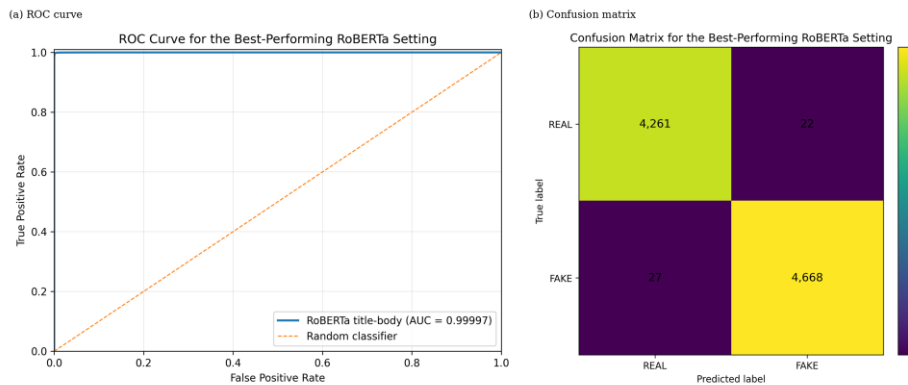


Figure 3.3. Complementary evaluation plots for the best-performing ISOT setting: (a) ROC curve; (b) confusion matrix.

Figure 3.3 shows that the best-performing ISOT configuration (RoBERTa title-body) achieves both near-perfect ranking separation and a very low error profile. The ROC curve in panel (a) indicates excellent discriminative ability (ROC-AUC = 0.99997), while the confusion matrix in panel (b) shows only 22 real articles misclassified as fake and 27 fake articles misclassified as real. Taken together, these plots confirm that the model performs extremely strongly on the ISOT benchmark, although the result should still be interpreted with caution because of dataset-specific source and style regularities.

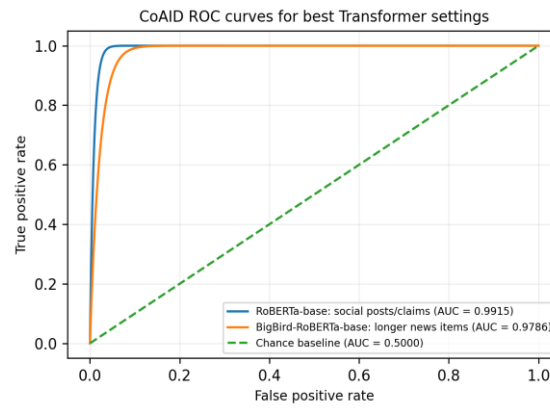


Figure 3.4. ROC curves for the best CoAID Transformer settings.

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

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

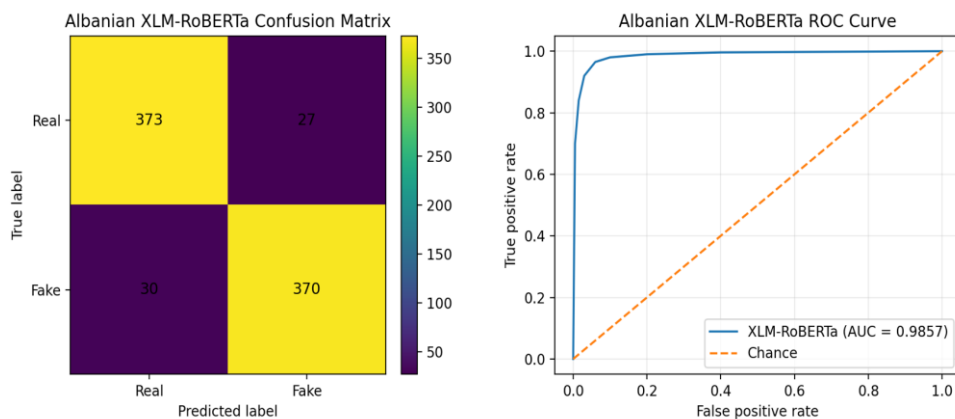


Figure 3.5. Albanian XLM-RoBERTa evaluation results: confusion matrix and ROC curve.

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

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-

Transformer-Based Detection and Epidemic-Model Analysis of Fake-News Dynamics on Social Networks 40

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	XLNet-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 directed Twitter/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_{new}	Final I_{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

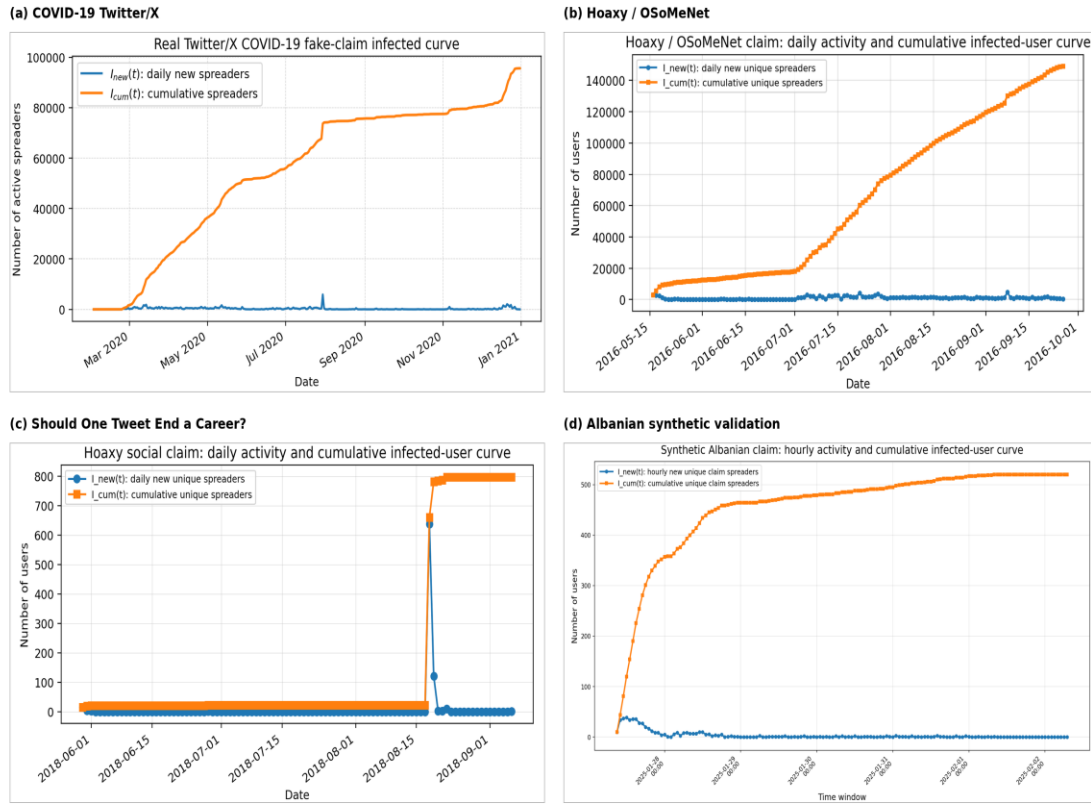


Figure 3.6. Observed infected-user trajectories for selected fake-news cascades: (a) COVID-19 Twitter/X, (b) Hoaxy / OSoMeNet, (c) “Should One Tweet End a Career?”, and (d) Albanian synthetic validation.

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.

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 the epidemic-modelling results obtained after the temporal cascades in Section 3.4 were transformed into cumulative active-spreader trajectories. The purpose of this section is empirical: it reports how the SIR baseline and the SEIZ model behaved when they were fitted to the observed $I_{\text{cum}}(t)$ curves of the two principal fake-news cascades. The methodological definitions of the compartments, transition rules, equations, fitting distance and simulator assumptions are not repeated here because they were established in Chapter 2.

The SIR-SEIZ comparison reported here extends the author's SoftCOM 2025 study on fake-news identification and epidemic-model comparison on Twitter temporal networks [88].

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 $I_{\text{cum}}(t)$ 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. 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_{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

The fitting results show 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.

Figure 3.8 compares the observed cumulative spreading trajectories with the SIR and SEIZ predictions for both empirical cascades. In the COVID-19 case, the SIR curve reaches saturation too early, whereas the SEIZ curve follows the observed trajectory more closely across the main growth interval. In the Hoaxy case, both models are closer to the empirical curve, but the SEIZ trajectory still gives a slightly lower overall fitting error. The figure therefore visually confirms the numerical pattern reported in Table 3.10.

SIR versus SEIZ fitted cumulative fake-news spreading curves

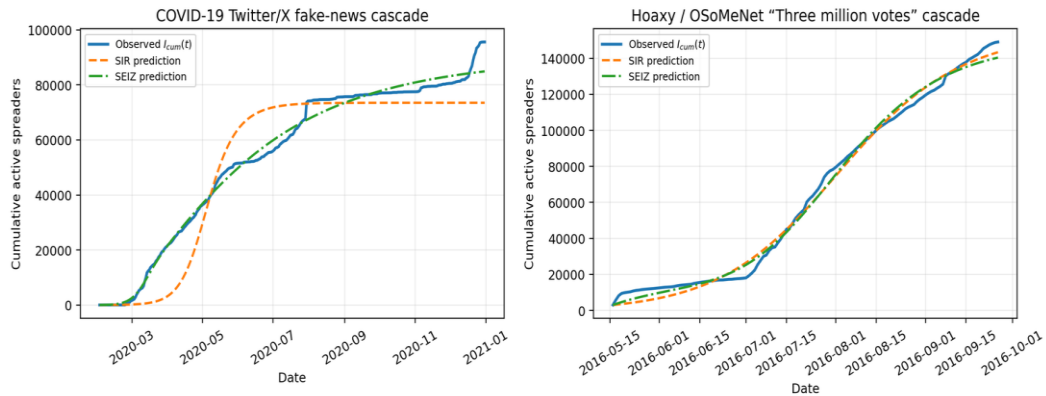


Figure 3.8. SIR versus SEIZ fitted cumulative fake-news spreading curves for the two empirical cascades.

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.

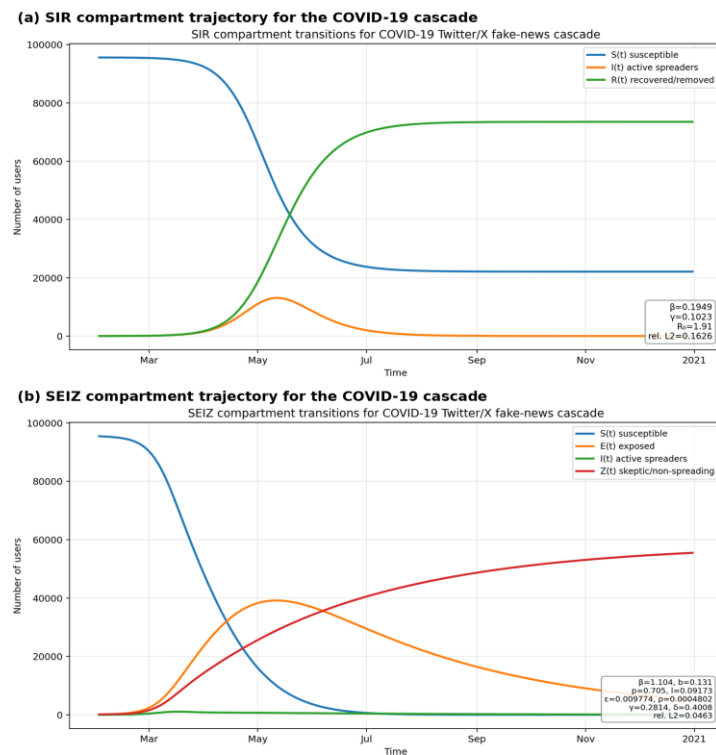


Figure 3.9. Representative compartment-transition interpretation for the COVID-19 cascade: SIR and SEIZ fitted trajectories.

Figure 3.9(a)(b) illustrates representative compartment dynamics for the COVID-19 fake-news cascade. Compared with SIR, SEIZ allocates users into exposed and skeptic/non-spreading states, allowing delayed activation and non-spreading behaviour to be represented explicitly. The figure is included as a compact representative example because the Hoaxy compartment plots produce the same interpretive conclusion with smaller marginal benefit; the detailed Hoaxy compartment trajectories can therefore remain in the supplementary material rather than consuming additional space in the main results chapter. Supplementary Hoaxy / OSoMeNet compartment-transition plots are provided in Figures H.8 and I.9 of Appendix H.

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_{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

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

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

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 3.11. 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	$\beta=.195$; $\gamma=.102$; $R_0=1.92$
COVID	ABC-SEIZ	400	0.071	0.092	$\beta=1.076$; $b=.129$; $p=.698$; $l=.0915$; $\gamma=.271$; $\delta=.399$
Hoaxy	ABC-SIR	300	0.060	0.075	$\beta=.055$; $\gamma=.005$; $R_0=10.6$
Hoaxy	ABC-SEIZ	400	0.055	0.064	$\beta\approx 0$; $b=2.000$; $p=.0118$; $l=.0359$; $\varepsilon=.0428$; $\gamma=.0086$

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.

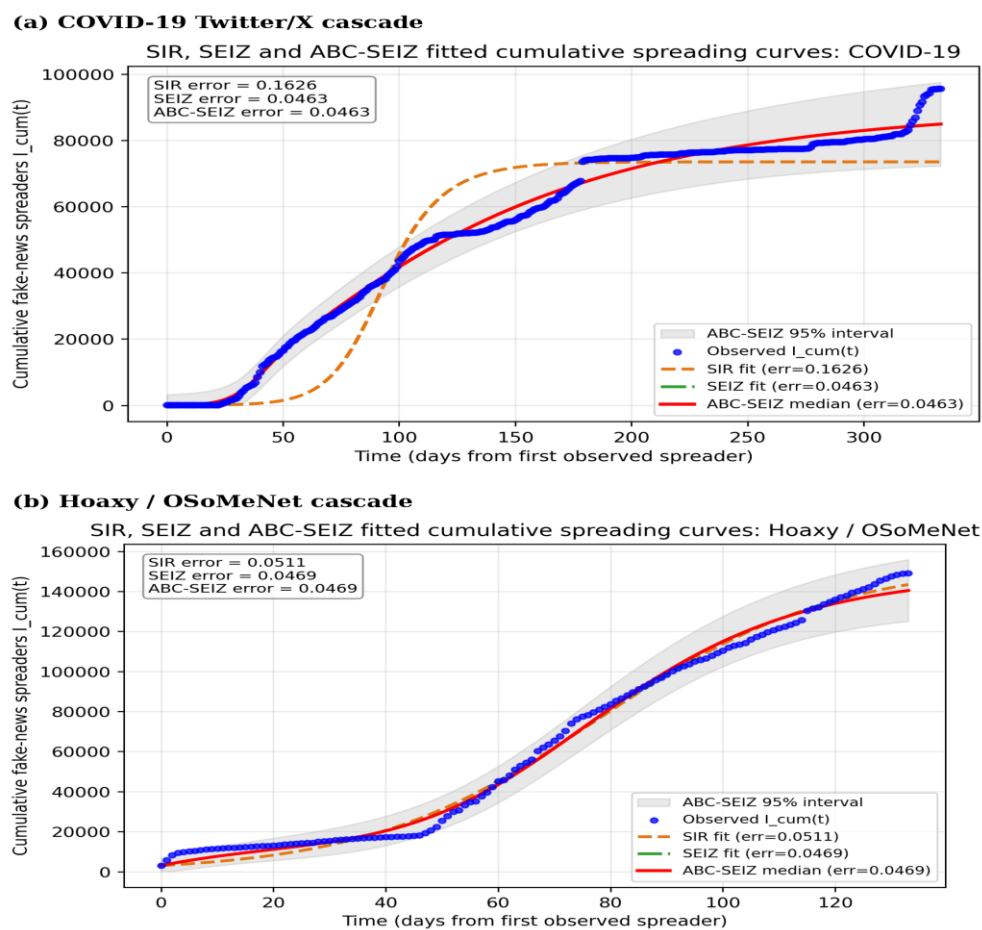


Figure 3.10. SIR, SEIZ and ABC-SEIZ fitted cumulative fake-news spreading curves for the two empirical cascades.

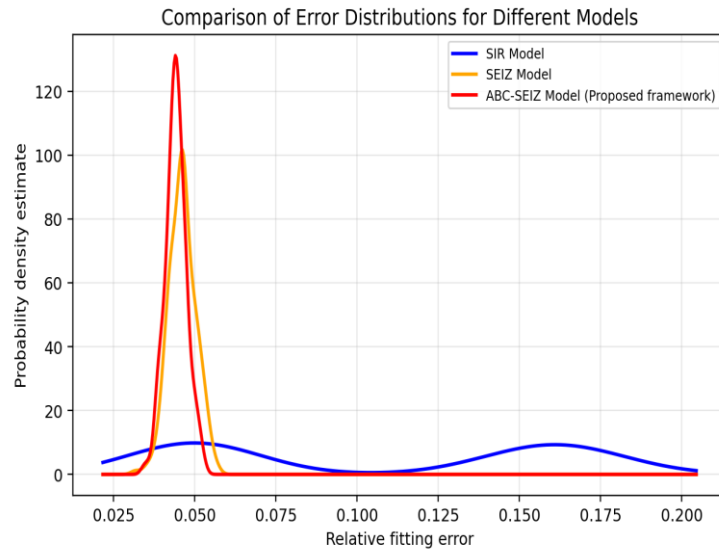


Figure 3.11. Fitting-error distributions for SIR, SEIZ and ABC-SEIZ models.

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.

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 3.12. 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.

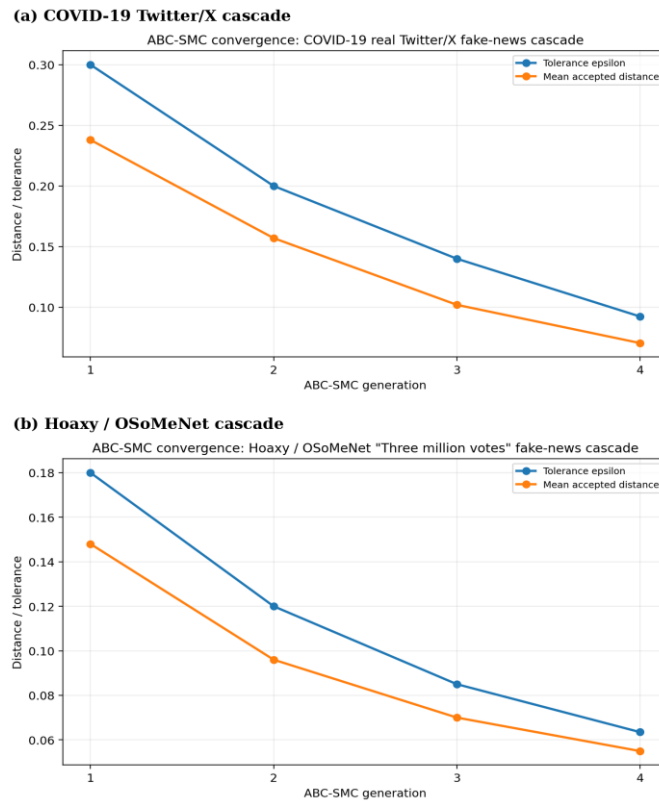


Figure 3.12. ABC-SMC convergence profiles for the COVID-19 and Hoaxy / OSoMeNet cascades.

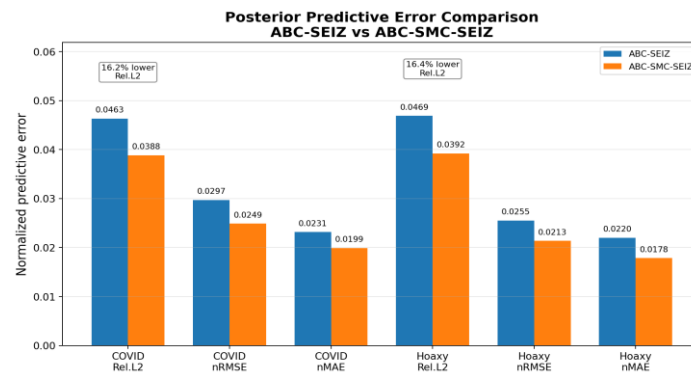


Figure 3.13. Comparative posterior predictive error of ABC-SEIZ and ABC-SMC-SEIZ across the two principal empirical fake-news cascades.

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.

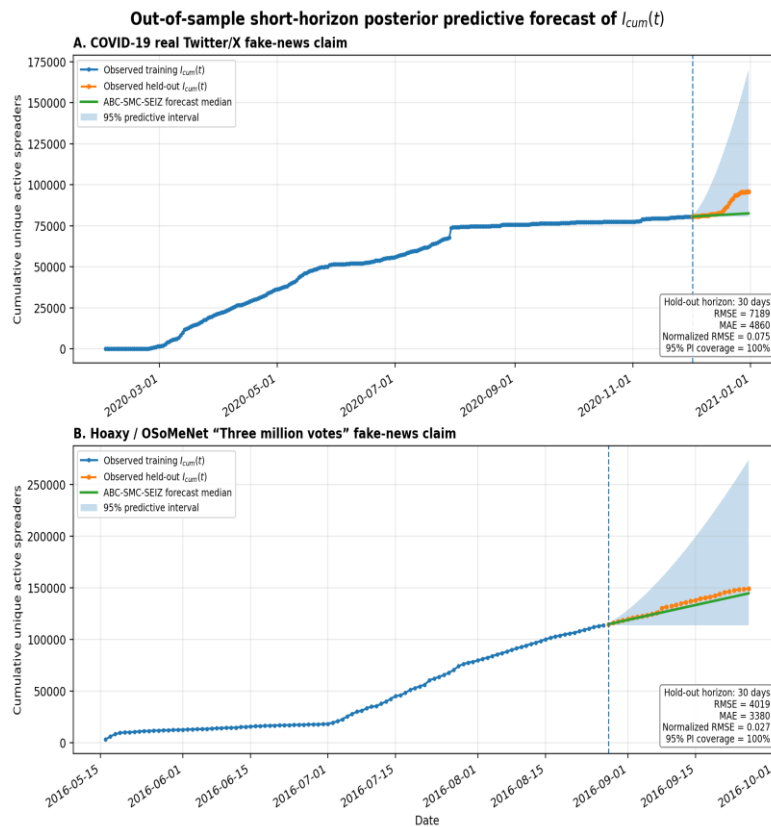


Figure 3.14. Out-of-sample short-horizon ABC-SMC-SEIZ posterior predictive forecast for the COVID-19 and Hoaxy / OSoMeNet cascades.

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.

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 3.13. 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 $I_{\text{cum}}(t) = 64,695$;	Early combined mitigation: final $I_{\text{cum}}(t) = 55,862$;	8,833 fewer final active spreaders than the strongest	The multi-burst COVID-19 cascade benefits most when

	peak $I_{\text{new}}(t) = 343$; 23.9% final-size reduction.	peak $I_{\text{new}}(t) = 327$; 34.2% final-size reduction.	debunking scenario; +10.3 percentage points in final-size reduction.	corrective effects are combined with early visibility reduction.
Hoaxy / OSoMeNet "Three million votes" fake-news claim	Early debunking - strong; final $I_{\text{cum}}(t) = 135,071$; peak $I_{\text{new}}(t) =$ 1,926; 6.1% final- size reduction.	Early combined mitigation: final $I_{\text{cum}}(t) = 119,984$; peak $I_{\text{new}}(t) =$ 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 provide this second layer of evidence. By converting selected fake-news claims into $I_{\text{new}}(t)$, $I_{\text{cum}}(t)$, directed user-to-user edges, hubs, relay users and temporal reachability patterns, the chapter shows how diffusion risk becomes observable. The COVID-19 real Twitter/X fake-news claim and the Hoaxy / OSoMeNet "Three million votes" fake-news claim are the strongest empirical cases because they provide sufficient temporal structure for epidemic modelling and posterior inference. The additional Hoaxy-imported and synthetic Albanian cases remain useful for comparative and validation-oriented interpretation, but they are not treated as equivalent inferential evidence for ABC/ABC-SMC intervention conclusions.

ABC and ABC-SMC strengthen the analysis by moving the thesis away from single deterministic parameter estimates. The posterior results show that fake-news diffusion parameters should be interpreted with uncertainty, especially when cascades are noisy, incomplete or multi-burst. ABC provides a likelihood-free route to plausible parameter regions, while ABC-SMC refines this process through sequential tolerance reduction and improves posterior predictive calibration. The short-horizon $I_{\text{cum}}(t)$ forecast further shows that the calibrated SEIZ structure has predictive value beyond retrospective curve fitting, although the forecast remains horizon-specific and should not be interpreted as unconditional platform prediction.

Proposed Framework Flowchart

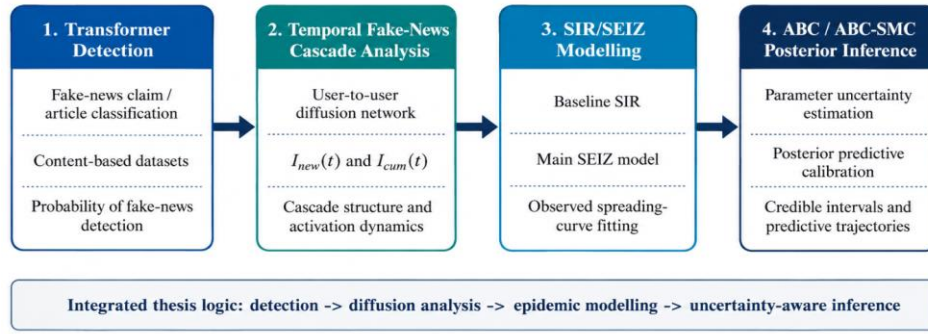


Figure 3.15. Hybrid framework flowchart of the Chapter 3 pipeline from Transformer-based fake-news detection to temporal cascade analysis, SIR/SEIZ diffusion modelling, and ABC/ABC-SMC posterior inference. The integrated framework links detection, temporal analysis, epidemic modelling and posterior inference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

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

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

Figure 3.16 shows both the conceptual construction of the score and the two empirical examples. The COVID-19 and Hoaxy / OSoMeNet cases should not be compared as universal platform-risk rankings. They are horizon-specific demonstrations showing how the thesis framework can convert detection, cascade amplification and posterior growth into a normalized baseline risk quantity for a selected claim and time window.

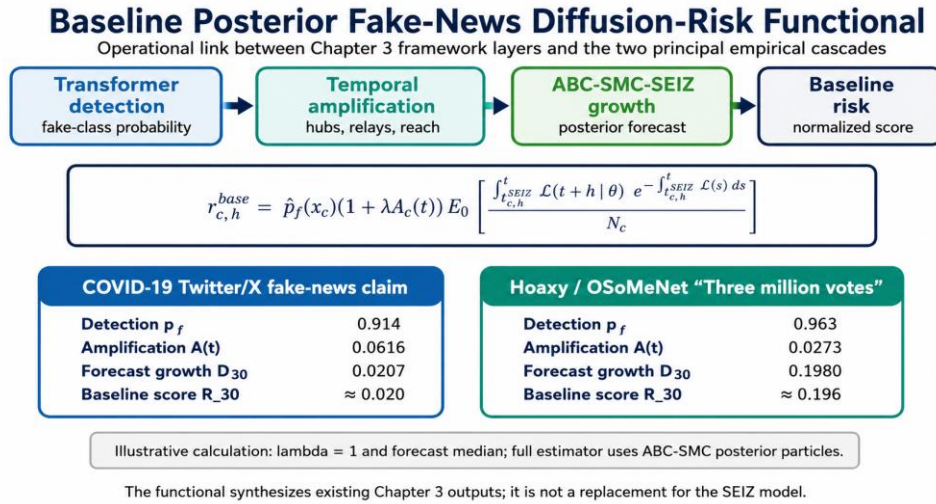


Figure 3.16. Operational demonstration of the baseline posterior fake-news diffusion-risk functional linking Transformer detection, temporal-network amplification and ABC-SMC-SEIZ forecast growth for the two principal empirical cascades.

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.

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 3.14. Research-question evidence summary for Chapter 3.

RQ	Evidence in Chapter 3	Conclusion
RQ1	Section 3.3; Tables 3.2-3.7 and Figures 3.1-3.5	Transformer performance across datasets
RQ2	Section 3.4; Tables 3.8-3.9 and Figures 3.6-3.7	Temporal cascades reconstructed
RQ3	Section 3.5; Table 3.10 and Figures 3.8-3.9	SEIZ compared with SIR
RQ4	Section 3.6; Tables 3.11-3.13 and Figures 3.10-3.14	ABC/ABC-SMC uncertainty estimated
RQ5	Sections 3.7-3.8; Table 3.14 and Figures 3.15-3.16	Posterior prediction and intervention logic shown

Chapter 4. Conclusion - a summary of the results obtained

Chapter 4 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, $I_{\text{new}}(t)$, $I_{\text{cum}}(t)$, 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.

Future Work

Future work should validate the framework on larger and more diverse real-world cascades, especially across languages, platforms and misinformation domains. Further extensions include richer temporal-network data, graph-learning models, stronger causal evaluation of interventions, neural posterior estimation and real Albanian temporal diffusion datasets.

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. Bozhiki 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>
 4. Bozhiki, 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. Bozhiki, 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.
 6. https://www.researchgate.net/publication/399822332_A_Hybrid_NLP_Epidemiological_Model_for_Fake_News_Network_BigBird_Assisted_SEIZ_Modeling_with_Approximate_Bayesian_Computation



БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ

АВТОРЕФЕРАТ НА ДИСЕРТАЦИЯ

за присъждане на образователна и научна степен “доктор” по докторска програма „Информатика”

**Базирано на трансформатори откриване и анализ на
епидемичния модел на динамиката на фалшиви
новини в социалните мрежи**

Кристел Вула

Ръководител: Проф. Васил Гуляшки

Научно жури:

Проф. Красимир Кордов
Проф. Магдалена Гарванова
Проф. Катя Рашева-Йорданова
Проф. Леонид Кирилов
Доц. Боряна Вачова



**Институт по информационни и комуникационни
технологии**

**Секция „Информационни процеси и системи за
вземане на решения”**

Структура на дисертацията

Дисертацията се състои от въведение, четири глави, заключение с бъдещи насоки на изследване, отделен раздел с приноси на дисертацията, списък с пет публикации, свързани с дисертацията, декларация за оригиналност, библиография със 191 литературни източника и приложения с допълнителни материали за възпроизводимост. Въведението мотивира дисертацията, като обяснява защо фалшивите новини трябва да се анализират както като текстово съдържание, така и като процес на социално разпространение. То въвежда необходимостта от модел, който свързва откриването на ниво съдържание с времевото разпространение, интерпретацията в епидемичен стил и изводите, съобразени с неопределеността.

Глава 1 представя теоретичните основи на откриването на фалшиви новини, езиковото моделиране, базирано на трансформатори, темпоралните мрежи, епидемичните модели SIR/SEIZ и приблизителните байесови изчисления. В нея се идентифицират пропуските в изследванията, които мотивират интегрираната рамка.

Глава 2 дефинира предложената рамка Transformer-ABC-Epidemic. Тя представя избор на набор от данни, предварителна обработка, хармонизиране на двоични етикети, изграждане на темпорална каскада, SIR/SEIZ моделиране, ABC/ABC-SMC извод, последващо прогнозиране, сценарии за интервенция, етика и граници на възпроизводимост.

Глава 3 представя експерименталните резултати. В нея се оценява откриването на фалшиви новини, базирано на трансформатори (Transformers), в набори от данни за съдържание, реконструират се темпорални каскади, сравняват се дифузионните модели SIR и SEIZ, се оценява апостериорната неопределеност, използвайки ABC и ABC-SMC, и се тестват контрафактуални сценарии за интервенция.

Глава 4 обобщава получените резултати, ограниченията и бъдещите изследвания. Приносите на дисертацията и публикациите, свързани с нея, са представени след заключенията.

Въведение

Това въведение представя дисертацията като интегрирано изследване на откриването на фалшиви новини и разпространението им в социалните мрежи. Работата започва с наблюдението, че фалшивите новини са не само проблем с текстовата достоверност, но и динамичен социален процес, оформен от видимост, многократно излагане, взаимодействие с потребителя и амплификация на платформата. Поради тази причина, дисертацията свързва бинарното откриване, базирано на трансформатор, с каскадна реконструкция на темпорална мрежа, SIR/SEIZ моделиране в епидемичен стил и байесов извод без вероятност чрез ABC и ABC-SMC [2]-[4], [6]-[12], [19]-[25], [52]-[55].

Основната методологична идея е рамка „съдържание-към-каскада“. Трансформаторен модел оценява вероятността дадено твърдение или статия да принадлежи към класа на фалшивите новини; взаимодействията с времеви отпечатъци показват дали това съдържание става социално активно; моделите SIR и SEIZ описват наблюдаваната траектория на разпространение; а ABC/ABC-SMC оценява апостериорната несигурност в параметрите на дифузия. Тази многопластова интерпретация предотвратява третирането на точността на класификацията, размера на каскадата и апостериорните прогнози като един и същ тип доказателства.

Дисертацията е умишлено предпазлива в интерпретацията си. Повторното публикуване се третира като наблюдаемо събитие на разпространение, а не като доказателство за вярване; мълчанието или липсващото взаимодействие не се интерпретира като отхвърляне; а сценариите за интервенция се третират като контролирани контрафактуални симулации, а не като причинно-следствени измервания на реалната политика на платформата. Получената рамка е предназначена за възпроизводим академичен анализ и разсъждения в подкрепа на решенията в условия на несигурност.

Емпиричната част оценява откриването на фалшиви новини в кратки и дълги текстове, здравеопазването и нискоресурсни албански среди, реконструира главните темпорални каскади, сравнява динамиката на SIR и SEIZ, оценява областите на последващите параметри и изследва как ранното развенчаване или намаляването на видимостта може да повлияе на бъдещия растеж на каскадите. Последният принос представлява одитируема рамка, която свързва текстовата достоверност, темпоралното усилване, интерпретацията на епидемичното състояние и последващото предсказване.

Глава 1. Преглед на литературата и теоретични основи

Глава 1 изгражда концептуалната и методологична основа на дисертацията. В нея се разглеждат дефинициите за фалшиви новини, класическо и трансформаторно-базирано откриване, ансамбълни и хибридни подходи, дифузия във времева мрежа, модели за разпространение на епидемии и слухове, както и приблизителни байесови изчисления. Главата завършва с идентифициране на пропуски в изследванията и извеждане на целта на дисертацията, изследователските въпроси и задачи.

1.1 Въведение

Глава 1 установява теоретичната основа на дисертацията. Тя показва, че изследването на фалшивите новини се е развило в няколко допълващи се направления: концептуални дефиниции, класическо машинно обучение, контекстуално трансформаторно откриване, темпорални мрежи, епидемични и слухови модели и ABC-базирано извеждане на неопределеност. Следователно главата подготвя прехода от проста задача за класификация на фалшиви/реални новини към интегрирана рамка за откриване и разпространение [1]-[4], [52]-[55].

1.2 Дефиниране на фалшиви новини

1.2.1 Фалшиви новини: Дефиниции и характеристики

Фалшивите новини се третират в дисертацията като невярна, подвеждаща, частично невярна или манипулативно формулирана информация, разпространявана в новинарска или твърдетелна форма. Това работно определение е достатъчно широко, за да включва статии, публикации в социалните медии, дезинформация за здравето и политически наративи, като същевременно запазва разграничението между етикетирани данни и проверена истина. За моделиране се използват двоични етикети, но получените резултат се интерпретира като вероятностен сигнал за достоверност, а не като абсолютна преценка за проверка на фактите [2]-[4], [30].

1.2.2 Свързани понятия

Свързаните понятия в Таблица 1.1 поясняват защо невярната информация, дезинформацията, малинформацията, слуховете, сатирата, кликбейтът, съмнителните новини и инфодемията не могат да бъдат обобщени без методологична предпазливост. Таблицата е запазена, защото тя определя речника, използван по-късно за хармонизиране на етикетите, интерпретация на наборите от данни и разграничението между откриване на ниво съдържание и моделиране на ниво дифузия [2], [3], [30], [41].

Таблица 1.1. Основни понятия, свързани с фалшивите новини, и тяхното значение за разкриването им.

Концепция	Значение	Значение за откриването на фалшиви новини
Невярна информация	Невярна или неточна информация, която може да бъде споделена без умишлено намерение за заблуда.	Важно, когато намерението на източника е неизвестно, недостъпно или не може да се наблюдава директно в набора от данни [41].
Дезинформация	Невярна или подвеждаща информация, създадена или разпространена умишлено с цел заблуда.	Най-близко до тясното определение за фалшиви новини и силно релевантно към умишлената манипулация [2], [3].
Малинформация	Вярна или частично вярна информация, използвана в подвеждащ контекст или с вредоносно намерение.	Изисква контекстуален анализ, тъй като проблемът може да се крие в рамкирането, времето или селективното разкриване [41].
Слухове	Непроверена информация, която циркулира, преди нейният истинен статут да бъде потвърден или отхвърлен.	От значение за времевата проверка, развиващите се етикети и разпространението преди потвърждението [111], [53].
Пропаганда	Стратегически структурирана комуникация, предназначена да повлияе на убеждения, нагласи или поведение.	Може да се припокрива с фалшиви новини, когато се използват селективни факти, емоционално рамкиране или изфабрикувани твърдения, за да се повлияе на аудиторията.
Сатира / пародия	Съдържание, което умишлено имитира новинарски стил с цел хумор, критика или преувеличение.	Може да бъде погрешно класифицирано като фалшива новина, освен ако не се вземат предвид жанрът, намерението и контекстът на публикуване [111].
Кликбейт	Сензационни или подвеждащи заглавия, предназначени да привлекат вниманието и да увеличат ангажираността.	Може да е невярно, частично вярно или просто манипулативно; полезно за откриване на дезинформация, породена от ангажираност.
Съмнителни новини	Съдържание, чиято истинност е несигурна, смесена, слабо подкрепена или недостатъчно проверена.	Важно за класификация, съобразена с несигурността и със забавено вземане на решения, а не за принудително бинарно етикетирание.
Инфодемия	Мащабна смесица от точна, неточна и подвеждаща информация, особено по време на кризи.	Силно релевантно за изследванията на дезинформацията в общественото здравеопазване и фалшивите новини за COVID-19 [30].

1.2.3 Предизвикателства при дефинирането на фалшиви новини

Основното предизвикателство пред дефинициите е, че истинността може да зависи от контекста, времето, източниците и по-късната проверка. Неяснотата в анотациите, сарказмът, сатирата, липсващият контекст и частичната истина могат да причинят разногласия между проверяващите факти или аотиращите набори от данни. Това оправдава избора на дисертацията да се запазят оригиналните етикети като метаданни, да се изключат двусмислени записи, когато е необходимо, и да се интерпретират резултатите от модела като оценки за достоверност, съобразени с неопределеността [3], [4], [30], [111].

1.3 Класически подходи за машинно обучение за откриване на фалшиви новини

1.3.1 Методи за представяне на текст

Класическото разпознаване на фалшиви новини представя текста чрез Bag-of-Words, TF-IDF, n-грами на думи или символи и стилметрични характеристики. Тези представяния са ефикасни и интерпретируеми, но са ограничени, когато дезинформацията зависи от контекст, перифраза, ирония или дългосрочни семантични отношения. Те остават важни в дисертацията, защото предоставят прозрачни базови линии за оценка на разпознаването, базирано на трансформатор (Transformer) [3], [45], [46].

1.3.2 Класически модели за класификация

Таблица 1.2 обобщава основните класически класификатори, използвани като базова литература: Логистична регресия, Наивен Байесов метод, SVM, Случайна гора и методи за бустирание. Тяхната ценност е в интерпретируемостта и възпроизводимостта, докато ограничението им е слабото контекстуално разбиране и чувствителността към промяна на домейна [42]-[46].

Таблица 1.2. Класически подходи за машинно обучение за откриване на фалшиви новини.

Подход	Типични характеристики	Представителна литература	Силни страни	Ограничения
Логистична регресия	TF-IDF, n-грами, метаданни	Шу и др. [3]; Агарвал и др. [45]; Джанг и др. [46]	Просто, интерпретируемо и силно като линейна базова линия.	Ограничена способност за моделиране на нелинейни семантични отношения.

Наивният Байес	Брой думи, двоични термини, TF-IDF	Шу и др. [3]; Агарвал и др. [45]; Джанг и др. [46]	Бърз, мащабируем и ефективен за малки текстови набори от данни.	Разчита на силно предположение за условна независимост.
Машина за опорни вектори	TF-IDF, n-грами от думи и символи	Йоаким [42]; Agarwal и др. [45]; Джанг и др. [46]	Ефективен във високоразмерни разреждени пространства.	Изисква внимателен избор на ядро и параметри за големи набори от данни.
Случайна гора	Стилометрични, лексикални, четимостни и метаданни характеристики	Брейман [43]; Агарвал и др. [45]; Джанг и др. [46]	Устойчиви на шум инженерни характеристики и полезни за важни функции.	По-малко подходящо за много оскъден текст без намаляване на функционалността.
Градиентно усилване / XGBoost	Лексикални, стилистични, метаданни и структурни характеристики	Чен и Гестрин [44]; Джанг и др. [46]	Силна предсказваща производителност с проектирани функции.	Изисква настройване и може да пренапасне с шумни или предубедени етикети.

1.3.3 Предимства и ограничения

Класическите модели остават полезни, защото разкриват дали даден набор от данни е разделен чрез повърхностни лексикални или стилистични сигнали за източника. Текстовете с фалшиви новини обаче често включват имплицитно противоречие, внимателно рамкиране и прехвърляне на тема, които изискват контекстуални представяния. Това ограничение мотивира трансформаторния слой на предложената рамка, като същевременно запазва класическите методи като точки за сравнение.

1.4 Трансформаторни модели за откриване на фалшиви новини

1.4.1 Архитектура на трансформатора и механизъм за внимание

Трансформаторните модели използват самовнимание, за да представят връзките между токени, обекти, числа и твърдения в рамките на една последователност. Този механизъм е ключов за откриването на фалшиви новини, тъй като подвеждащото съдържание често зависи от контекстуални връзки, а не от изолирани думи.

Прегледаният принцип на внимание обяснява защо моделите в стил BERT могат да се подобрят спрямо оскъдните лексикални базови линии [6]-[8].

1.4.2 BERT и RoBERTa-базирано откриване

BERT и RoBERTa са разгледани като основни контекстуални архитектури за класификация на двоични фалшиви новини. BERT осигурява двупосочно контекстуално кодиране, докато RoBERTa подобрява избора на обучение и често увеличава устойчивостта. В рамките на дисертацията, тяхната роля е да оценят вероятността за фалшиви новини, която по-късно може да бъде свързана с доказателства за времева каскада [6], [8].

1.4.3 Многоезични модели на трансформатори

Многоезичните BERT и XLM-RoBERTa са от значение, защото изследванията на дезинформацията не трябва да се ограничават само до английски набори от данни. Техните споделени представяния на поддуми позволяват да се оцени откриването на фалшиви новини на албански с ниски ресурси, като същевременно се запазва същата двоична логика за откриване, използвана за английски и здравни набори от данни [9], [11].

1.4.4 Модели с дълги последователности

Откриването на фалшиви новини в дълги документи изисква модели, които намаляват загубата на информация, причинена от съкращаване. BigBird е разгледан, защото оскъдното внимание позволява по-ефективна обработка на доказателства от по-дълги статии, което го прави подходящ, когато подвеждащи сигнали се появяват след първата част на статията [12].

1.4.5 На ниво изречения, LLM и нововъзникващи насоки

SBERT и моделите на големи езици се третират като релевантни, но спомагателни насоки. Вграждането на изречения поддържа сравнението на твърдения, докато LLM могат да помогнат за обяснение или контролирано допълване. Генерираните изходи обаче не се третират като етикети или основна истина, тъй като халюцинациите, пристрастията и свръхувереността остават методологични рискове [10], [47]-[49], [181]-[183].

Таблица 1.3. Основни модели, базирани на трансформатори, свързани с откриването на фалшиви новини.

Модел	Основна идея	Значение за откриването на фалшиви новини
BERT	Двупосочен трансформаторен енкодер, предварително обучен с моделиране на маскиран език.	Силна основа за класификация на контекстуални фалшиви новини [6].
RoBERTa	Оптимизирано BERT обучение с повече данни, динамично маскиране и без предвиждане на следващото изречение.	Често подобрява устойчивостта и класификационните характеристики [8].
mBERT	Многоезичен BERT, обучен на 104 езика.	Полезно за откриване на фалшиви новини на различни езици или на много език [9].
BigBird	Трансформатор с разрежено внимание, който комбинира локално/прозоречно, случайно и глобално внимание, за да моделира дълги последователности с линейна сложност на вниманието.	Полезно за откриване на фалшиви новини на пълни статии, защото намалява загубата на информация, причинена от отрязването на дълги статии до първите 512 токена [12].
SBERT	Вграждания на ниво изречения, базирани на сиамски/триплетни BERT мрежи.	Полезно за семантично сходство, сравнение на твърдения и класификация на ниво изречения [10].
XLM-RoBERTa	Междоезичната RoBERTa е обучена върху големи многоезични данни от CommonCrawl.	Полезно за многоезично откриване и трансфер между езици [11].
Големи езикови модели	Моделите на генеративни трансформатори, използвани за допълване, обяснение и задачи с нулев резултат.	Поддържа допълване и откриване на нулев изстрел; изисква повишено внимание поради халюцинации и пристрастия.

1.4.6 Ограничения на трансформаторно-базираното откриване

Трансформаторните детектори могат да бъдат изчислително скъпи, по-малко прозрачни от класическите базови линии и уязвими към състезателен парафраз или изместване на домейна. Следователно техните вероятности трябва да бъдат валидирани и интерпретирани като оценки за риск от съдържание, особено когато по-късно се използват като входни данни за каскадно и епидемично моделиране [50], [51].

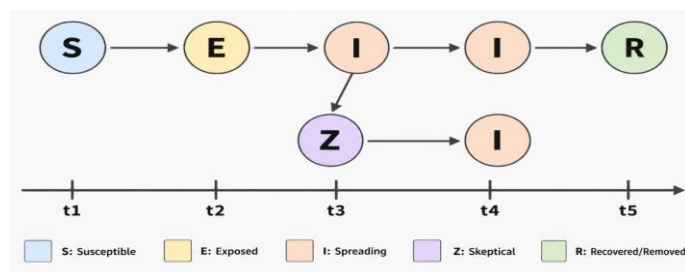
1.5 Ансамблови и хибридни подходи за откриване на фалшиви новини

Разглеждат се ансамбъл и хибридни детектори, защото те могат да комбинират лексикални, невронни, метаданни или графово-базирани сигнали. Дисертацията не имплементира ансамбъл за гласуване или подреждане като основен детектор, тъй като низходящият дифузионен слой изисква ясна и одитируема трансформаторна

вероятност. Рамката е хибридна на аналитично ниво: откриването на съдържание, темпоралните мрежи, динамиката на SEIZ и ABC неопределеността са свързани като отделни слоеве с доказателства.

1.6 Временно-мрежови подходи към разпространението на фалшиви новини

Временните мрежи осигуряват мост между откриването на фалшиви новини и моделирането на социалния риск. За разлика от статичните графики, те запазват хронологичния ред на ретуитовете, отговорите, повторните публикации или споделянията, позволявайки само пътища, съобразяващи се с времето. Запазените каскадни величини, включително нови активни разпространители, кумулативни разпространители, време до пик и междусъбитийно забавяне, се превръщат в емпирични обобщения, използвани по-късно за SIR/SEIZ фитинг и ABC извод [52]-[55].



Фигура 1.1. Временна каскада от фалшиви новини с подредени във времето потребителски състояния.

1.7 Модели на разпространение на епидемии и фалшиви новини

Моделите на епидемии и слухове предлагат интерпретируем език за моделиране на разпространението на фалшиви новини чрез състояния като податлив, изложен, активно разпространяващ се, възстановен и скептичен. Таблица 1.4 е запазена, защото съпоставя биологичната терминология с наблюдаеми показатели в социалните медии, като същевременно подчертава, че онлайн споделянето е само приближение на заразата [31]-[36].

Таблица 1.4. Съпоставяне между епидемичната терминология и разпространението на фалшиви новини.

Концепция за епидемия	Тълкуване на болестта	Тълкуване на фалшиви новини	Наблюдаем заместител в социалните данни
Чувствителен индивид	Лице, което може да бъде заразено	Потребител, който все още не е ангажирал с твърдението, но може да е бил разкрит	Връзка с последователите, членство в общността, видимост в хронологията

Изложен индивид	Лицето е заразено, но все още не е заразно	Потребител, който е видял твърдение, но все още не го е споделил	Преглед/импресия, когато е налична; забавен отговор; забавяне преди повторно публикуване
Заразен индивид	Лице, което в момента е предавател на заболяването	Потребителят активно разпространява твърдението	Ретуит, публикувай отново, цитирай, коментирай, споделяй
Възстановен/Премахнат индивид	Лицето вече не е заразно	Потребителят спира да разпространява, губи интерес или бива коригиран	Няма по-нататъшно споделяне след корекция или след каскаден разпад
Скептик	Не е стандартно в SIR; използва се в модели на слухове	Потребител, запознат с твърдението, но не го разпространява или го отхвърля/опровергава	Опровергаващ отговор, корекция, неангажиране след разкриване

1.7.1 Класически модели на епидемия и слухове

Класическите модели SI, SIS, SIR и SEIR предоставят речника за преход на състоянията, докато моделите на слухове на Дейли-Кендъл и Маки-Томпсън въвеждат поведение на насищане, отхвърляне и спиране. Тази литература оправдава използването на SIR като прозрачна базова линия и SEIZ като по-богат модел за каскади от фалшиви новини [31], [32], [67], [71]-[75].

Таблица 1.5. Класически епидемични модели и тяхната интерпретация за разпространение на фалшиви новини.

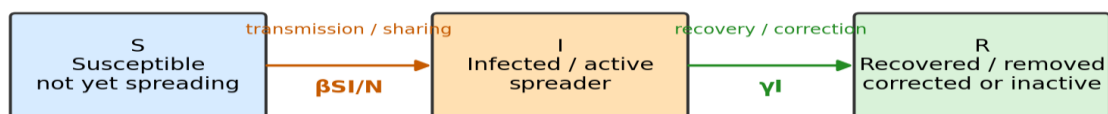
Модел	Основни състояния	Представителна литература	Тълкуване на фалшиви новини	Основна сила и ограничение
SI	$S \rightarrow I$	Пастор-Саторрас и Веспиняни [71]; Пастор-Саторрас и др. [72]	Потребителите преминават от потенциално излагане към активно разпространение.	Полезно за ранна неконтролирана дифузия, но няма състояние на корекция, забравяне или възстановяване.
SIS	$S \rightarrow I \rightarrow S$	Пастор-Саторрас и др. [72]	Потребителите могат да спрат да разпространяват вируса и по-късно да станат отново податливи след	Полезно за повтарящо се споделяне, но не представлява трайна корекция.

			многократно излагане.	
SIR	$S \rightarrow I \rightarrow R$	Кермак и Маккендрик [31]; Джин и др. [33]; Мураяма и др. [39]	Потребителите споделят фалшиви новини и по-късно спират, забравят, биват коригирани или премахвани от активната каскада.	Интерпретируема базова линия с параметри на предаване и възстановяване, но без изрично забавяне на експозицията или скептицизъм.
SEIR	$S \rightarrow E \rightarrow I \rightarrow R$	Пастор-Саторрас и др. [72]; Чентола [74]; Уанг (Wang) и др. [75]	Потребителите могат да разгледат съдържанието, преди да решат дали да го споделят.	Улавя забавянето на експозицията, но не представя изрично скептично или опровергаващо поведение.

1.7.2 SIR и SEIZ за разпространение на фалшиви новини

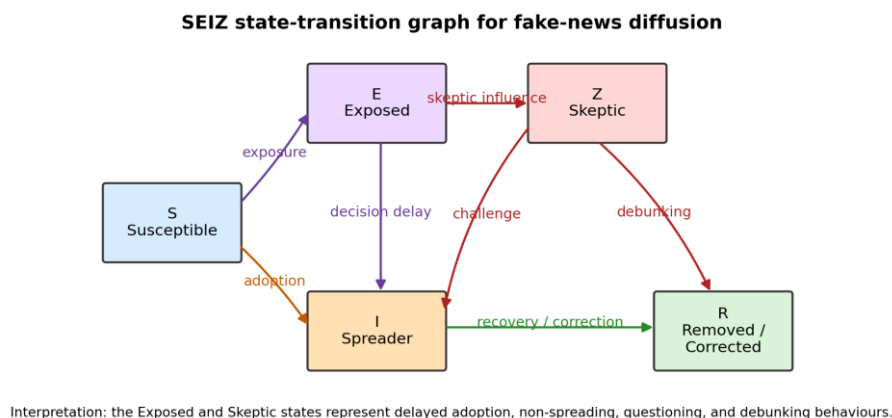
SIR представя потребителите като податливи, активни разпространители и възстановени или неактивни участници. SEIZ разширява тази структура с открити и скептични или неразпространяващи се състояния, което я прави по-подходяща за процеси на дезинформация, включващи забавяне, колебание и корекция. Фигурите и таблицата с коефициентите се запазват, защото те определят интерпретацията на моделирането, използвана в Глава 3 [33]-[36], [73].

SIR state-transition graph for fake-news diffusion



Interpretation: users move from potential exposure, to active spreading, to correction, inactivity, or removal.

Фигура 1.2. Графика на прехода на състоянието на SIR за разпространение на фалшиви новини.



Фигура 1.3. Графика на прехода на състоянието на SEIZ за разпространение на фалшиви новини.

Таблица 1.6. Основни коефициенти в моделите за разпространение на фалшиви новини SIR и SEIZ.

Коефициент	Модел	Значение на разпространението на фалшиви новини	Възможен емпиричен заместител
β	SIR / SEIZ	Скорост на предаване от активни разпространители към уязвими потребители	Процент на повторни публикации, процент на експозиция, тежест на ребрата, централност на инфлуенсъра
γ	SIR / SEIZ	Процент на възстановяване, забравяне или спиране	Време до края на споделянето; каскаден разпад
b	SEIZ	Процент на контакт между уязвими потребители и скептични потребители	Отговори от развенчаващи; коригиращи взаимодействия
p	SEIZ	Вероятност за директно приемане след контакт с разпространител	Вероятност за незабавно повторно публикуване след първото излагане
l	SEIZ	Вероятност да станете скептични след контакт със скептици	Корекционни отговори, проверка на фактите, липса на ангажираност след опровергаване
ε	SEIZ	Скорост на преход от изложено към активно разпространение	Забавяне от публикуването или отговор до повторно публикуване
ρ	SEIZ	Влияние на активните разпространители върху изложените на риск потребители	Многократно излагане; дискусийни теми, които водят до приемане
δ	Разширен SEIZ	Процент на корекция или опровержение	Предупредителни етикети, проверки на фактите, официални корекции

1.7.3 Граници на епидемичната аналогия

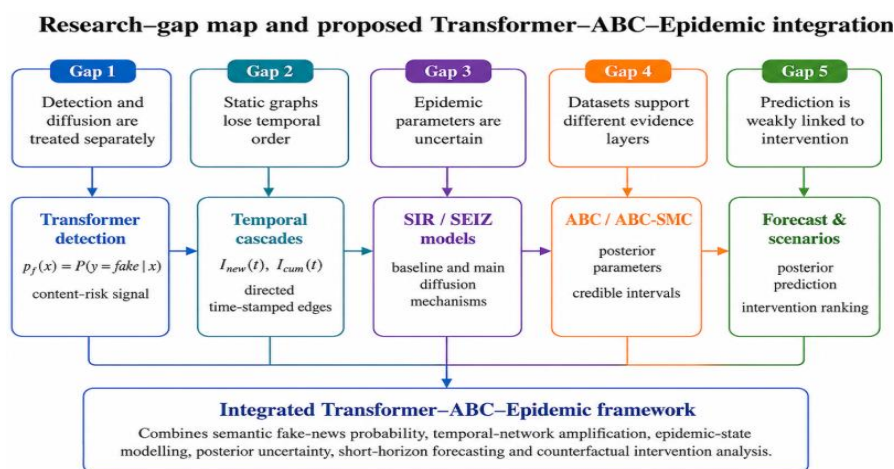
Аналогията с епидемията е полезна, но непълна. Данните от платформата не разкриват всяко излагане, лично споделяне, изтрита публикация или мълчаливо отхвърляне, а повторното публикуване не доказва достоверност. Това ограничение мотивира по-скоро оценката на неопределеността от последен път, отколкото детерминистичното точково напасване.

1.8 Приблизително байесово изчисление за дифузионен извод

ABC е разгледан, защото разпространението на фалшиви новини може да бъде симулирано, но точната му вероятност е трудна за определяне. Кандидатите за параметри SIR или SEIZ генерират синтетични каскади; симулираните обобщения се сравняват с наблюдаваните обобщения; а приетите частици апроксимират апостериорната неопределеност. ABC-SMC усъвършенства този процес чрез последователно намаляване на толеранса [19]-[25], [40], [77]-[80].

1.9 Пропуски в изследванията и открити проблеми

Липсите в литературата са свързани: откриването често е отделено от разпространението, статичните графики скриват времевия ред, параметрите на епидемията са несигурни, наборите от данни поддържат различни слоеве доказателства, а моделирането на интервенцията обикновено е отделено от апостериорната несигурност. Тези празнини мотивират рамката Transformer-ABC-Epidemic, където вероятността за съдържание, доказателствата за темпорална каскада, динамиката на SEIZ и апостериорната прогноза ABC/ABC-SMC се анализират заедно [86]-[89].



Фигура 1.4. Карта на пропуските в научните изследвания и предложена интеграция Transformer-ABC-Epidemic.

Таблица 1.7. Пропуски в изследванията, методологични отговори и очаквани доказателства.

Пропуски в изследванията	Методологичен отговор в Глава 2	Очаквани доказателства в Глава 3	Принос към дисертацията
Детекцията и дифузията често се третираят отделно.	Използвайте вероятността за фалшиви новини от Transformer като сигнал за риск от съдържание и я свържете с конструкцията на темпоралната каскада.	Метрики за откриване, вероятности на ниво претенция и каскадни изходи за допустимост.	Свързва достоверността на съдържанието с анализа на дифузионния риск.
Статичните графики могат да скрият пътища на дифузия, съобразяващи се с времето.	Представете твърденията като насочени темпорални каскади с подредени взаимодействия и криви на разпространение, индексирани във времето.	$I_{new}(t)$, $I_{cum}(t)$, размер на каскадата, време до достигане на пик и показатели за времевата мрежа.	Запазва хронологията, необходима за моделиране на разпространението на фалшиви новини.
Параметрите SIR/SEIZ са несигурни и само частично наблюдаеми.	Оценете неопределеността на параметрите чрез ABC и ABC-SMC, използвайки наблюдавани каскадни обобщения.	Постериорни оценки, достоверни интервали и постериорна предсказваща калибрация.	Предоставя извод за епидемичен модел, съобразен с несигурността, за каскади от фалшиви новини.
Наборите от данни за съдържание и наборите от данни за разпространение поддържат различни форми на доказателства.	Разделете наборите от данни, ориентирани към съдържание, от наборите от данни, ориентирани към дифузия, и дефинирайте ясни роли в наборите от данни.	Различни експерименти за откриване и експерименти за дифузия с използване на незаменяеми набори от данни.	Подобрява методологичната валидност и предотвратява твърдения за неподдържани набори от данни.
Анализът на смекчаването изисква повече от класификационни етикети.	Използвайте сценарии за последващо прогнозиране и контрафактуална интервенция при ABC-калибрирани дифузионни модели.	Прогнози, сценарии за интервенция и базови функционални резултати, свързани с риска от дифузия.	Подкрепя интервенционно-ориентирано разсъждение от интегрираната рамка.

1.9.1 Изследователски въпроси, произтичащи от липсата на литература

- 1) B1. Колко ефективно трансформаторните модели могат да класифицират фалшивите и истинските новини в кратки текстове, дълги текстове, многоезични и нискоресурсни набори от данни?
- 2) B2. Как твърденията за фалшиви новини могат да бъдат представени като времеви каскади, подходящи за моделиране в стил епидемия?
- 3) RQ3. SEIZ предоставя ли по-подходящо представяне от SIR за разпространение на фалшиви новини, когато забавянето на експозицията и скептицизмът са от значение?
- 4) RQ4. Могат ли ABC и ABC-SMC да оценят неопределеността в параметрите на разпространение на фалшиви новини от частично наблюдавани каскади?
- 5) RQ5. Може ли комбинирана рамка „Трансформатор-Темпорална-Мрежа-ABC“ да подпомогне последващо прогнозиране и интервенционно-ориентиран анализ на риска от фалшиви новини?

1.10 Заключение на главата

Глава 1 заключава, че анализът на фалшивите новини изисква интегрирани доказателства, а не изолирана точност на класификацията. Следователно целта на дисертацията е да се разработи и оцени рамка, която свързва езикови модели, базирани на Transformer, представяне във времева мрежа, SIR/SEIZ епидемично моделиране и ABC/ABC-SMC апостериорна inferенция. Изследователските задачи, произтичащи от тази цел, обхващат преглед на литературата, разделяне на набори от данни, бинарно откриване, каскадно конструиране, SIR-SEIZ сравнение, извод, съобразен с неопределеността, и интервенционно-ориентиран синтез на риска.

Таблица 1.8. Компактен синтез на прегледаната литература и нейното значение за предложената дисертация.

Литературно направление	Основна сила	Оставащо ограничение	Релевантност към тази дисертация
Дефиниции и анотации на фалшиви новини	Изяснява фалшиви, съмнителни, подвеждащи и свързани с тях понятия.	Етикетите могат да бъдат субективни, непълни или нестабилни във времето.	Мотивира внимателно двоично картографиране и обсъждане на неяснотите в анотациите.
Класическо машинно обучение	Интерпретируема, ефикасна и полезна като базова	Ограничено контекстуално разбиране и слаб	Предоставя исторически контекст и обясними точки за

	методология.	трансфер при промяна на домейна.	сравнение.
Детекция на базата на трансформатор	Улавя контекстуалната семантика и поддържа вероятно оценяване.	Изисква изчисления, качество на данните и адаптация към домейна.	Формира основния слой за откриване на съдържание на предложената рамка.
Калибриране на вероятностите и устойчивост	Поддържа надеждна интерпретация на резултатите от модела и стабилност между различните набори от данни.	Не замества необходимостта от моделиране на времева дифузия.	Мотивира използването на калибрирана вероятност на трансформатора като вход към слоя на дифузионния риск.
Анализ на темпоралната мрежа	Запазва подредбата, закъсненията и маршрутите, спазващи времето.	Често откъснати от етикетите за достоверност на съдържанието.	Поддържа каскадно изграждане и функции за темпорална дифузия и памет по пътищата.
Модели на епидемия и слухове	Представете състояния на разпространение, прагове и каскадна динамика.	Човешкото поведение и скритата експозиция са само приблизителни.	Осигурява SIR и SEIZ основи за анализ на разпространението на фалшиви новини.
ABC и ABC-SMC извод	Оценява несигурни параметри, когато вероятностите са неразрешими.	Чувствителен към обобщенията, идентифицируемостта и изчислителните разходи.	Позволява калибриране с отчитане на неопределеността на дифузионните модели SIR/SEIZ.
Интервенционно-ориентирано моделиране	Свързва прогнозирането със стратегии за смекчаване.	Често отсъства от стандартните подходи само за откриване.	Мотивира последващо прогнозиране и интервенционен анализ в Глава 2.

Глава 2. Предложена рамка и методология за трансформатор-ABC-епидемия.

2.1 Обща предложена рамка

Глава 2 операционализира теоретичната рамка. Предложената методология има четири слоя: слой данни за набори от данни за съдържание и дифузия, слой за двоично откриване, базиран на Transformer, слой за моделиране на темпорални мрежи и епидемии и слой за извод ABC/ABC-SMC. Таблица 2.1 и алгоритъм 2.1 са запазени, защото осигуряват проследимост между изследователските въпроси, методологичните компоненти и резултатите от Глава 3.

Таблица 2.1. Методологична матрица за проследимост на внедрената рамка на дисертацията

Изследователски въпрос / цел	Методологичен компонент	Тип набор от данни	Глава 3 доказателства или резултати
RQ1: откриване на двоични фалшиви новини в различни текстови типове и езици	BERT, RoBERTa, BigBird, mBERT, XLM-RoBERTa и класически базови линии	Ограничение@AAAI2021, FRN, ISOT, CoAID и Албански корпус на фалшивите новини	Точност, прецизност, пълнота на отчитане, F1-оценка, ROC-AUC и анализ на грешките
RQ2: времево каскадно представяне на твърдения за фалшиви новини	Нормализиране на времеви отпечатъци, насочени ребра $I_{\text{new}}(t)$ и $I_{\text{cum}}(t)$ каскадни характеристики	Каскада от Twitter/X на COVID-19, измами / OSoMeNet, каскада, внесена от измами, и синтетична албанска каскада	Графики на размера, дълбочината, широчината, времето до достигане на пик и времево-мрежовата каскада
RQ3: Моделиране на дифузия на SIR спрямо SEIZ	Картографиране на базовата линия на SIR и състоянието на SEIZ за активно разпространение, експозиция и скептицизъм	Основни емпирични каскади от фалшиви новини	Напаснати криви на разпространение, траектории на отделенията и таблици за сравнение на модели
RQ4: извод за параметри на епидемията, съобразен с неопределеността	ABC и ABC-SMC апостериорна инференция, използваща наблюдавани каскадни обобщения	Емпирични дифузионно-ориентирани каскади	Постериорни оценки, достоверни интервали, обобщения на конвергенцията и калибрирането
RQ5: последващо прогнозиране и интервенционен анализ	Постериорна прогнозна симулация, краткосрочно прогнозиране и смущение на параметри от контрафактуален подход	Емпирични каскади с извлечени времеви редове	Прогнозни резултати, сценарии за интервенция и функционални доказателства за риска от дифузия на базовата линия

Алгоритъм 2.1. Трансформатор-ABC-Епидемичен тръбопровод.

Входни данни: базирани на съдържание набори от данни за фалшиви новини, дифузионно-ориентирани каскадни записи и документираните правила за предварителна обработка.

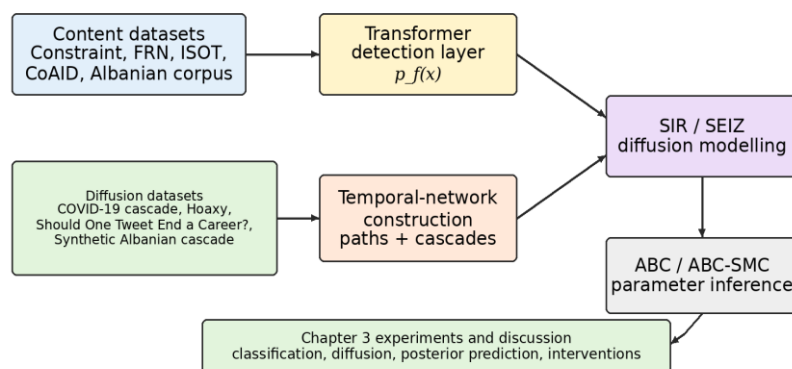
1. Почистете текстовите полета и хармонизирайте етикетите в проверяемо двоично представяне фалшив/реален код.
2. Токенизирайте съдържанието според избраното семейство Transformer и оценете слоя за двоично откриване.

3. Извлекете или запишете вероятността за фалшиви новини, оценена от модела, $p_f(x)$ за допустими твърдения.
 4. Запазете времеви отпечатъци, потребителски или идентификатори на източник-цел, каскадни идентификатори и типове действия за дифузионно-ориентирани записи.
 5. Конструирайте насочени темпорални каскади и изчислете $I_{\text{new}}(t)$, $I_{\text{cum}}(t)$ както и обобщена статистика на мрежата.
 6. Напаснете дифузионните модели SIR и SEIZ към емпиричните криви на разпространение.
 7. Изпълнете ABC и ABC-SMC изводи, използвайки наблюдавани каскадни обобщения и генерирани от симулатор траектории.
 8. Генерирайте последващи прогностични криви, краткосрочни прогнози и хипотетични сценарии за интервенция.
- Резултат: класификационни доказателства, каскадно представяне, оценки на епидемичните параметри, апостериорна неопределеност и интервенционно-ориентиран дифузионен анализ на риска.

2.2 Критерии за избор на набор от данни

Изборът на набор от данни разделя доказателствата, базирани на съдържание, от доказателствата, ориентирани към дифузия. Наборите от данни за съдържание трябва да предоставят текст и фалшиви/реални етикети за контролирано откриване, докато наборите от данни за дифузия трябва да предоставят времеви марки, взаимодействия и каскадна структура за SIR/SEIZ и ABC моделиране. Това разграничение предотвратява интерпретирането на наборите от данни за класификация като доказателства, базирани на времева мрежа, и поддържа възпроизводимостта чрез документация, етика и ограничения за достъп [30], [81]-[85], [90]-[94].

За дифузионно-ориентиран анализ изискванията за наборите от данни са по-строги. Моделирането на SIR, SEIZ, ABC и ABC-SMC изисква наблюдаеми времеви сигнали, взаимодействия с потребителите, каскадно членство и мрежова структура. Запазеният дифузионен слой от Глава 3 следователно използва каскадата от твърдения за фалшиви новини от Twitter/X на реални данни за COVID-19 „Коронавирусът не само влияе на начина, по който живеем, но и драматично влияе на начина, по който умираме.“, каскадата от твърдения за политически фалшиви новини Ноаху / OSoMeNet, импортираното от Ноаху социално твърдение за фалшиви новини „Трябва ли един твит да сложи край на кариерата?“ и синтетичната албанска времева каскада. Тези набори от данни се запазват, защото поддържат криви на разпространение с времеви прозорци, насочени взаимоотношения между потребители и $I_{\text{new}}(t)$ и $I_{\text{cum}}(t)$ количествата, необходими за SEIZ и ABC/ABC-SMC моделиране [90], [86]-[89].



Фигура 2.1. Логика за избор на набор от данни и методология за рамката Transformer-ABC-Epidemic, състояща се от четири глави

Таблица 2.2. Критерии за избор на набор от данни и релевантност към предложената рамка за бинарно-фокусирана дисертация

Критерий	Изискване за избор	Роля в откриването на двоични елементи	Роля в дифузията / ABC
Качество на етикетите и текста	Надеждни фалшиви/истински етикети и достъпно съдържание на текст, заглавие, твърдение, публикация или статия.	Необходимо за контролирано обучение и сравнение на модели.	Определя кои каскади се третират като дезинформация или като надеждно съдържание.
Баланс на бинарния клас	Достатъчно представяне на фалшиви и реални класове или документирана стратегия за дисбаланс.	Подобрява F1-резултата, прецизността, извикването на информация и анализа на грешките.	Избягва моделирането само на най-големите или най-видимите каскади от фалшиви новини.
Домейн, език и дължина	Отразяване на здравеопазване, политика, общи новини, дълги статии, кратки публикации и езици с ограничени ресурси.	Поддържа междудомейна, дълготекстова и многоезична оценка на Transformer.	Позволява сравнение на поведението на разпространение в различните области на дезинформация.
Временна информация	Време на публикуване, време на твитване, забавяне на ретуитване, време за отговор или хронология на събитието.	Полезно за експерименти за ранно откриване.	От съществено значение за каскадния растеж, времето до достигане на пик и последващото предсказване.
Взаимодействие и каскадна структура	Ретуитове, отговори, повторни публикации, коментари, дървета на разпространение или записи за ангажираност на	Може да се използва като метаданни, когато са налични етикети и текст.	Необходимо за темпорални мрежи, преходи между състояния и калибриране на SIR/SEIZ.

	потребителите.		
Етика и възпроизводимост	Лицензиране, ограничения за поверителност, политика на платформата, анонимизация и документация за набори от данни.	Осигурява отговорно използване на етикети и текст.	Осигурява отговорно използване на данни на ниво времева мрежа и потребител.

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

В уравнение (2.1), $Q(D)$ представлява общия резултат за пригодност на набор от данни D . Терминът $C_{\text{text}}(D)$ измерва текстовата наличност и качество; $C_{\text{label}}(D)$ измерва надеждността на етикетите; $C_{\text{binary}}(D)$ измерва съвместимостта с двоичната задача фалшив/реален формат; $C_{\text{time}}(D)$ измерва времевата информация; $C_{\text{network}}(D)$ измерва дали потребителските взаимодействия или каскади могат да бъдат реконструирани; и $C_{\text{ethics}}(D)$ измерва ограниченията за документация, достъп, поверителност и отговорно използване. Теглата могат да се коригират в зависимост от това дали експериментът е предимно експеримент за откриване на трансформатор или експеримент за дифузионно-инференсен извод.

2.3 Основни набори от данни

2.3.1 Набори от данни, базирани на съдържание, за откриване на трансформатори

Наборите от данни за съдържание поддържат двоично откриване при различни условия: Constraint@AAAI2021 за кратки публикации за COVID-19, FRN и ISOT за дълги публикации или настройки в стил статия, CoAID за здравна дезинформация и албанският корпус за многоезична оценка с ниски ресурси. Таблица 2.3 е запазена като обобщение на ролята на компактния набор от данни [30], [81]-[85].

Таблица 2.3. Сравнителен преглед на основните двоични набори от данни, базирани на съдържание, използвани в дисертацията

Набор от данни	Ключови характеристики	Основна употреба в тази дисертация
Constraint@AAAI2021	Дезинформация за COVID, кратък текст	Откриване на бинарни фалшиви новини в условия на кризисна информация.
FRN	обща класификация на фалшиви/истински	Откриване на фалшиви новини с дълъг текст и оценка от BigBird.
ISOT	дълги новинарски статии	Масшабен двоичен бенчмаркинг и междудоменно тестване за устойчивост.
CoAID	Съдържание, свързано с COVID	Откриване на дезинформация в здравеопазването и частичен анализ на социалния контекст.
Albanian Corpus	език с ниски ресурси	Многоезична оценка на трансформатор в среда с ниски ресурси.

2.3.2 Дифузионно-ориентирани набори от данни за моделиране на темпорални каскади

Раздел 2.3.2 представя накратко дифузионно-ориентирани набори от данни, използвани в Глава 3. Запазените набори от данни са: реалната каскада от фалшиви новини за COVID-19 в Twitter/X „Коронавирусът не само влияе на начина, по който живеем, но и драматично влияе на начина, по който умираме.“; каскадата от фалшиви новини за Ноаху / OSoMeNet „Три милиона гласа на президентските избори бяха подадени от нелегални имигранти.“; внесена от Ноаху каскада от социални фалшиви новини „Трябва ли един туит да сложи край на кариерата си?“, и синтетичното албанско фалшиво новино „Shqipëria do të presë 100 mijë palestinezë nga Gaza.“ Тези набори от данни предоставят записи с времеви отпечатък от потребител до потребител или на ниво твърдение за насочени времеви каскади, $I_{\text{new}}(t)$, $I_{\text{cum}}(t)$, SIR/SEIZ входни данни и ABC/ABC-SMC резюмета. Подробни описания, правила за филтриране и разширени метаданни са предоставени в Приложение D.

2.3.3 Обобщение на ролите в наборите от данни

Стратегията за набор от данни умишлено приписва различни роли на доказателствата за съдържание и разпространение. Наборите от данни за съдържание обучават и оценяват вероятностния слой на фалшивите новини; наборите от данни за разпространение предоставят криви на разпространение, подредени във времето, и мрежова структура. Това разделяне е методологична защита срещу неподкрепени твърдения.

2.4 Предварителна обработка на данни и експериментална подготовка

Предварителната обработка произвежда два координирани изхода: готови за съдържание записи за класификация на Transformer и готови за дифузия записи за моделиране на темпорална каскада. Неоднозначните етикети се изключват от основните бинарни експерименти, когато е необходимо, докато разширените примери са ограничени до данни за обучение, за да се предотврати изтичане. Таблица 2.4 обобщава запазените цели на предварителната обработка.

Таблица 2.4. Цели на предварителната обработка за откриване на двоични фалшиви новини и анализ на темпоралната дифузия

Цел на предварителната обработка	Съответни набори от данни	Методологична цел
Надеждност на текста	Constraint@AAAI2021, FRN, ISOT, CoAID, Албански корпус на фалшиви новини	Премахнете текстовия шум, като същевременно запазете сигналите за достоверност за двоична класификация.
Последователност на етикетите	Всички набори от данни, базирани на	Съпоставете хетерогенните етикети в общо фалшиво/реално представяне за експерименти

	съдържание	от Глава 3.
Готовност за писане на дълги текстове	FRN и ISOT	Подгответе по-дълги статии за Transformer, включително BigBird или стратегии, базирани на прозорци.
Нормализиране на ниските ресурси	Albanian Corpus на фалшивите новини	Запазете специфичните за албански символи, именувани обекти и местен речник за многоезичните Трансформърс.
Временна цялост	Дифузионно-ориентирани набори от данни	Запазете времевите отметки, реда на събитията и структурата на взаимодействие с потребителя за каскадно изграждане.
Възпроизводимост и предотвратяване на течове	Всички набори от данни	Правилата за почистване, разделяне и допълване на документа се записват, така че експерименталните резултати да могат да бъдат възпроизведени.

2.4.1 Почистване и нормализиране на текст

Почистването на текста е консервативно: дубликатите и празните полета се контролират, докато URL адресите, хаштаговете, споменаванията и специфичните за езика подсказки се запазват, когато могат да носят достоверност или дифузионно значение. Това е особено важно за кратки текстове в социалните медии и албански записи, където прекомерната нормализация може да премахне полезна информация [86]-[89], [99], [112].

Таблица 2.5. Основни операции за почистване и нормализиране на текст, запазени в основната методология

Операция	Лечение	Причина
Контрол на дублирани и празни текстове	Премахнете точните дубликати и изпразнете важните текстови полета.	Предотвратява завишени резултати и невалидни тренировъчни записи.
Обработка на URL адреси, споменавания и хаштагове	Стандартизирайте, а не изтривайте автоматично, когато токенът може да носи достоверност или значение за разпространение.	Запазва сигналите за тема, източник и платформа.
Нормализиране на пунктуацията и главните букви	Нормализирайте прекомерната пунктуация и използването на главни и малки букви, като същевременно запазвате границите на изреченията.	Подобрява токенизацията без премахване на стилистични доказателства.
Запазване на специфичните за езика символи	Запазете албанските символи като ё и џ.	Поддържа информация за език с ограничени ресурси.

2.4.2 Хармонизация на етикетите

Хармонизирането на бинарни етикети картографира хетерогенни етикети на набори от данни в обща фалшива/реална цел, като същевременно запазва оригиналните етикети като метаданни. Уравненията и Таблица 2.6 документируют принципа на картографиране и поддържат прозрачна интерпретация на всички по-късни резултати от откриването [81], [83].

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

В уравнение (2.2) y_{bin} означава хармонизираната двоична цел, използвана в експериментите за откриване на трансформатор. Съпоставянето не изтрива оригиналния етикет на набора от данни; вместо това оригиналният етикет се запазва като метаданни за сурови етикети. Това предотвратява методологична грешка, при която несигурни, спомагателни или специфични за дифузията етикети се третират като идентични със стандартни фалшиви/реални класификационни етикети.

Таблица 2.6. Примери за компактно бинарно картографиране, запазени в основната методология

Набор от данни	Оригинални етикети	Двоично картографиране
Constraint@AAAI2021	fake / real	fake -> 1; real -> 0
ISOT	Fake.csv / True.csv	Fake.csv -> 1; True.csv -> 0
Hoaxy / OSoMeNet	твърдение (claim) и проверка на фактите (fact-checking) - записи за ретуитване - твърдението може да се третира като ориентирано към дезинформация; проверката на фактите като коригиращо/надеждно в дифузионния анализ	claim-> 1 ; fact_checking->0

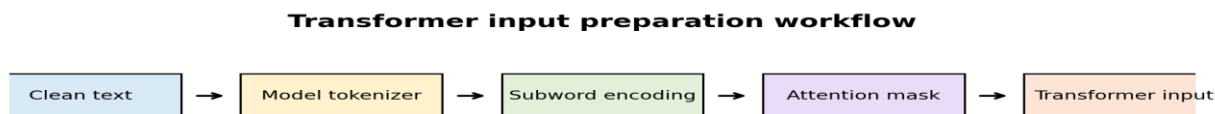
2.4.3 Токенизация и подготовка на входа на трансформатора

Токенизацията преобразува почистения текст в готов за модел вход, съгласно избраното семейство Transformer. BERT/mBERT, RoBERTa/XLM-R и BigBird използват различни политики за поддуми и последователности; следователно изборът на токенизатор се третира като част от експерименталния дизайн, а не като неутрален технически детайл.

$$\text{Tokenizer}_m(x_i) \rightarrow (\text{input_ids}_i, \text{attention_mask}_i, \dots) \quad (2.3)$$

В уравнение (2.3), Tokenizer_m означава токенизатора, свързан със семейството модели m , x_i е почистеният текстов запис, input_ids_i е последователността от идентификатори на поддуми и attention_mask_i маркира валидни входни позиции. Специфични за модела технически подробности, като token_type_ids , динамично допълване, специална обработка на токени, [CLS], [SEP] и варианти на отрязване, са преместени в

Приложение F, защото те са спецификации за имплементация, а не основни методологични аргументи.



Фигура 2.2. Работен процес за токенизация на трансформатора и подготовка на входните данни за модела.

Таблица 2.7. Стратегия за токенизация по семейство модели

Моделно семейство	Основна логика на токенизатора	Използване в тази дисертация
BERT / mBERT	Токенизация на WordPiece с фиксирана максимална дължина на последователността.	Откриване на фалшиви новини с кратка и средна дължина, включително многоезична поддръжка.
RoBERTa / XLM-R	Моделиране на поддуми на ниво байт BPE или SentencePiece в стил SentencePiece.	Надеждна обработка на поддуми за шумен или многоезичен текст.
BigBird	Вход с дълга последователност с разредено внимание.	Дълги набори от данни в стил статии, като например FRN.

2.4.4 Обработка на шумни и небалансирани данни

Шумът и дисбалансът се обработват чрез документирано филтриране, оценка, съобразена с класа, опционална загуба с претегленост на класа и контролирано допълване само с обучение. Запазените уравнения и таблицата определят коефициента на дисбаланс и претеглената загуба като методологични опции, докато основните резултати наблягат на извикването на фалшив клас, макро-F1 и ROC-AUC [100], [101].

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

Уравнение (2.4) определя съотношението на класов дисбаланс за двоичен набор от данни D , където n_{fake} и n_{real} означават броя на фалшивите и реалните примери. Стойност, близка до 1, показва приблизителен баланс, докато по-високите стойности показват по-силен дисбаланс. Това съотношение определя дали трябва да се използва класово претегляне, контролирано увеличение или оценка, съобразена с дисбаланса [101].

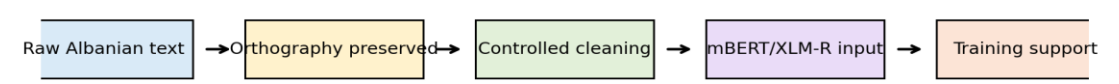
$$\mathcal{L}_{\text{WBCE}} = -[w_1 y \log(p) + w_0 (1 - y) \log(1 - p)] \quad (2.5)$$

Уравнение (2.5) дава претеглена бинарна цел за кръстосана ентропия, която може да се използва, когато дисбалансът е значителен. Теглата w_1 и w_0 компенсират разликите в честотата на класовете, докато p означава прогнозираната вероятност за фалшиви новини. Тази цел се третира като опция за обучение, а не като задължителен избор за всеки набор от данни.

Таблица 2.8. Компактен набор от стратегии за шумни и небалансирани двоични данни за фалшиви новини.

Стратегия	Използвайте
Загуба, претеглена по клас	Опция по подразбиране, когато е видим дисбаланс в класовете.
Контролирано увеличаване	Опция за подкрепа на обучението само за условия с ниски ресурси.
Оценка на макро-F1 и PR-AUC	Опции за оценка, които намаляват прекомерната зависимост от точността.

Low-resource Albanian preprocessing workflow



Фигура 2.3. Нискоресурсен работен процес за предварителна обработка на албански фалшиви новини за многоезичен Transformer вход

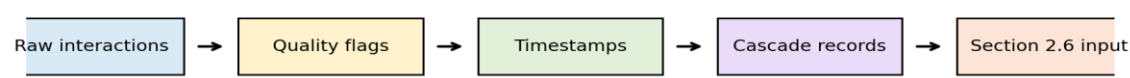
Таблица 2.9. Предиизвикателства и отговори на предварителната обработка на Compact Albanian

Предиизвикателство	Препоръчителен отговор	Причина
Специфични за езика знаци	Запазете ё и ç; избягвайте принудително ASCII преобразуване.	Предотвратява загубата на албанска правописна информация.
Местни имена и политически речник	Запазете именувани обекти и локални термини.	Поддържа специфични за домейна сигнали за достоверност.
Малък размер на набора от данни	Използвайте контролирано допълване само в рамките на обучителните данни.	Поддържа обучението без да променя резултатите от тестовите.

2.4.5 Работен процес за предварителна обработка

За дифузионно-ориентирани записи, предварителната обработка е по-малко агресивна, отколкото за текстова класификация, тъй като времевите марки, потребителските идентификатори, полетата за взаимодействие и каскадната принадлежност трябва да бъдат запазени. Резултатът е слой от събития с контролирано качество, който се превръща в вход за изграждането на темпорална каскада, SIR/SEIZ фитинг и изчисляване на ABC разстояние.

Diffusion-oriented preprocessing bridge



Фигура 2.4. Работен процес за предварителна обработка на дифузионно-ориентирани данни преди изграждане на темпорална каскада

Таблица 2.10. Основни цели на предварителната обработка на дифузионно-ориентирани данни

Цел	Методологична роля
Запазване на каскадната идентичност	Гарантира, че всяко събитие остава свързано с твърдението или нишката за фалшиви новини.
Запазване на потребителите и взаимодействията	Позволява по-късно изграждане на насочени отношения източник-цел.
Запазване на времеви отпечатъци	Поддържа темпорално подреждане, $I_{\text{new}}(t)$, $I_{\text{cum}}(t)$, SIR/SEIZ фитинг и ABC изчисляване на разстояние.
Присвояване на флагове за качество	Липсващи документи, дублирани или несъответстващи записи, без ненатрапчива промяна на доказателствата.

Таблица 2.10 е запазена, защото определя минималния изход, готов за дифузия, изискван от по-късните етапи на моделиране. Специфичните за набора от данни схеми за дифузия, решенията за контрол на качеството и съпоставянията на състоянията са преместени в приложенията, за да се избегне повтарянето на една и съща логика на конвейера както в секциите за предварителна обработка, така и в секциите за изграждане на времева каскада.

2.5 Конструиране на темпорална каскада и мрежово представяне

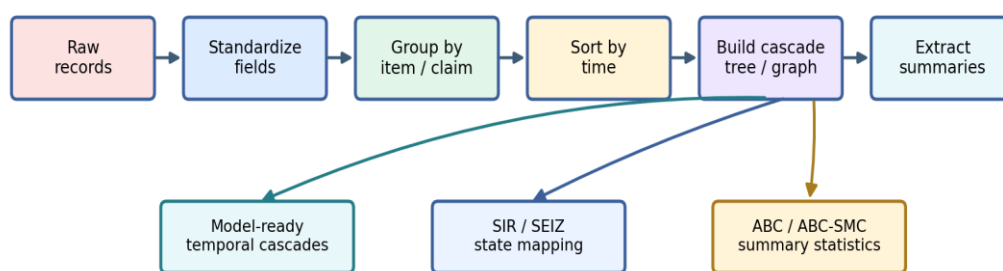
Раздел 2.5 определя преобразуването от дифузионни записи към насочени времеви каскади. Всяко събитие запазва източник, цел, времеви отпечатък, тип действие, идентификатор на каскада и етикет на твърдението. Това представяне запазва хронологична информация и позволява извличането на наблюдаваните криви на разпространение за по-късно епидемично и байесово моделиране [52]-[55].

Вместо невронни архитектури с темпорални графи е избрано представяне на темпорална мрежа, тъй като дисертацията изисква интерпретируема каскадна реконструкция и извод за епидемични параметри, а не научени динамични вграждания на възли или модели за предсказване на връзки. Наличните дифузионно-ориентирани набори от данни предоставят взаимодействия с времеви отпечатъци и каскадни следи, подходящи за пътища, спазващи времето, $I_{\text{new}}(t)$ картографиране $I_{\text{cum}}(t)$ на състоянията на SIR/SEIZ и извод за ABC/ABC-SMC. Невронните модели с темпорални графи, като TGN, TGAT, DyRep и EvolveGCN, следователно се третираат като бъдещи разширения, които биха изисквали по-богати данни за темпорални графи на потребителско ниво, стабилни атрибути на възли и ребра и допълнителни контролирани цели за обучение на графи [128]-[129], [153]-[155].

$$e_k = (u_s, u_t, t_k, a_k, c_k, y_k) \quad (2.6)$$

В уравнение (2.6), u_s означава източника или предишния възел, u_t означава целевия или отговарящия възел, t_k е времето на събитието, a_k е типът наблюдавано действие, c_k е идентификаторът на каскадното или фалшиво новинарско твърдение и y_k е наличният технически етикет, като например твърдение, covid_fake_claim или fact_checking. Това представяне поддържа етикетите на съдържанието и времевите доказателства свързани, без да третира всички записи като еквивалентни на епидемични състояния.

Temporal cascade construction from diffusion-oriented records



Фигура 2.5. Работен процес за изграждане на темпорална каскада от дифузионно-ориентирани записи до каскади, готови за моделиране

2.5.1 Каскадна предварителна обработка

Каскадната предварителна обработка групира събитията по твърдение или източник, филтрира запазените случаи на фалшиви новини и запазва полетата, необходими за времева реконструкция. Схемата, готова за моделиране, и каскадната дефиниция поддържат вероятността на съдържанието свързана с времевите доказателства, без да третират всеки запис като пряко психологическо състояние.

$$C_c = (V_c, E_c^T, x_c, y_c, p_f(x_c)) \quad (2.7)$$

В уравнение (2.7), V_c е множеството от потребители или изходни възли, участващи в каскадата, E_c^T е множеството от насочени взаимодействия с времеви отпечатък, x_c е свързаното текстово твърдение или статия за фалшиви новини, y_c е хармонизираният етикет на твърдението и $p_f(x_c)$ е вероятността за фалшиви новини, базирана на трансформатор, използвана като сигнал за риск от съдържание в по-широката рамка.

Таблица 2.11. Основни елементи на представянето на темпоралната мрежа

Елемент	Дефиниция в тази дисертация	Пример	Използване в по-късно моделиране
Възел	Потребител, публикация, твърдение за източник, статия или реакция в	Потребител на Twitter/X в Ноаху; туйт от източника в каскада от ретуитове.	Определя единиците, които могат да станат податливи, изложени

	зависимост от набора от данни.		на риск, разпространяващи се, възстановени или скептични при епидемична интерпретация.
Насочено ребро	Временно подредено взаимодействие от изходен възел към целеви възел.	from_user_id -> to_user_id в данните за ретуитване.	Представява възможни пътища за предаване на информация.
Времева марка	Времето, в което се случва взаимодействието, или забавянето от първоначалното събитие.	tweet_created_at в Twitter/X или записи в стил Noaxu.	Позволява времево подреждане, анализ от времето до пика и пътища, спазващи времето.
Тегло	Стойност, измерваща силата, честотата или релевантността на взаимодействието.	Брой ретуитове между двама потребители или повтарящи се взаимодействия.	Поддържа претеглено дифузионно моделиране и анализ на потребителското влияние.
Каскаден идентификатор	Етикет, свързващ всички събития, принадлежащи към едно и също твърдение за фалшиви новини или източник.	Идентификационен номер на иска, туйт на източника или етикет на външен елемент.	Групира събития за каскадна статистика и ABC обобщаващи функции.

Таблица 2.12. Унифицирана схема, готова за моделиране, за дифузионно-ориентирани записи

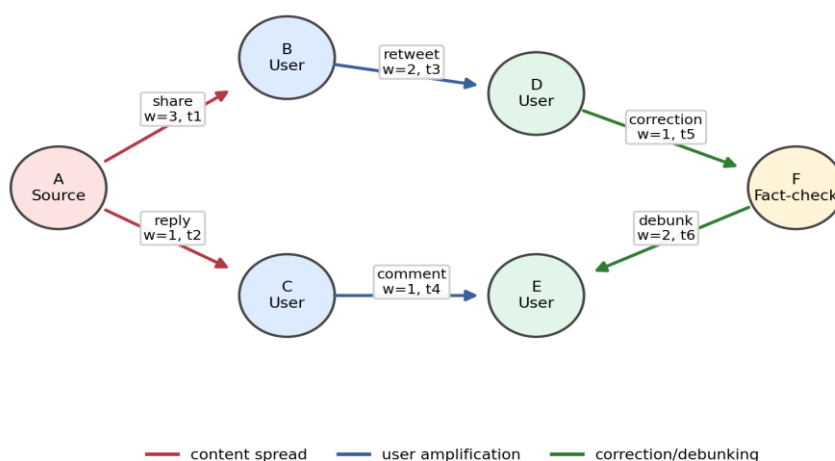
Поле	Определение	Задължителна роля в дисертацията
каскаден_идентификатор	Идентификатор, свързващ всички записи, принадлежащи към едно и също твърдение за фалшиви новини или източник.	Групира записите преди извличане на характеристики и подготовка за SIR/SEIZ.
източник	Първоначален потребител, източник на публикация, статия или предишен възел.	Определя началната страна на насочено темпорално ребро.
цел	Получаване, повторно публикуване, отговаряне или взаимодействие с потребител/възел.	Определя активираната или засегнатата страна на насочно темпорално ребро.
време	Времева маркировка или относително време на събитието.	Позволява хронологично подреждане и изграждане на каскади, спазващи времето.
тип_действие	Наблюдаван тип взаимодействие,	Поддържа интерпретация на показатели

	като например ретуит, отговор, повторна публикация, твърдение или запис за проверка на фактите.	за разпространение, експозиция или корекция.
етикет_на_претенцията	Технически данни, като например claim , covid_fake_claim или fact_checking.	Запазва разграничението между редове за твърдения за фалшиви новини и спомагателни/коригиращи редове.
текст_или_заглавие	Текстово твърдение, заглавие на статията или описание на външен елемент, когато е налично.	Свързва темпоралната каскада със слоя със съдържание, базиран на Transformer.

2.5.2 Изграждане на темпорална мрежа

Всяка последователност от събития на ниво твърдение е представена като насочена времева мрежа. Ребрата следват наблюдаема посока на взаимодействие и са валидни само когато времевите марки спазват хронологичния ред. Това предотвратява невъзможни пътища в статичен граф и поддържа анализ на времето до пик, достижимостта и каскадната дълбочина.

Weighted temporal-network representation of fake-news diffusion



Фигура 2.6. Илюстрация на възли, насочени ребра, тегла и времеви отпечатъци в мрежа за разпространение на фалшиви новини

$$P = [(u_0, u_1, t_1), \dots, (u_{m-1}, u_m, t_m)], \quad t_1 < t_2 < \dots < t_m \quad (2.8)$$

В уравнение (2.8), P означава път, съобразяващ се с времето; заявката може да премине от един потребител към следващия само когато времената на събитията се увеличат.

Таблица 2.13. Сравнение между статични и времеви мрежови представления.

Аспект	Статична мрежа	Временна мрежа	Извод за тази дисертация
Интерпретация на ребрата	Ребро съществува, ако взаимодействие се случи поне веднъж по време на прозореца за наблюдение.	Ребро съществува в определен момент или интервал от време.	Временните мрежи са подходящи за каскадна реконструкция и дифузионно време.

Валидност на пътя	Всички свързани пътища може да изглеждат валидни след агрегиране.	Само пътища, спазващи времето, са валидни.	Предотвратява невъзможни пътища за поток от информация.
Полезни функции	Степен, гъстота, клъстеризация и общности.	Време между събитията, време до пик и достъпност, съобразена с времето.	Временните характеристики поддържат SIR/SEIZ симулация и ABC обобщена статистика.
Ограничение	Губи реда на събитията.	Изисква точни времеви отметки и подреждане на събитията.	Временното моделиране е по-взискателно, но по-реалистично за разпространението на фалшиви новини.

$$s_{\text{out}}(i) = \sum_j \sum_t w_j(t) \quad (2.9)$$

където $s_{\text{out}}(i)$ представлява претеглената изходяща активност на потребител i , а $w_{ij}(t)$ е теглото на взаимодействието от потребител i към потребител j във време t . Тази мярка подпомага идентифицирането на потребители, които многократно разпространяват твърдение за фалшива новина чрез наблюдаеми взаимодействия.

$$R_{\text{cc}} = \frac{E_{\text{cross}}}{E_{\text{total}}} \quad (2.10)$$

където E_{cross} е броят на каскадните ребра, свързващи различни общности, а E_{total} е общият брой на каскадните ребра. Съотношението се използва като структурен индикатор за това дали твърдението за фалшива новина остава локализирано или се разпространява между общностите. Централността и откриването на общности се третират като установени инструменти за мрежов анализ, докато подробната им теоретична основа е запазена в прегледа на литературата и Приложение Е [53], [58]-[61].

2.5.3 Каскадни характеристики за SIR/SEIZ/ABC

Основните характеристики на каскадата са нови активни разпръсквачи, кумулативни активни разпръсквачи, краен размер на каскадата, време до достигане на пик, скорост на растеж и избрани мрежови дескриптори. Тези характеристики свързват описателния времеви анализ с обобщената статистика, използвана при калибриране на SIR/SEIZ и ABC/ABC-SMC извод.

Двете централни мерки за времева активация са $I_{\text{new}}(t)$ и $I_{\text{cum}}(t)$. $I_{\text{new}}(t)$ е броят на новите уникални активни разпространители, наблюдавани в даден времеви прозорец, докато $I_{\text{cum}}(t)$ е кумулативният брой уникални активни разпространители, наблюдавани до този момент. Тези две величини формират директния мост между изграждането на времева мрежа и напасването на епидемичния модел. Този мост е важен, защото ABC и ABC-SMC по-късно сравняват симулираните и наблюдаваните

криви на разпространение чрез обобщена статистика, а не чрез вероятност в затворена форма [76], [106].

Таблица 2.14. Каскадни характеристики за извличане на обобщение на SIR/SEIZ/ABC

Функция	Определение	Използвайте в SIR/SEIZ/ABC
Размер на каскадата	Общ брой участващи възли или уникални активни разпространители.	Използва се за сравняване на крайния размер на епидемията и последващата прогнозна сатурация.
$I_{\text{new}}(t)$	Брой новоактивни уникални разпространители във времеви прозорец t .	Използва се като наблюдавано увеличение на времето на заразяване за динамиката на дифузия.
$I_{\text{cum}}(t)$	Кумулативен брой уникални активни разпространители до момента t .	Използва се като основна кумулативна крива за моделиране и прогнозиране.
Време за достигане на пик	Време между старта на каскадата и максимума $I_{\text{new}}(t)$.	Поддържа обобщена статистика по ABC и сравнение на времето.
Време между събитията	Времеви пропуски между последователни събития в каскадата.	Описва спукване и забавено активиране.
Обхват и широта	Брой потребители или клонове, достигнати от каскадата.	Поддържа описателно сравнение и последваща прогнозна проверка.
Претеглена изходяща активност	Временна сума на теглата на изходящите ребра за всеки потребител.	Идентифицира влиятелни разпространители и възможни релейни хъбове.
Съотношение между общностите	Дял на ребрата, пресичащи границите на общността.	Обобщава структурното разпространение отвъд първоначалната общност.

$$G_c^T = (V_c, E_c^T) \quad (2.11)$$

В уравнение (2.11), G_c^T означава насочената времева каскадна мрежа, свързана с твърдение за фалшиви новини s . Множеството V_c съдържа потребителите, изходните акаунти или изходните обекти, които се появяват като възли в каскадата. Множеството E_c^T съдържа насочените взаимодействия с времеви отпечатък, наблюдавани за същото твърдение, където ребро може да бъде записано като $(u_i, u_j, t, w_{ij}(t))$, за да се посочи взаимодействие от възел u_i към възел u_j във време t с опционално тегло $w_{ij}(t)$. Горният индекс T показва, че наборът от ребра е индексирен по време и подреден по време; не се използва като транспонирана матрица. Тази формулировка предоставя формалния обект, от който се извличат $I_{\text{new}}(t)$, $I_{\text{cum}}(t)$, характеристики на каскадно ниво и SIR/SEIZ входове [52]–[55], [76].

Подредените активации на целевите потребители генерират $I_{\text{new}}(t)$ и $I_{\text{cum}}(t)$, докато мрежови свойства, като например претеглена изходяща активност и пресичане на общността, предоставят опционални структурни обобщения. На този етап ретуитовете,

отговорите, повторните публикации и записите за проверка на фактите не се интерпретират като директни наблюдения на латентни убеждения. Те са наблюдаеми заместители, които подкрепят моделирането на допускания и последващото сравнение. Методологичната граница на Раздел 2.5 е умишлено тясна. Дисертацията не имплементира модели на темпорална мрежа, зависими от паметта, като емпирични компоненти. Вместо това, тя използва насочени темпорални каскади, защото те се поддържат от запазените набори от данни и се свързват директно с експерименталния конвейер: визуализация на темпорална мрежа, $I_{\text{new}}(t)$, $I_{\text{cum}}(t)$, SIR/SEIZ фитинг, ABC/ABC-SMC посттериорна инференция и краткохоризонтална прогнозна оценка. По-сложните подходи, базирани на пътна памет, остават възможни бъдещи разширения, а не невалидирани компоненти на внедрената методология [52], [76].

2.6 SIR/SEIZ, ABC/ABC-SMC и методология за последващо прогнозиране

Раздел 2.6 определя ядрото на моделирането в дисертацията. SIR предоставя прозрачна базова линия, SEIZ моделира експозицията и скептицизма, ABC оценява несигурността без изрична вероятност, а ABC-SMC прогресивно подобрява последващата концентрация. След това от калибрирания SEIZ слой се извличат последващи прогнози и сценарии за интервенция.



Фигура 2.7. ABC работен процес за калибриране на параметрите на епидемията от фалшиви новини от времеви каскади

2.6.1 Базов модел на SIR

SIR картографира разпространението на фалшиви новини като податливи потребители, активни разпространители и възстановили се или неактивни потребители. Той се запазва, защото е интерпретируем и предоставя базова траектория, спрямо която може да се оцени по-богатият SEIZ модел.

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

2.6.2 SEIZ основен дифузионен модел

SEIZ е основният дифузионен модел, защото включва експонирани и скептични или неразпространяващи се състояния. Тези състояния са латентни моделиращи показатели за забавено приемане, колебание, корекция и липса на ангажираност, а не директни измервания на потребителското доверие.

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

Уравнения (2.12) и (2.13) дефинират векторите на параметрите SIR и SEIZ; SEIZ разширява SIR с термини за експозиция, скептицизъм, корекция и начално състояние.

2.6.3 ABC извод

Нека D_{obs} означава наблюдаваната каскада от фалшиви новини, а θ означава вектора на параметрите на епидемията. В стандартна байесова формулировка, апостериорното разпределение е пропорционално на произведението на вероятността и априорното разпределение:

$$p(\theta | D_{\text{obs}}) \propto p(D_{\text{obs}} | \theta)p(\theta) \quad (2.14)$$

Трудността в уравнение (2.14) е членът на вероятността $p(D_{\text{obs}} | \theta)$. За каскада от фалшиви новини, този член би изисквал вероятността на наблюдаваната последователност от потребителски активации, времената между събитията, повторните публикации, корекциите и скритите неотговори. Тъй като тези елементи не са напълно наблюдавани и не са лесно представени чрез затворен вероятностен модел, ABC замества оценката на вероятността със симулация и обобщаващо-статистическо сравнение [19], [23]-[26].

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

В уравнение (2.15), D_{sim} е симулирана каскада, генерирана от симулатора SIR или SEIZ, $S(D)$ е вектор от обобщени статистики, ρ е функция на разстоянието, а ε е прагът на толерантност ABC. Симулирана стойност на параметъра се приема, когато разстоянието между симулираните и наблюдаваните обобщения е достатъчно малко. Приетите стойности на параметрите апроксимират апостериорното разпределение по параметри на епидемията, което може да възпроизведе наблюдавания модел на разпространение на фалшиви новини.

Таблица 2.15. Кандидатски априорни разпределения за ABC-базирано извеждане на параметрите на епидемията от фалшиви новини

Параметър	Модел	Предишен пример	Интерпретация при разпространението на фалшиви новини
β	SIR/SEIZ	Uniform(0.01, 0.80)	Базова скорост, с която активните разпространители влияят на податливите потребители.
γ	SIR/SEIZ	Uniform(0.01, 0.50)	Скорост, с която разпространителите спират да споделят, забравят съдържанието или губят интерес.
ε	SEIZ	Uniform(0.01, 0.50)	Скорост, с която изложените на риск потребители стават активни разпространители.
p	SEIZ	Beta(2, 2)	Вероятност за незабавно споделяне след

			контакт с активен разпространител.
l	SEIZ	Beta(2, 2)	Вероятност за скептично отношение или неразпространение след експозиция.
δ	SEIZ-modified	Uniform(0.00, 0.30)	Процент на корекции или опровержения, причинени от предупреждения, проверки на фактите или контраинформация.
ρ	SEIZ	Uniform(0.00, 0.60)	Коефициент на взаимодействие между изложените на риск и заразените потребители.

ABC инференцията взема проби от кандидат параметри, симулира каскади и приема стойности на параметрите, които възпроизвеждат наблюдаваните каскадни обобщения в рамките на толеранс. Това позволява на дисертацията да оцени правдоподобните области на предаване, възстановяване, експозиция и скептицизъм, когато точната вероятност за каскадата не е налична [19]-[25].

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

В уравнение (2.16), C_T е краен размер на каскадата, t_{peak} е време до достигане на пик, r_{max} е максималната скорост на разпространение, d_c е дълбочина на каскадата, b_c е ширина на каскадата, τ_{mean} е средното време между събитията и $\bar{p}_f$ е средната вероятност за фалшиви новини, получена от Transformer, за твърдението или каскадният елемент. Разстоянието между наблюдаваните и симулираните обобщения се изразява като:

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

Теглата w_k в уравнение (2.17) определят важността на всяка обобщена статистика. Ако крайният размер на каскадата е основната цел, C_T тя получава по-голяма тежест. Ако е по-важно поведението за ранно предупреждение t_{peak} и r_{max} може да се подчертае, то е по-важно. В Глава 3 тези обобщени статистики подкрепят последваща калибрация, последваща прогнозна симулация и краткосрочно прогнозиране извън извадката. Пълната таблица с обобщени статистики за кандидати е предоставена в Приложение Г.3.

$$p_f(x_i) = P(y_i = \text{fake} | x_i) \quad (2.18)$$

В уравнение (2.18), $p_f(x_i)$ означава вероятността, зададена от детектора на трансформатора, че текстовият елемент x_i принадлежи към класа на фалшивите новини. Резултатът не се използва като компонент на ансамбъла и не замества моделирането на времевата дифузия. Вместо това, той предоставя калибриран сигнал за съдържание-риск, който може да модулира налягането на разпространение, присвоено на наблюдавано времево взаимодействие.

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

В уравнение (2.19), ефективната скорост на предаване се коригира чрез теглото на реброто, вероятността за фалшиви новини от трансформатора и влиянието на локалната мрежа.

2.6.4 ABC-SMC извод

ABC-SMC подобрява отхвърлянето ABC чрез използване на последователни популации от частици и намаляване на толерансите. Методът е подходящ за каскади от фалшиви новини, защото концентрира апостериорната маса около параметрични области, които по-добре възпроизвеждат наблюдавания времеви растеж, като същевременно отчитат несигурността.



Фигура 2.8. Работен процес ABC-SMC за сравняване на SIR и SEIZ върху времеви каскади от фалшиви новини

$$\mathcal{M} = \{\text{SIR}, \text{SEIZ}\} \quad (2.20)$$

Уравнение (2.20) определя двата кандидат-модела за дифузия, сравнени при едни и същи наблюдавани каскадни данни.

$$P_{\text{ABC}}(M \mid D_{\text{obs}}) \propto P(M)P_{\text{ABC}}(D_{\text{obs}} \mid M) \quad (2.21)$$

Уравнение (2.21) не се интерпретира като доказателство, че един модел е универсално превъзходен. Вместо това, дисертацията оценява дали структурата SEIZ осигурява по-добро постфериорно предсказващо калибриране за емпиричните каскади от фалшиви новини, разгледани в Глава 3. Приетите разстояния, постфериорните разпределения на параметрите и постфериорните предсказващи криви се използват за сравняване на практическото калибриране на SIR и SEIZ при едни и същи наблюдавани $I_{\text{new}}(t)$ и $I_{\text{cum}}(t)$ траектории.

2.6.5 Постериорно предсказващо симулиране и прогнозиране

Постериорната прогнозна симулация разпространява приетите параметри напред, за да оцени каскадния растеж в краткосрочен план. Резултатът не е детерминистична прогноза, а интервално представяне на правдоподобното бъдещо разпространение при настроените допускания на модела.

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

В уравнение (2.22) всяка бъдеща траектория се генерира чрез начертаване на приет вектор на апостериорни параметри и симулиране на дифузионния модел напред през прогнозния хоризонт.

Фигура 2.13. ABC калибриране на каскада от фалшиви новини, използвайки изходи на трансформатор и обобщения на темпорална мрежа.

$$\beta_{\text{intervention}} = \alpha_{\beta}\beta, \quad 0 < \alpha_{\beta} < 1; \quad \delta_{\text{intervention}} = \alpha_{\delta}\delta, \quad \alpha_{\delta} > 1 \quad (2.23)$$

Уравнение (2.23) представлява сценарии на интервенция: намаляването на видимостта понижава бета, докато по-силното развенчаване или корекция увеличава делта.

2.6.6 Базов функционален риск от разпространение на фалшиви новини в задна линия

Базовият функционал за дифузионно-рисков риск комбинира вероятността за фалшиви новини от типа Transformer, амплификацията на темпоралната мрежа и апостериорния прогнозен растеж на ABC-SMC-SEIZ. Целта му е синтез и одитируемост: той прави връзката между риска от съдържание, каскадната активност и очакваното бъдещо разпространение изрична.

За твърдение за фалшива новина c , текстов вход x_c , наблюдавана каскада S_c^{obs} , хоризонт на прогнозиране h и размер на наблюдаваната каскадна популация N_c , базовият риск се определя като:

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

В уравнение (2.24), $\hat{p}_f(x_c)$ е базирана на трансформатор вероятност текстовият елемент да е твърдение за фалшива новина или статия с фалшива новина. $A_c(t)$ е резултатът от темпорално-мрежовата амплификация, изчислен от наблюдаваната каскада, λ контролира влиянието на този член на амплификация, $\pi_{\epsilon}^{\text{SMC}}(\theta | S_c^{\text{obs}})$ е ABC-SMC апостериорното разпределение на SEIZ параметрите, $I_{\text{cum},c}^{\text{SEIZ}}(t + h; \theta)$ е прогнозиранят кумулативен брой активни разпространители в хоризонт h , $I_{\text{cum},c}^{\text{obs}}(t)$ е наблюдаваният кумулативен брой активни разпространители до времето t и N_c нормализира резултата според размера на наблюдаваната каскадна популация. Следователно функционалът измерва очаквания нормализиран бъдещ кумулативен растеж при базовата апостериорна линия, мащабиран според риска за съдържанието и мрежовата амплификация.

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

В уравнение (2.25), $I_{\text{new},c}^{\text{norm}}(t)$ е нормализираният брой нови уникални активни разпространители в текущия времеви прозорец, $H_c^{\text{norm}}(t)$ е нормализиран резултат за активност на хъба и $B_c^{\text{norm}}(t)$ е нормализиран резултат за активност, базиран на реле или посредничество. Когато се изисква минимален оперативен резултат, $A_c(t)$ може да се апроксимира, като се използва само нормализираният $I_{\text{new}}(t)$ сигнал. Това поддържа функционала съвместим с емпиричните каскади в Глава 3, където $I_{\text{new}}(t)$, $I_{\text{cum}}(t)$, изходните хъбове и потребителите на релета са наблюдаемите дифузионни изходи.

Ако ABC-SMC върне M апостериорни частици $\{(\theta_m, w_m)\}$ за $m = 1, \dots, M$, уравнение (2.24) може да бъде оценено във форма на частици като:

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

Уравнение (2.26) оценява функционала на риска с ABC-SMC частици чрез осредняване на прогнозирания растеж в последващи проби и техните тегла.

2.6.7 Интеграция с рамката

Интеграцията на рамката свързва четирите слоя в един възпроизводим конвейер. Слой за съдържание идентифицира подозрителни твърдения, темпоралният слой реконструира наблюдавания процес на разпространение, епидемичният слой интерпретира преходите на състоянията, а байесовият слой количествено определя неопределеността на параметрите и риска от последващо прогнозиране.

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

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

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

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

Уравнения (2.27)-(2.30) определят мерките за грешка, използвани за сравняване на наблюдаваните и симулираните криви на разпространение: относително разстояние $L2$, RMSE, MAE и нормализирана RMSE.

2.7 Граници на пригодност, етика и възпроизводимост на набора от данни

Раздел 2.7 определя границите на валидност на рамката. Пригодността на наборите от данни се оценява отделно за задачи за откриване и разпространение, докато етичната обработка изисква анонимизация, съответствие с политиката на платформата, липса на заключения за лични убеждения и ясно разграничение между емпирично наблюдение и контрафактуална симулация.

2.7.1 Набор от данни, подходящ за откриване на фалшиви новини чрез трансформатори

Наборът от данни е подходящ за откриване на трансформатори, когато съдържа надеждни етикети, използваем текст, адекватно представяне на класовете и документиран разделения. Оценката за пригодност се използва като прозрачна рубрика, а не като заместител на експертната оценка.

За систематично сравнение, пригодността на базиран на съдържание набор от данни D_i за откриване на трансформатори може да се изрази качествено като:

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

В уравнение (2.31), $S_T(D_i)$ означава пригодността на набора от данни за откриване на трансформатор D_i . Терминът L_i представлява надеждност на етикета, T_i представлява наличност на текст и пригодност на дължината на документа, B_i представлява баланс на двоичния клас, Q_i представлява качество на текста след предварителна обработка и N_i представлява пригодност на езика или домейна. Теглата w_L , w_T , w_B и w_N са качествени тегла, дефинирани от изследователя. Изразът не се използва като строга числена оценка; той формализира критериите, използвани при определяне кои набори от данни трябва да останат в оперативния слой, базиран на съдържание $.w_Q$

Таблица 2.16. Крайни набори от данни, базирани на съдържание, запазени за откриване чрез Transformer и класическо машинно обучение.

Набор от данни	Основна роля в откриването	Причина за задържане	Методологична граница
Constraint@AAAI2021	Разпознаване на фалшиви/истински кратки текстови съобщения, свързани с COVID.	Поддържа кратка класификация на дезинформацията в стил социални медии при условия на пандемична информация [30].	Използва се за откриване на съдържание, а не за моделиране на времева дифузия.
FRN	Разпознаване на фалшиви/истински статии с дълга форма.	Предоставя дълги новинарски статии и поддържа оценка на трансформатори с дълга последователност, включително моделиране в стил BigBird [12], [81], [95].	Използва се за откриване на ниво статия; не се прави каскадна реконструкция само от корпуса.
ISOT	Голям бенчмарк за сравнение на фалшиви/истински статии на английски език.	Осигурява голям двоичен бенчмарк за сравнение на класически и трансформаторни системи.	Използва се само за анализ на устойчивост на ниво съдържание.
CoAID	Откриване на дезинформация в здравеопазването.	Свързва разкриването на фалшиви новини с дезинформация, свързана със здравето, и частични доказателства за социалния контекст [83], [97].	Използва се главно като набор от данни за съдържание; социалните сигнали не се третират като пълни времеви каскади.
Albanian Fake News Corpus	Разкриване на фалшиви новини в Албания с ниски ресурси.	Поддържа многоезична и нискоресурсна оценка на трансформатори [84],	Използва се за откриване и синтетична

		[85], [98].	валидация, а не като емпирично доказателство за Twitter/X или Ноаху.
--	--	-------------	--

Практическият резултат от това решение за пригодност е контролиран слой за риск от съдържание. Трансформаторните модели, обучени или оценени върху избраните набори от данни за съдържание, могат да генерират вероятностен резултат за фалшиви новини $p_f(x)$, но този резултат е само един компонент от пълната рамка на дисертацията. Той става оперативно значим за дифузионния анализ само когато е свързан с твърдение за фалшиви новини, което също има доказателства за времево взаимодействие.

2.7.2 Удобство за моделиране на темпорална дифузия, SIR/SEIZ и ABC/ABC-SMC

Дифузионният набор от данни трябва да съдържа времеви отметки, каскадни идентификатори, насочени взаимодействия и достатъчен времеви обем, за да поддържа обобщена статистика. Това обяснява защо само главните емпирични каскади се използват за последващ извод, докато по-малките или синтетични случаи остават допълнителни.

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

В уравнение (2.32), $S_D(D_i)$ означава пригодността на набора от данни за темпорална дифузия D_i . Терминът T_i представлява наличност на времеви отпечатък, C_i представлява каскадна структура или структура на ниво твърдение, E_i представлява насочена наличност на ребра, M_i представлява съвместимост със състоянията на SIR/SEIZ, A_i представлява наличност на обобщени статистики на ABC/ABC-SMC и Q_i представлява качеството на данните след предварителна обработка във времето. Както при уравнение (2.31), изразът е методологично ръководство, а не фиксирано числено правило.

Таблица 2.17. Окончателни дифузионно-ориентирани набори от данни, запазени за времево, епидемично и ABC/ABC-SMC моделиране

Набор от данни / каскада	Твърдение или материал за фалшиви новини	Основна дифузионна опора	Роля
COVID-19 истински Twitter/X	„Коронавирусът не само влияе на начина ни на живот, но и драстично влияе на начина, по който умираме.“	Взаимодействията между потребители с времеви отпечатък са филтрирани до <code>site_type =</code>	Основна каскада от емпирични дезинформации в здравеопазването за темпорална мрежа,

		covid_fake_claim.	SEIZ/ABC-SMC и $I_{cum}(t)$ прогнозиране.
Политически измами / OSoMeNet	„Три милиона гласа на президентските избори бяха подадени от нелегални имигранти.“	Масштабна каскада от ретуитове с насочени взаимодействия между потребители.	Основна емпирична каскада от твърдения за политически фалшиви новини за SIR/SEIZ, ABC/ABC-SMC и сравнение на прогнозите.
Социална каскада, внесена от измама	„Трябва ли човек да прекрати кариерата си с туйт?“	Социална каскада с етикет „Измама“ и записи за разпространение, маркирани с времеви отпечатък.	Допълнителна сравнителна каскада от твърдения за социални фалшиви новини за интерпретация във времевата мрежа.
Синтетична албанска темпорална каскада	„Shqipëria do të presë 100 mijë palestinezë nga Gaza.“	Генерирана каскада с времеви отпечатък със записи за проверка на твърдения и факти.	Подкрепа за валидиране и симулация на албанския тръбопровод с ниски ресурси; не се третира като емпирично доказателство от Twitter/X.

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

В уравнение (2.33), C_T е крайният размер на каскадата, t_{peak} е времето до достигане на пик на дифузия, r_{max} е максималната скорост на разпространение, d_c е дълбочината на каскадата, b_c е широчината на каскадата, q_f е вероятността за фалшиви новини от трансформатора, свързана с елемента, и n_Z е броят на скептичните, коригиращите или свързаните с проверка на фактите събития, когато такава информация е налична. Точният вектор се адаптира към всеки набор от данни, но ролята на вектора остава същата: той преобразува наблюдавана каскада от твърдения за фалшиви новини в обобщени доказателства, които могат да бъдат сравнени със симулирани траектории на SIR/SEIZ в процедурите ABC и ABC-SMC [19]-[26], [35], [86]-[89].

2.7.3 Етични съображения, съображения за поверителност и възпроизводимост

Рамката използва отговорно наблюдаеми публични или документиран записи за взаимодействия. Потребителските идентификатори се обработват съгласно ограниченията на набора от данни, а резултатите от модела не се интерпретират като лични убеждения, убеждения или намерения.

2.7.4 Методологични допускания и граници на валидност

Основните допускания са, че вероятностите на трансформатора са оценени от модела рисковни оценки, каскадните взаимодействия са наблюдаеми заместители, състоянията

на SEIZ са латентни моделиращи конструкции, а сценариите за интервенция са симулации. Тези граници правят резултатите защитими, а не преувеличени.

Таблица 2.18. Внедрени компоненти и граници само за литература или бъдеща работа

Компонент	Статус в тази дисертация	Причина за статуса
BERT, RoBERTa, BigBird, mBERT и XLM-RoBERTa	Внедрено и оценено	Те осигуряват оперативния слой за откриване на Transformer и генерират моделно оценената вероятност за фалшиви новини, използвана от рамката.
Дифузионни модели SIR и SEIZ	Внедрено и оценено	Те предоставят интерпретируеми представяния на епидемичното състояние за активно разпространение, спиране, експозиция и скептицизъм.
ABC и ABC-SMC	Внедрено и оценено	Те количествено определят неопределеността в параметрите на дифузия, когато точната оценка на вероятността не е осъществима.
GNN и обучение по темпорални графи	Преглед на литературата и бъдеща работа	Те изискват по-богати графични данни на потребителско ниво, надеждни мрежи от последователи или взаимодействия и последователни етикети за разпространение, отколкото са налични във всички запазени набори от данни [121]-[129], [153]-[155].
Обясним изкуствен интелект	Бъдещо обяснително разширение	LIME, SHAP и обяснимите системи за фалшиви новини са релевантни, но пълен XAI слой би изисквал отделен протокол за обяснение и валидиране [115]-[117], [152], [172].
Големи езикови модели	Само спомагателна или поддържаща роля	LLM могат да подпомогнат допълването или анализа, но крайните етикети, наборите за оценка и дифузионните интерпретации не се генерират от LLM изхода [47]-[49], [135]-[136], [143]-[145].
Ансамбли за гласуване, подреждане и класификатор	Само преглед на литературата	Те могат да подобрят точността, но биха намалили проследимостта между оценката на Трансформър и интерпретацията на риска от дифузия надолу по веригата.

2.8 Резюме на главата

Глава 2 превръща прегледа на литературата в оперативна методология. Тя определя ролите на наборите от данни, предварителната обработка, времевите каскади, SIR/SEIZ моделирането, ABC/ABC-SMC изводите, посттериорното прогнозиране, сценариите за интервенция и етичните граници. Глава 3 след това оценява този процес емпирично.

Глава 3. Експериментални резултати и Дискусия

3.1 Въведение

Глава 3 представя емпиричното изпълнение на рамката. Първо се оценява бинарното откриване, базирано на трансформатор, след което се реконструират избрани каскади от фалшиви новини, се напасват SIR и SEIZ модели, се оценява апостериорната неопределеност с ABC/ABC-SMC и се оценяват сценарии за контрафактуална интервенция. Главата разделя доказателствата на ниво съдържание и на ниво дифузия, преди да ги интегрира в окончателното обсъждане.

3.2 Експериментална настройка

Експерименталната настройка следва последователността „детектиране-дифузия-извод“. Наборите от данни за съдържание се оценяват на ниво елемент, използвайки точност, прецизност, изчерпаемост, F1-оценка и ROC-AUC; наборите от данни за дифузия се оценяват на каскадно ниво, използвайки криви на разпространение, подредени по време, SIR/SEIZ съпоставяне и posterior prognostic behavior. Двете основни емпирични каскади са случаят с COVID-19 Twitter/X и случаят с политическото твърдение Ноаху / OSoMeNet [30], [81]-[85], [90], [108].

Таблица 3.1. Експериментални доказателства и набори от данни

Блок	Набори от данни	Изход
Откриване	Ограничение, FRN, ISOT, CoAID, албански	F1, ROC
Времеви	COVID в Twitter, Измама	Каскади
Епидемия	COVID в Twitter, Измама	SIR/SEIZ
Байесов	COVID в Twitter, Измама	Заден (апостериорен)

3.3 Резултати от бинарно откриване на базата на трансформатори и сравнителен анализ

Раздел 3.3 представя слоя за откриване. Целта не е да се класират всички възможни детектори на фалшиви новини, а да се покаже как се държат избраните трансформаторни архитектури в набори от данни за кратки текстове, дълги документи, здравеопазване и албански набори от данни с ниски ресурси. Резултатите предоставят вероятностни изходи за класа на фалшивите новини за по-късния каскаден гейт.

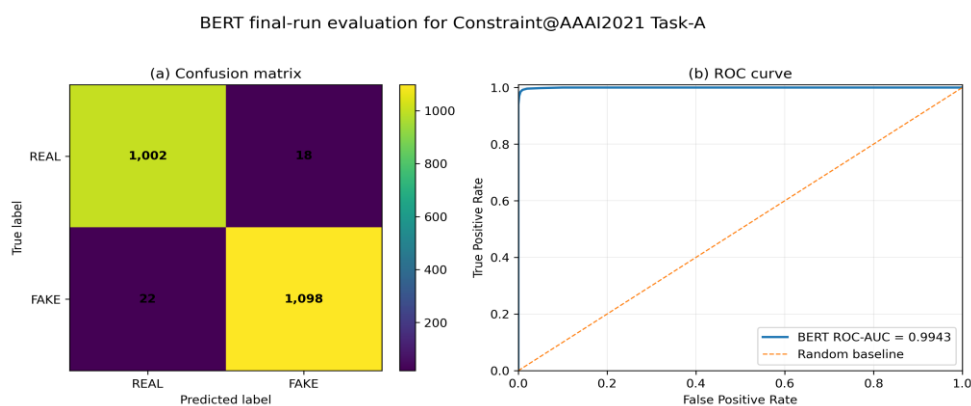
3.3.1 Резултати върху набори от данни, базирани на съдържание

Таблица 3.2. Производителност на генерирано от автора откриване на Constraint@AAAI2021 Задача-A

Модел / метод	Точност	Прецизност	Припомняне	Резултат от F1
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%
Логистична регресия (LR)	92,07%	92,11%	92,00%	92,00%
Машина с опорни вектори (SVM)	91,09%	91,13%	91,00%	91,00%
Наивен Байес (NB)	93,01%	93,06%	93,00%	93,00%
NB + LR	93,01%	93,05%	93,00%	93,00%

В задача А на Constraint@AAAI2021, BERT постигна най-силния резултат за дезинформация за COVID-19 в кратки текстове, с точност от 98,11% и F1-оценка от 98,00%. Матрицата на объркване и ROC кривата подкрепят заключението, че контекстуалните представяния се подобряват в сравнение със силните класически базови линии за тази задача.



Фигура 3.1. Оценка на финалния цикъл на BERT за Constraint@AAAI2021 Задача А: (а) матрица на объркване; (б) ROC крива.

Фигура 3.1 комбинира двете допълващи се диагностики на най-добре представящия се BERT детектор в Constraint@AAAI2021 Task-A. Панел (а) показва матрица на объркване с ниска степен на грешки, с 1002 реални съобщения и 1098 фалшиви съобщения, правилно класифицирани, и само 18 фалшиви положителни и 22 фалшиви отрицателни резултата. Панел (б) показва силно, независимо от прага, разделяне, с ROC-AUC = 0,9943. Заедно двата панела потвърждават, че слоят за откриване на кратък текст осигурява надежден сигнал за риск от съдържание преди времеви каскаден анализ.

3.3.2 Резултати от откриване на дълъг текст и специфични за домейна данни

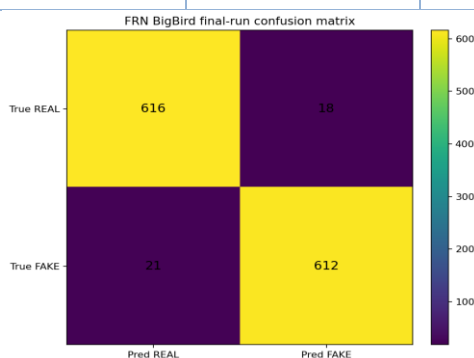
За дълги и специфични за дадена област набори от данни, производителността на Transformer зависи от дължината на текста и темата. BigBird-RoBERTa се представи най-добре при по-дългите новини на FRN и CoAID, RoBERTa се представи най-добре при ISOT и кратките твърдения за здравеопазване, а класическите базови линии на TF-IDF останаха полезни отправни точки. Фигурите и таблиците са запазени, защото показват качеството на класирането и разпределението на грешките.

Таблица 3.3. Компактно сравнение на най-силните класически базови линии в FRN, ISOT и CoAID.

Набор от данни / подмножество	Най-добра класическа базова линия	Фалшива Формула 1	Макро-F1	ROC-AUC
ФРН	Линеен SVM + TF-IDF	0.9531	0.9526	0,9903
ИЗОТ	Линеен SVM + TF-IDF	0,9924	0,9921	0,9996
Публикации/искания за социални мрежи, свързани с CoAID	Логистична регресия + TF-IDF	0.9358	0.9488	0.9830
По-дълги новини за CoAID	Логистична регресия + TF-IDF	0.4615	0.7243	0.9328

Таблица 3.4. Сравнителна производителност на откриване на трансформатори при FRN, ISOT и CoAID.

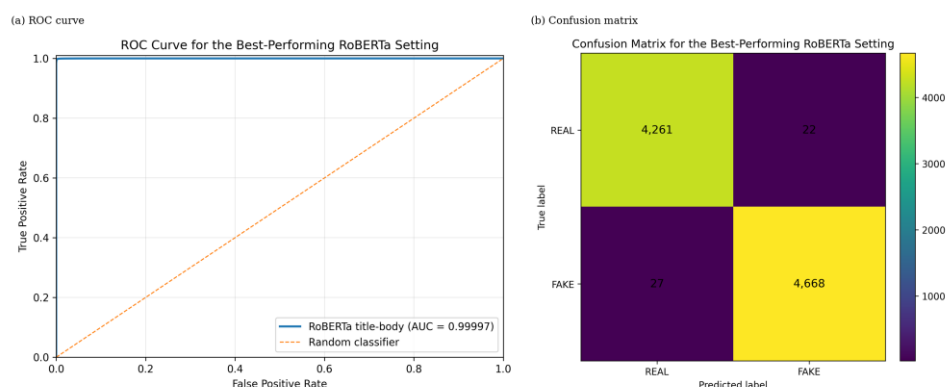
Набор от данни / подмножество	Тип текст	Най-добра настройка на трансформатора	Фалшива Формула 1	Макро-F1	ROC- AUC
ФРН	Дълги политически/общии статии	BigBird- RoBERTa-base	0.9691	0.9692	0,9946
ИЗОТ	Референтен показател за големи статии на английски език	Заглавие на RoBERTa	0,9948	0,9945	0,99997
Публикации/искания за социални мрежи, свързани с CoAID	Кратки твърдения/публикации за здравеопазване	RoBERTa-база	0.9615	0.9701	0,9915
По-дълги новини за CoAID	По-дълги новини за здравеопазването	BigBird- RoBERTa-base	0.8205	0.9081	0.9786



Фигура 3.2. Матрица на объркване в последния цикъл на BigBird за FRN

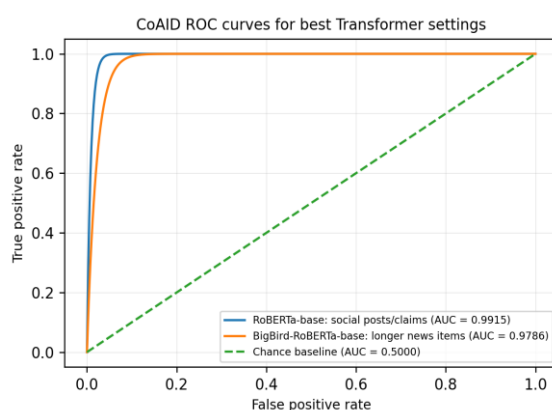
Фигура 3.2 потвърждава силно разделяне на класовете за модела FRN BigBird. Само 18 реални статии са неправилно маркирани като фалшиви, а 21 фалшиви статии са пропуснати. Този модел на грешки показва силно разделяне на класовете и балансиран компромис между отзоваването на фалшиви класове и фалшиво положителния

контрол, което е важно, когато изходният резултат от откриването по-късно се интерпретира като сигнал за риск от съдържание.



Фигура 3.3. Допълнителни графики за оценка на най-добре представящата се ISOT настройка: (а) ROC крива; (б) матрица на объркване

Фигура 3.3 показва, че най-добре представящата се ISOT конфигурация (RoBERTa title-body) постига както почти перфектно разделяне на класирането, така и много нисък профил на грешки. ROC кривата в панел (а) показва отлична дискриминационна способност (ROC-AUC = 0.99997), докато матрицата на объркване в панел (б) показва само 22 реални статии, погрешно класифицирани като фалшиви, и 27 фалшиви статии, погрешно класифицирани като реални. Взети заедно, тези графики потвърждават, че моделът се представя изключително добре по ISOT бенчмарка, въпреки че резултатът все пак трябва да се тълкува с повишено внимание поради специфичните за набора от данни закономерности на източника и стила.



Фигура 3.4. ROC криви за най-добрите настройки на CoAID трансформатора

Фигура 3.4 обобщава най-силните резултати от класацията на CoAID Transformer за кратки публикации/твърдения в здравеопазването и по-дълги новинарски материали.

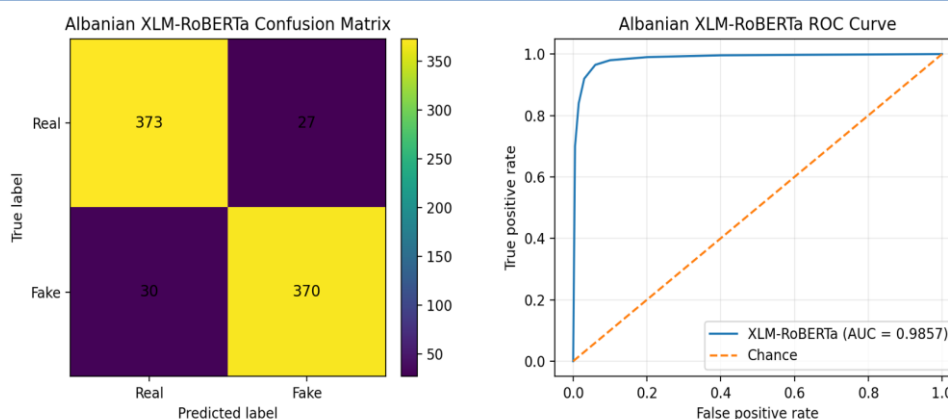
3.3.3 Многоезични резултати с ограничени ресурси

Албанският експеримент показва, че многоезичните контекстуални модели са необходими за откриване на фалшиви новини с ниски ресурси. XLM-RoBERTa даде

най-силен резултат, превъзхождайки както класическата базова линия на TF-IDF, така и mBERT, което подкрепя включването на многоезичната оценка на Transformer в рамката.

Таблица 3. 4. Производителност на многоезичното разпознаване на албански език

Модел	Фалшива Формула 1	ROC-AUC
Линейна SVM + дума/знак TF-IDF	0.8802	0.9516
Многоезичен mBERT-базиран	0.9057	0.9741
XLM-RoBERTa-база	0.9285	0.9857



Фигура 3.5. Резултати от оценката на албанския XLM-RoBERTa: матрица на объркване и ROC крива. Фигура 3.5 представя матрицата на объркване и ROC кривата на най-добре представящата се конфигурация XLM-RoBERTa.

3.3.4 Сравнителен анализ на откриването

В различните набори от данни няма един-единствен модел или политика за въвеждане, която да е оптимална. BERT е ефективен за кратки публикации, RoBERTa за бенчмаркове и твърдения за здравни проблеми, BigBird за по-дълги статии и XLM-RoBERTa за албански текст. Сравнението подкрепя използването на вероятността на Transformer като контролиран сигнал за риск от съдържание, а не като универсална окончателна преценка.

Таблица 3.5. Сравнение на най-силните настройки за откриване на трансформатори, генерирано от автора, между набори от данни

Набор от данни / подмножество	Тип текст	Най-добра настройка на трансформатора	Фалшива Формула 1	Макро-F1	ROC-AUC
Ограничение@AAAI2021	Кратки публикации за COVID-19	BERT-база	0.9800	0.9800	0,9943
ФРН	Дълги политически/обществени статии	BigBird-RoBERTa-base	0.9691	0.9692	0,9946

ИЗОТ	Референтен показател за големи статии на английски език	Заглавие на RoBERTa	0,9948	0,9945	0,99997
Социални публикации/твърдения за CoAID	Кратки твърдения/публикации за здравеопазване	RoBERTa-база	0.9615	0.9701	0,9915
По-дълги новини за CoAID	По-дълги новини за здравеопазването	BigBird-RoBERTa-base	0.8205	0.9081	0.9786
Албански корпус на фалшивите новини	Албански новини с ниски ресурси	XLM-RoBERTa-база	0.9285	0.9287	0.9857

Таблица 3. 6. Основни модели на фалшиво положителни и фалшиво отрицателни резултати в наборите от данни за откриване, базирани на съдържание

Набор от данни / група	Основен модел на фалшиво положителни резултати	Основен фалшиво отрицателен модел	Импликация на рамката
Кратки публикации и твърдения за COVID-19	Истинските здравни публикации с алармиращ речник могат да наподобяват дезинформация.	Фалшиви твърдения, написани с предпазлив или официално звучащ език, може да бъдат пропуснати.	Необходима е висока степен на отзоваване на фалшиви класове преди времево наблюдение.
Дълги английски статии	Статиите с истински мнения може да изглеждат като ненадеждни източници.	Дългите фалшиви статии могат да крият подвеждащи доказателства извън съкращения вход.	Плъзгащите се прозорци и BigBird намаляват загубата на доказателства, свързана с дължината.
По-дълги новини за здравеопазването	Легитимните доклади за COVID-19 може да съдържат рисков речник, подобен на фалшивите твърдения.	Неутралната биомедицинска формулировка може да прикрие подвеждащо съдържание.	По-дългите здравни новини изискват както качество на класиране, така и запомняемост от фалшив клас.
Албански новини с ниски ресурси	Емоционално натоварените истински албански текстове могат да бъдат маркирани като фалшиви.	Формалният журналистически стил може да направи фалшивите албански текстове да изглеждат достоверни.	Токенизирането, съобразено с езика, е необходимо за надеждно многоезично откриване.

3.4 Резултати от темпоралната каскада

Раздел 3.4 преминава от откриване на ниво елемент към времево разпространение. Твърдението се прехвърля към каскаден анализ само когато вероятността за фалшив

клас надвиши прага и когато съществуват достатъчно доказателства за взаимодействие. Този портал предотвратява използването само на записи със съдържание като входни данни за епидемичен модел.

3.4.1 Детекторен портал и избор на каскада

Порталът за откриване избира две основни емпирични каскади за моделиране: твърдението за фалшивите новини в Twitter/X за COVID-19 и твърдението на Ноаху / OSoMeNet за „три милиона гласа“. Случаите, внесени от Ноаху, и синтетичните албански случаи са запазени като помощни примери за сравнение или валидиране, а не като равностойни доказателства за последващ извод.

Таблица 3.7. Детекционен портал и решение за каскаден избор преди моделиране на темпорална мрежа

Случай	$p(fake x)$	Доказателства за взаимодействие	Решение
Твърдение за фалшиви новини за COVID-19 в Twitter/X	0.914	Насочени взаимодействия в Twitter/X с времеви отпечатък	Каскаден анализ; прехвърлен към SIR/SEIZ
Измама / OSoMeNet Твърдение за „три милиона гласа“	0.963	Каскада от ретуитове с обхват на твърдение	Каскаден анализ; прехвърлен към SIR/SEIZ
„Трябва ли човек да сложи край на кариерата си чрез туит?“ Твърдение, внесено от фалшиви източници	0.874	Малка темпорална каскада в стил Ноаху	Само сравнителна траектория
Твърдение за синтетични албански фалшиви новини	0.918	Генерирана темпорална схема в стил Ноаху	Само валидиране

3.4.2 Наблюдавани траектории на заразени потребители

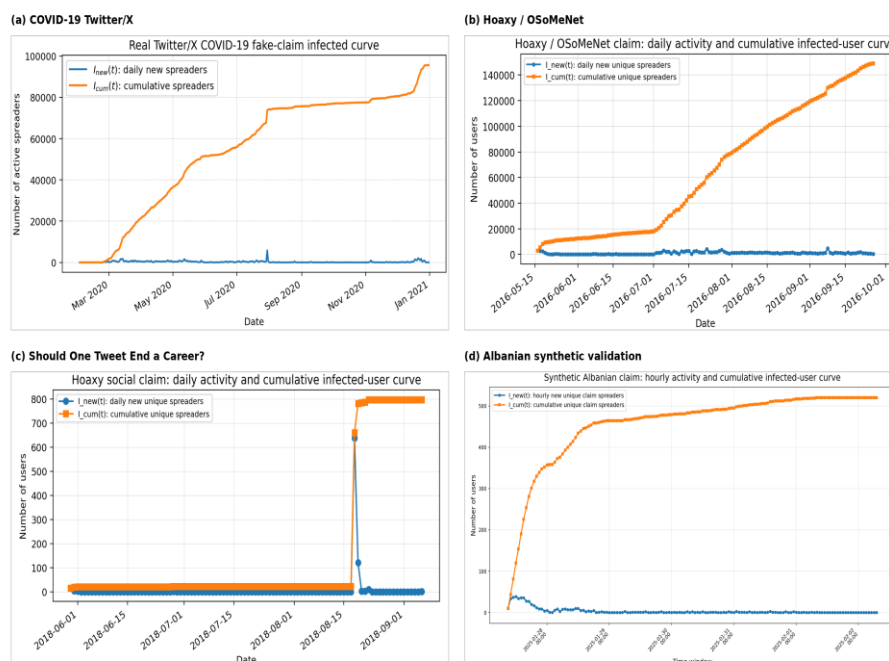
След каскаден подбор, времевите доказателства се редуцират до две директно наблюдавани серии. $I_{new}(t)$ означава броя на потребителите, които за първи път се появяват като активни разпространители в рамките на избрания прозорец за агрегиране, докато $I_{cum}(t)$ означава кумулативния брой уникални активни разпространители до този момент. В емпиричните случаи Twitter/X и Ноаху, активното разпространение се извлича от доказателства за ретуитване, цитиране или насочено споделяне; в албанския случай то се извлича от синтетичните записи, етикетирани с твърдения, използвани само за валидиране.

Наблюдаваните траектории показват, че каскадният риск не може да се изведе само от крайния размер. Случаят с COVID-19 показва повтарящи се изблици и достига 95 599 кумулативни активни разпространители, докато политическото твърдение „Ноаху“

достига 149 079 кумулативни активни разпространители. Тези криви се превръщат в емпирични входни данни за SIR/SEIZ фитинга.

Таблица 3.8. Каскадни обобщени статистики, използвани за прехода SIR/SEIZ

Случай	Врх I_{new}	Финал I_{cum}	Записи	Роля в по-късен анализ
Твърдение за фалшиви новини за COVID-19 в Twitter/X	5 892	95 599	133 374	Основна емпирична каскада
Измама / OSoMeNet Твърдение за „три милиона гласа“	4 660	149 079	1 048 575	Основна емпирична каскада
„Трябва ли човек да прекрати кариерата си с твит?“	638	797	817	Спомагателно сравнение
Твърдение за синтетични албански фалшиви новини	39	520	520	Само валидиране на тръбопровода



Фигура 3.6. Наблюдавани траектории на заразени потребители за избрани каскади от фалшиви новини:

(а) COVID-19 Twitter/X, (б) Hoaxy / OSoMeNet, (в) „Трябва ли един твит да сложи край на кариерата си?“ и (г) Албанска синтетична валидация

Фигура 3.6 обобщава наблюдаваните инкрементални и кумулативни траектории на заразените потребители, извлечени от избраните каскади. Случаят с Ноаху показва най-голям обхват на разпространение, докато случаят с COVID-19 показва повтарящи се пикове на активност за по-дълъг период на наблюдение. Каскадата „Трябва ли човек да сложи край на кариерата си?“ демонстрира забавено поведение на пикове, докато албанският пример е запазен само за валидиране на синтетичен конвейер.

3.4.3 Представителни структури на темпоралната мрежа

Представителните темпорални мрежи показват как разпространението на фалшиви новини се оформя от посоката, времето и позицията в мрежата. Фигурите са запазени, за да се демонстрира, че моделната единица не е изолиран текстов запис, а подредена във времето структура на взаимодействие на ниво твърдение.



Фигура 3.7. Примери за основни темпорални мрежи за двете основни емпирични каскади: (а) COVID-19 Twitter/X и (б) Hoaxy / OSoMeNet

Следователно, раздел 3.4 завършва моста от откриване към времево моделиране. Трансформаторните вероятности установяват кои твърдения преминават през скрининга за фалшиви новини; взаимодействията с времеви отпечатъци превръщат тези твърдения в каскади; $I_{\text{new}}(t)$ и $I_{\text{cum}}(t)$ предоставят емпиричните криви на разпространение; а представителните мрежи потвърждават, че основните случаи имат насочени дифузионни структури, подходящи за SIR/SEIZ анализ. Следващият раздел използва траекториите на COVID-19 и Hoaxy / OSoMeNet като наблюдавани $I_{\text{obs}}(t)$ криви за сравнение на епидемични модели.

3.5 Резултати от моделирането на разпространението на епидемията

Раздел 3.5 представя резултатите от моделирането на епидемията, получени след трансформирането на времевите каскади в Раздел 3.4 в кумулативни траектории на активния разпространител. Целта на този раздел е емпирична: той представя как са се държали базовата линия на SIR и моделът SEIZ, когато са били съобразени с наблюдаваните $I_{cum}(t)$ криви на двете основни каскади от фалшиви новини. Методологичните дефиниции на отделенията, правилата за преход, уравненията, разстоянието на съпоставяне и допусканията на симулатора не се повтарят тук, защото са установени в Глава 2. Сравнението SIR-SEIZ, представено тук, разширява изследването на автора SoftCOM 2025 относно идентифицирането на фалшиви новини и сравнението на епидемични модели във времеви мрежи на Twitter [88].

Раздел 3.5 оценява дали моделите на епидемични състояния апроксимират наблюдаваните каскади. SIR се използва като прозрачна базова линия, докато SEIZ се очаква да улови закъснението и скептицизма чрез допълнителни латентни състояния. Сравнението се основава на пригодени криви на разпространение и грешки на ниво траектория.

Следователно резултатите са представени чрез производителност на ниво крива, а не чрез ново теоретично извеждане. Ключовите величини са относителната грешка на L2 апроксимация, RMSE, MAE и крайният апроксимиран кумулативен брой активни разпръсквачи. По-ниските грешки показват, че моделът по-точно реконструира наблюдаваната $I_{cum}(t)$ траектория. Централният въпрос е дали допълнителната SEIZ структура води до емпирично подобрене, достатъчно голямо, за да оправдае използването ѝ в последващия етап на калибриране на ABC и ABC-SMC.

3.5.1 Резултати от моделното напаиване

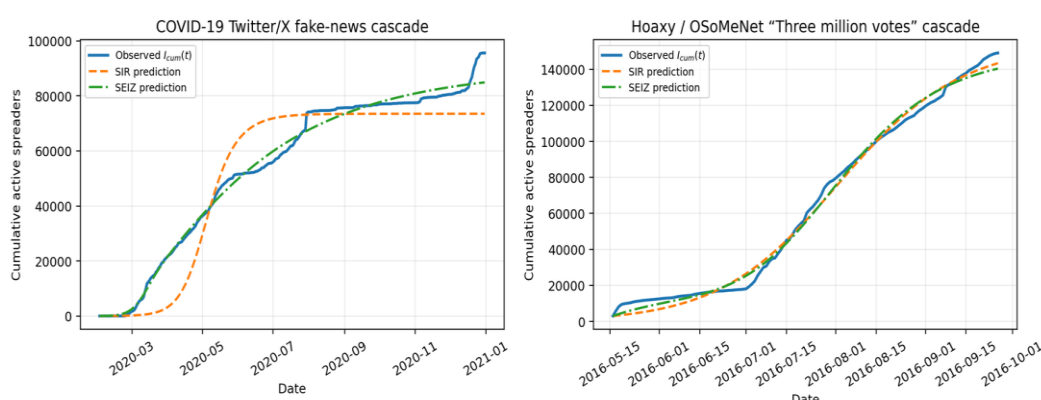
Таблица 3.9. Обобщение на резултатите от апроксимацията на епидемичния модел за двете емпирични каскади от фалшиви новини

Набор от данни / емпирична каскада	Модел	Относителен L2	RMSE	MAE	Крайно наблюдавано / монтирано I_{cum}
Твърдение за фалшиви новини за COVID-19 в Twitter/X	СЪР	0.1626	9 958,4	7 793,3	95 599 / 73 510
Твърдение за фалшиви новини за COVID-19 в Twitter/X	ЗАХВАЩАНЕ	0,0463	2 838,9	2 092,6	95 599 / 84 921
Измама / OSoMeNet Твърдение за „три	СЪР	0,0511	4 150,2	3 560,2	149 079 / 143 396

милиона гласа“ – фалшива новина					
Измама / OSoMeNet	ЗАХВАЩАНЕ	0,0469	3 809,7	3 200,6	149 079 / 140 408
Твърдение за „три милиона гласа“ – фалшива новина					

Резултатите от апроксимацията показват, че SEIZ следва емпиричните каскадни траектории по-точно от SIR, особено когато растежът е неравномерен или многократно изригващ. Това подкрепя тезата, че експонираните и скептични/неразпространяващи се състояния подобряват моделирането на разпространението на фалшиви новини.

SIR versus SEIZ fitted cumulative fake-news spreading curves



Фигура 3.8. SIR спрямо SEIZ, апроксимирани криви на кумулативно разпространение на фалшиви новини за двете емпирични каскади

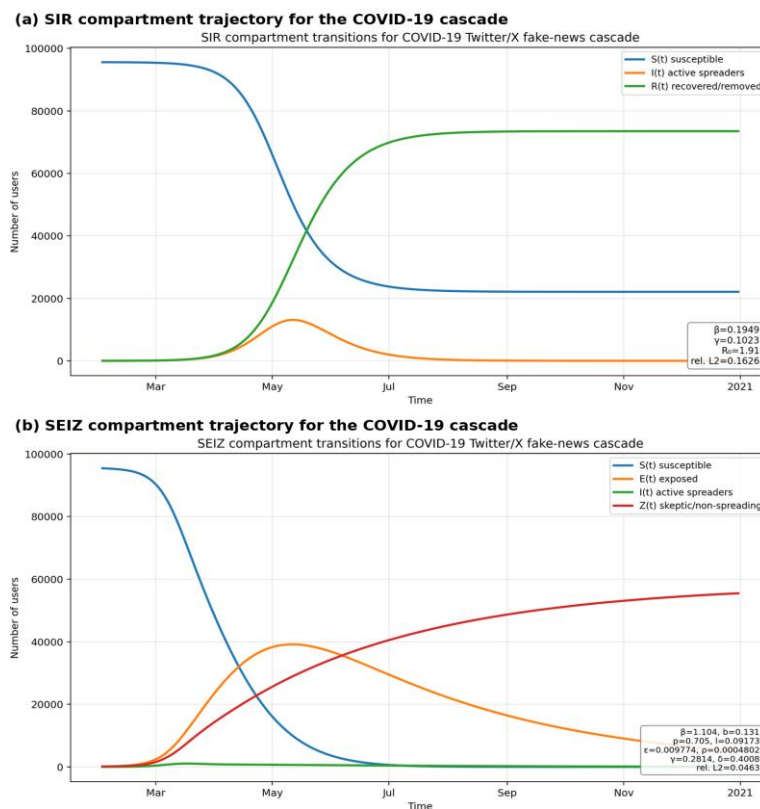
Фигура 3.8 сравнява наблюдаваните кумулативни траектории на разпространение с прогнозите на SIR и SEIZ за двете емпирични каскади. В случая на COVID-19, кривата на SIR достига насищане твърде рано, докато кривата на SEIZ следва наблюдаваната траектория по-точно в основния интервал на растеж. В случая на Ноаху и двата модела са по-близо до емпиричната крива, но траекторията на SEIZ все още дава малко по-ниска обща грешка при напасване. Следователно фигурата визуално потвърждава числения модел, представен в Таблица 3.10.

3.5.2 Сравнителен анализ на SIR-SEIZ

Сравнението SIR-SEIZ показва, че една проста структура „податливи-заразени-възстановени“ е полезна за интерпретация, но е твърде ограничаваща за много каскади от фалшиви новини. SEIZ осигурява по-добър баланс между интерпретируемост и поведенчески реализъм, защото включва забавяне на експозицията, колебание и спиращи ефекти.

3.5.3 Интерпретация на прехода между отделения

Интерпретацията на компартментите остава предпазлива: експонираните и скептичните състояния не са пряко наблюдавани потребителски убеждения. Те са моделиращи показатели, които помагат да се обясни видимото разпространение, потиснатата активация и свързаният с корекцията спад при непълно наблюдение на платформата.



Фигура 3.9. Представителна интерпретация на прехода в компартменти за каскадата COVID-19: траектории, съобразени с SIR и SEIZ

Фигура 3.9(a)(b) илюстрира представителната динамика на компартментите за каскадата от фалшиви новини за COVID-19. В сравнение със SIR, SEIZ разпределя потребителите в експонирани и скептични/неразпространяващи се състояния, което позволява забавеното активиране и неразпространяващото се поведение да бъдат представени изрично. Фигурата е включена като компактен представителен пример, защото Ноаху компартментните графики водят до същото интерпретативно заключение с по-малка пределна полза; подробните траектории на Ноаху компартментите следователно могат да останат в допълнителния материал, вместо да заемат допълнително място в главата с основните резултати. Допълнителни Ноаху / OSoMeNet компартментни преходни графики са представени на Фигури Н.8 и I.9 от Приложение Н.

Резултатът от сравнението на модели също така изяснява как Раздел 3.5 подкрепя по-широкия аргумент на дисертацията. Трансформаторното откриване идентифицира кандидат-съдържание на фалшиви новини, докато темпорално-мрежовият слой показва дали това съдържание се развива в наблюдаема каскада. Епидемичният слой след това тества дали получената $I_{\text{cum}}(t)$ крива може да бъде реконструирана чрез проста базова линия или дали е необходима допълнителна латентна дифузионна структура. И в двата емпирични случая SEIZ осигурява по-силната основа за този мост, защото запазва наблюдаваната кумулативна траектория по-ефективно от SIR и остава интерпретируем за по-късен последващ анализ.

3.6 Резултати от ABC и ABC-SMC изводи

Раздел 3.6 представя извод, съобразен с неопределеността. Вместо да се разчита на един вектор от параметри, ABC и ABC-SMC оценяват области на последващите параметри, които могат да генерират симулирани каскади, близки до наблюдаваните обобщения. Това е от съществено значение, тъй като експозицията, скептицизмът и корекцията са само частично наблюдаеми.

3.6.1 Настройка на инференциалния извод и апостериорна оценка

Резултатите от апостериорната оценка предоставят приети параметри, достоверни интервали и сравнение между основните емпирични каскади. Байесовите резултати показват, че няколко комбинации от параметри могат да бъдат правдоподобни, което оправдава отчитането на неопределеност, а не на детерминистични коефициенти на дифузия.

Таблица 3.10. Наблюдавани каскадни обобщени статистики, използвани за ABC извод

Набор от данни	Краен размер	Пиков ден	Връх I_{new}	Време до достигане на пик
Твърдение за фалшиви новини за COVID-19 в Twitter/X	95 599	179	5 892	179
Измама / OSoMeNet Твърдение за „три милиона гласа“	149 079	115	4 660	115

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

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

Уравнения (3.1) и (3.2) дефинират разстоянието по Глава 3 ABC: несъответствието на траекторията е комбинирано с несъответствие на обобщената статистика, като по-голяма тежест се дава на пълната крива.

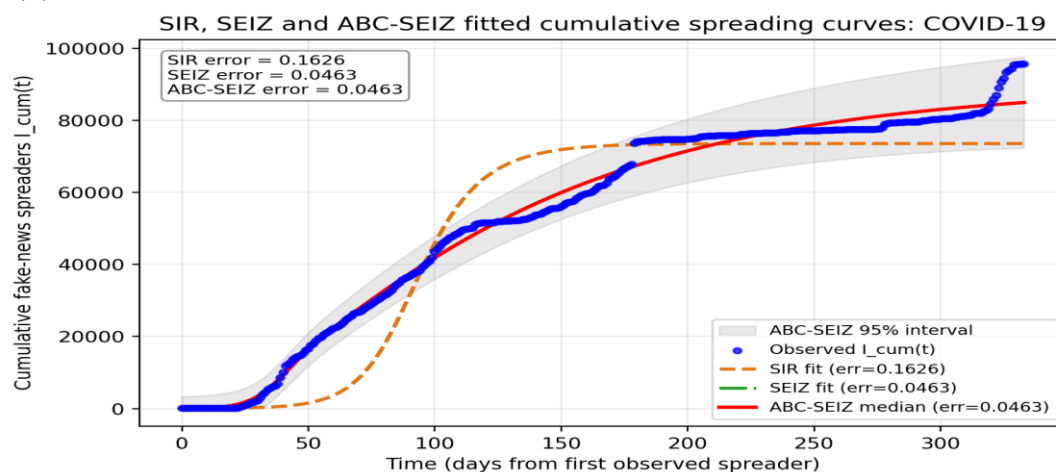
Таблица 3.11. Обобщение на последващите оценки за ABC-SIR и ABC-SEIZ

Набор от данни	Модел	Частици	Средно разстояние	Толерантност	Ключови параметри
COVID	ABC-SIR	300	0,175	0.186	$\beta=0,195$; $\gamma=0,102$; $R_0=1,92$
COVID	ABC-SEIZ	400	0,071	0,092	$p=1.076$; $b=.129$; $p=.698$; $l=.0915$; $\gamma=.271$; $\delta=.399$
Измама	ABC-SIR	300	0,060	0,075	$\beta=0,055$; $\gamma=0,005$; $R_0=10,6$
Измама	ABC-SEIZ	400	0,055	0,064	$\beta \approx 0$; $b=2.000$; $p=.0118$; $l=.0359$; $\varepsilon=.0428$; $\gamma=.0086$

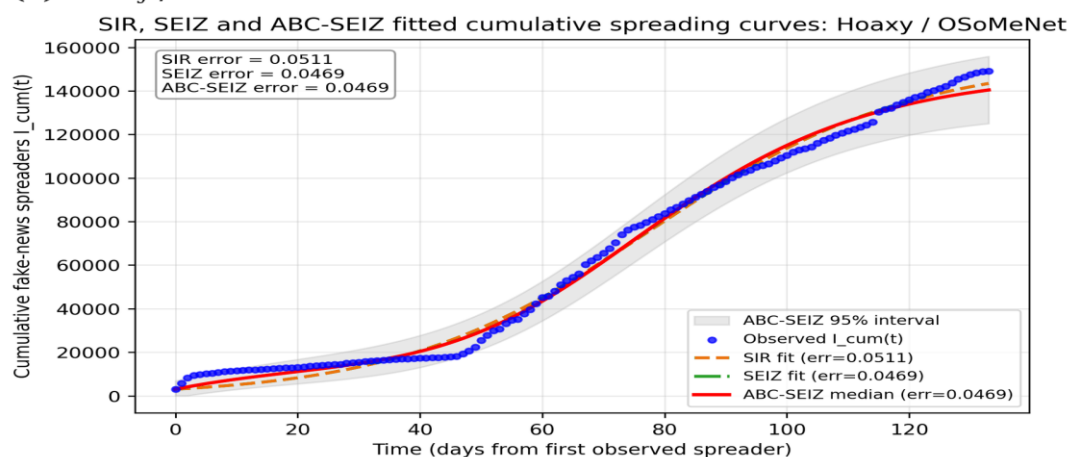
3.6.2 Постериорна предиктивна валидация

Постериорната прогнозна валидация оценява дали приетите параметри възпроизвеждат наблюдаваните криви на разпространение. Запазените фигури показват, че ABC-калибрираните SEIZ симулации предоставят по-информативна обвивка на неопределеността, отколкото единична детерминистична крива.

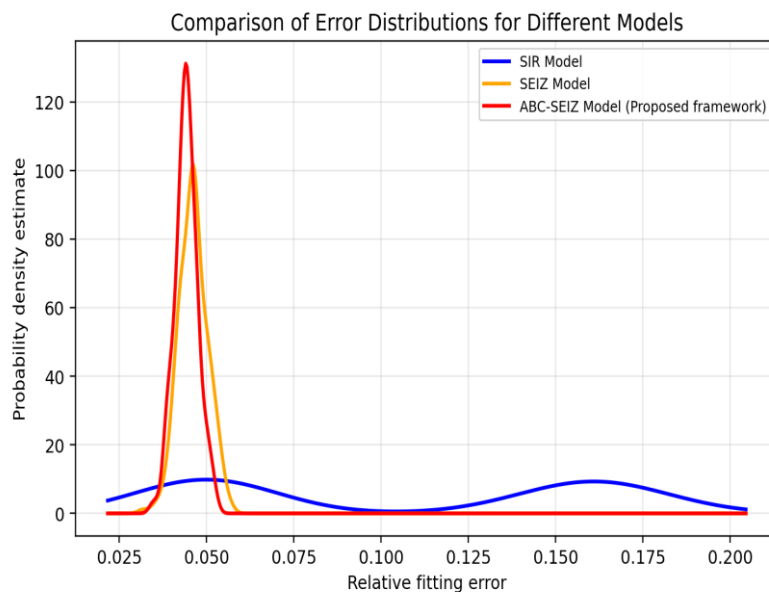
(a) COVID-19 Twitter/X cascade



(b) Hoaxy / OSoMeNet cascade



Фигура 3.10. SIR, SEIZ и ABC-SEIZ са изградили кумулативни криви на разпространение на фалшиви новини за двете емпирични каскади



Фигура 3.11. Разпределения на грешките при апроксимация за модели SIR, SEIZ и ABC-SEIZ.

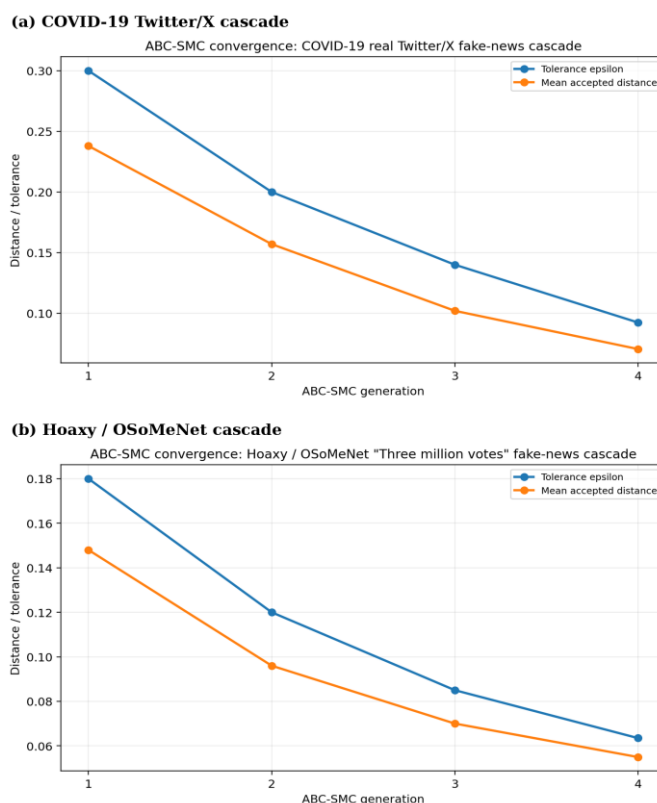
Фигура 3.11 сравнява разпределенията на грешките при напасване между различните модели. ABC-SEIZ дава най-тясното разпределение и най-ниската обща грешка, докато SIR е изместен към по-големите грешки при реконструкцията. Следователно фигурата добавя нови доказателства отвъд единичната траектория, защото обобщава стабилността на намаляването на грешките при последващи предсказващи симулации.

3.6.3 ABC-SMC конвергенция и калибриране на прогнозата

ABC-SMC подобрява калибрирането чрез прогресивно намаляване на толеранса и концентриране на частиците в по-добре напасващи се области. Данните за конвергенция и прогноза показват, че апостериорното предсказване е най-полезно като инструмент за краткосрочна неопределеност, отколкото като дългосрочна детерминистична прогноза.

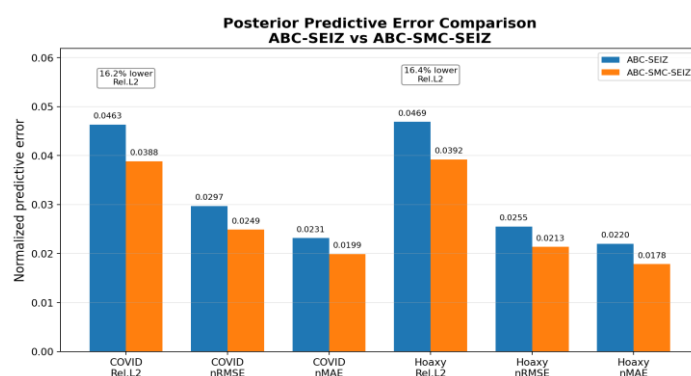
Таблица 3.12. Обобщение на конвергенцията ABC-SMC за модела за разпространение на фалшиви новини SEIZ

Набор от данни	Поколения	Начално ϵ	Крайно ϵ	Крайно разстояние
COVID	4	0.30	0,092	0,070
Измама	4	0.18	0,063	0,055



Фигура 3.12. Профили на конвергенция ABC-SMC за каскадите COVID-19 и Hoaxy / OSoMeNet.

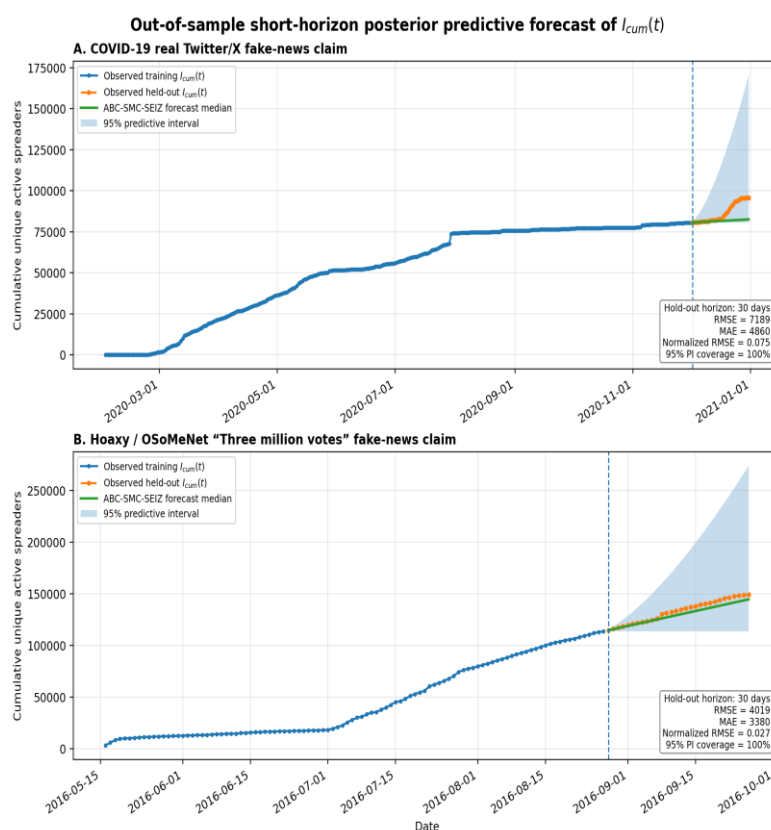
Фигура 3.12.(a)(b) показва, че ABC-SMC прогресивно намалява приетите разстояния през поколенията, което показва стабилна постопериорна конвергенция. Профилът Ноаху е по-стегнат до последното поколение, докато профилът COVID-19 остава по-широк, тъй като повтарящите се изблици на активност правят траекторията по-малко редовна. Това подкрепя използването на ABC-SMC като калибровъчно уточнение преди прогнозните и интервенционните анализи.



Фигура 3.13. Сравнителна апостериорна предсказваща грешка на ABC-SEIZ и ABC-SMC-SEIZ в двете основни емпирични каскади от фалшиви новини

Фигура 3.13 показва, че последователното калибриране намалява нормализираните грешки при прогнозиране и за двете каскади. За случая COVID-19, ABC-SMC-SEIZ намалява относителната L2 грешка с около 16,2% в сравнение с ABC-SEIZ. За случая Ноаху съответното намаление е около 16,4%. Подробни стойности на RMSE и MAE са

предоставени в Приложение D, докато основният текст запазва само сравнението, свързано с резултатите.



Фигура 3.14. Краткосрочна прогноза за каскадите COVID-19 и Hoaxy / OSoMeNet, базирана на ABC-SMC-SEIZ и извън извадката

Фигура 3.14 оценява краткосрочното прогнозиране извън извадката. Калибрираният модел ABC-SMC-SEIZ следва наблюдаваната тенденция на насищане, въпреки че прогнозата за COVID-19 подценява късния растеж. Резултатът показва, че калибрираната рамка поддържа ограничено краткосрочно прогнозиране при несигурност.

3.7 Анализ на сценарии за интервенция

Сценариите за интервенция сравняват развенчаването на несъответствия, намаляването на видимостта и ранното комбинирано смекчаване при калибрирана динамика на SEIZ. Резултатите подкрепят интерпретация, чувствителна към времето: ранната комбинирана интервенция намалява бъдещия кумулативен спред по-ефективно от отложените корекционни действия. Тези сценарии са базирани на модели контрафактуални сценарии, а не причинно-следствени оценки на платформено-политически аспекти.

Таблица 3.1 3. Кратко обобщение на сценария за контрафактуална интервенция при изходно ниво на постериорна медиана на ABC-SEIZ.

Набор от данни	Най-силният сценарий за развенчаване	Сценарий с най-силна видимост / смекчаване	Допълнителна печалба	Основна интерпретация
Твърдение за фалшиви новини за COVID-19 в Twitter/X	Ранно развенчаване - силно: крайно $I_{cum}(t) = 64\,695$; пик $I_{new}(t) = 343$; 23,9% намаляване на крайния размер.	Ранно комбинирано смекчаване: крайно $I_{cum}(t) = 55\,862$; пик $I_{new}(t) = 327$; 34,2% намаление на крайния размер.	8 833 по-малко крайни активни разпространители в сравнение с най-силния сценарий за развенчаване; +10,3 процентни пункта в намаляването на крайния размер.	Многократната каскада от COVID-19 е най-полезна, когато коригиращите ефекти се комбинират с ранно намаляване на видимостта.
Измама / OSoMeNet Твърдение за „три милиона гласа“ - фалшива новина	Ранно развенчаване - силно: крайно $I_{cum}(t) = 135\,071$; пик $I_{new}(t) = 1\,926$; 6,1% намаление на крайния размер.	Ранно комбинирано смекчаване: крайно $I_{cum}(t) = 119\,984$; пик $I_{new}(t) = 1\,584$; 16,6% намаление на крайния размер.	15 087 по-малко крайни активни разпространители в сравнение с най-силния сценарий за развенчаване; +10,5 процентни пункта в намаляването на крайния размер.	Каскадата от политически фалшиви новини е по-слабо чувствителна към промени, свързани само с корекции, и по-скоро към ранен контрол на видимостта и релейното излагане.

3.8 Интегрирана дискусия

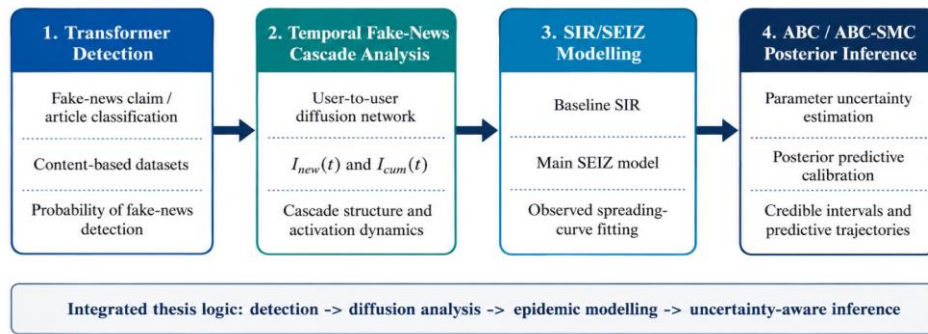
Интегрираната дискусия свързва емпиричните слоеве. Трансформаторните модели предоставят надеждни оценки за риск от съдържание, темпоралните каскади показват дали това съдържание става социално активно, SEIZ подобрява интерпретацията на динамиката на разпространение, а ABC-SMC предоставя апостериорна неопределеност. Базовият функционал за дифузия-риск е полезен, защото прави връзката между вероятността за съдържание, темпоралното усилване и очаквания апостериорен растеж изрична.

Резултатите от темпоралните мрежи предоставят този втори слой доказателства. Чрез преобразуване на избрани твърдения за фалшиви новини в $I_{new}(t)$, $I_{cum}(t)$, насочени потребителски ребра, хъбове, релейни потребители и модели на темпорална достъпност, главата показва как рискът от дифузия става наблюдаем. Твърдението за

реални фалшиви новини от Twitter/X за COVID-19 и твърдението за фалшивите новини от Ноаху / OSoMeNet „Три милиона гласа“ са най-силните емпирични случаи, защото осигуряват достатъчна темпорална структура за епидемично моделиране и последващи изводи. Допълнителните случаи, импортирани от Ноаху и синтетични албански, остават полезни за сравнителна и валидационно ориентирана интерпретация, но те не се третират като еквивалентни изводни доказателства за заключенията на интервенцията ABC/ABC-SMC.

ABC и ABC-SMC подсилват анализа, като отдалечават дисертацията от оценките на единични детерминистични параметри. Последните резултати показват, че параметрите на разпространение на фалшиви новини трябва да се интерпретират с несигурност, особено когато каскадите са шумни, непълни или многократно изригващи. ABC предоставя безвероятно направление до правдоподобни области на параметрите, докато ABC-SMC усъвършенства този процес чрез последователно намаляване на толеранса и подобрява постериорното предсказващо калибриране. Прогнозата за кратък хоризонт $I_{cum}(t)$ допълнително показва, че калибрираната SEIZ структура има предсказваща стойност отвъд ретроспективното напасване на кривата, въпреки че прогнозата остава специфична за хоризонта и не трябва да се тълкува като безусловно предсказване на платформата.

Proposed Framework Flowchart



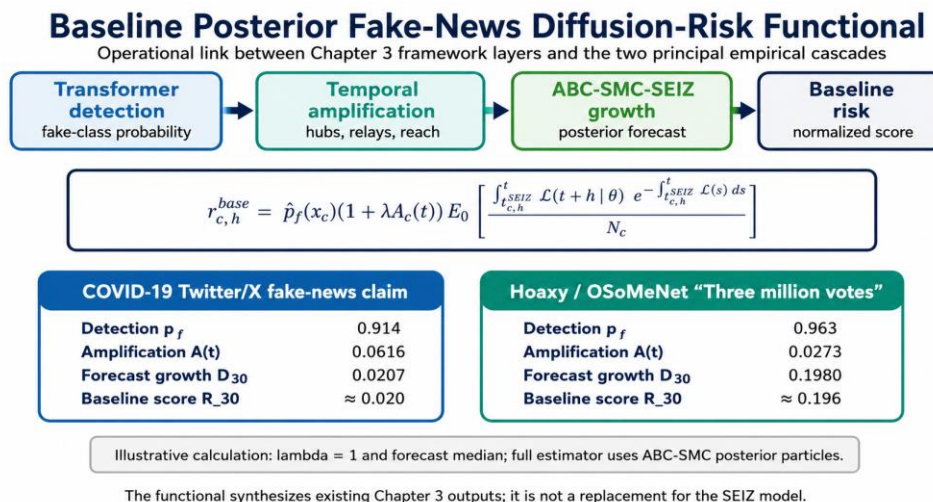
Фигура 3.15. Блок-схема на хибридната рамка на конвейера по Глава 3 от откриване на фалшиви новини, базирано на Transformer, до времеви каскаден анализ, дифузионно моделиране SIR/SEIZ и последващ извод ABC/ABC-SMC. Интегрираната рамка свързва откриването, времевия анализ, епидемичното моделиране и последващия извод в единен конвейер.

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

Уравнение (3.3) прилага базовия функционал на дифузионния риск към случаите от Глава 3, като комбинира вероятност за съдържание, времево усилване и очакван посттериорен растеж.

Фигура 3.16 показва както концептуалната конструкция на оценката, така и двата емпирични примера. Случаите COVID-19 и Ноаху / OSoMeNet не трябва да се

сравняват като универсални класации на риска за платформата. Те са специфични за хоризонта демонстрации, показващи как рамката на дисертацията може да преобразува откриването, каскадното усилване и последващия растеж в нормализирано базово количество риск за избрано твърдение и времеви прозорец.



Фигура 3.16. Оперативна демонстрация на функционално свързване на базовата функция за дифузия на фалшиви новини и риск от последваща дифузия, трансформаторно откриване, усилване на темпоралната мрежа и прогнозиран растеж по ABC-SMC-SEIZ за двете основни емпирични каскади

Блокът Transformer предоставя вероятността за фалшиви новини, блокът temporal-network предоставя усилване, а блокът ABC-SMC-SEIZ предоставя очаквания прогнозен растеж. Умножаването на тези компоненти дава базова оценка за риск от последваща дифузия за избран прогнозен хоризонт. Фигурата пояснява, че оценката не е допълнителен невалидиран модел; тя е компактен мост между емпиричните резултати, вече генерирани от главата.

3.9 Резюме на главата

Глава 3 отговаря на изследователските въпроси емпирично. Трансформаторните модели се представят добре, но зависят от типа на текста и езика; темпоралните каскади са необходими за дифузионния анализ; SEIZ се подобрява спрямо SIR за основните каскади; ABC/ABC-SMC количествено определя неопределеността на параметрите; а сценариите за интервенция показват как калибрираната рамка може да подпомогне предпазливия анализ на риска.

Таблица 3.14. Обобщение на доказателствата по изследователските въпроси за Глава 3

РК	Доказателства в Глава 3	Заклучение
РК1	Раздел 3.3; Таблици 3.2-3.7 и Фигури 3.1-3.5	Производителност на трансформатора в различни набори от данни
РК2	Раздел 3.4; Таблици 3.8-3.9 и Фигури 3.6-3.7	Реконструирани времеви каскади
РК3	Раздел 3.5; Таблица 3.10 и Фигури 3.8-3.9	SEIZ в сравнение със SIR
РQ4	Раздел 3.6; Таблици 3.11-3.13 и Фигури 3.10-3.14	Оценена неопределеност на ABC/ABC-SMC
РQ5	Раздели 3.7-3.8; Таблица 3.14 и Фигури 3.15-3.16	Показана е последваща прогноза и логика на интервенция

Глава 4. Заключение - обобщение на получените резултати

Глава 4 синтезира резултатите от дисертацията. Дисертацията демонстрира, че анализът на фалшивите новини е по-силен, когато текстовото откриване, темпоралната дифузия, епидемичното моделиране и байесовата несигурност са комбинирани в една проследима рамка. Рамката не претендира да измерва частните убеждения или реалните ефекти на модерирането; тя предоставя възпроизводим академичен метод за свързване на риска от съдържание с каскадно поведение и последваща предсказваща несигурност.

Заклучения на дисертацията

Заклученията са организирани съгласно RQ1-RQ5 и синтезират доказателствата на дисертацията, без да повтарят подробните резултати от Глава 3.

RQ1. Трансформаторните модели формират силен двоичен слой за откриване на фалшиви новини. Експериментите показват надеждна производителност на класа фалшиви новини в кратки публикации за COVID-19, дълги статии, дезинформация в здравеопазването и албански данни с ниски ресурси, като най-добрата архитектура зависи от дължината на текста, езика и домейна.

RQ2. Социалният риск не може да се изведе само от вероятността за фалшиви новини. Насочените каскади с времеви отпечатъци, $I_{\text{new}}(t)$, $I_{\text{cum}}(t)$, размерът на каскадата и времето до достигане на пик свързват текстовата достоверност с наблюдаемо разпространение.

RQ3. SIR е полезен като прозрачна базова линия, но SEIZ е по-подходящ за разпространение на фалшиви новини, тъй като експонираните и скептични/неразпространяващи се състояния представляват забавяне, колебание и

корекция. Следователно SEIZ обяснява по-добре нередовните и многократните каскади.

RQ4. ABC и ABC-SMC предоставят необходимите изводи, отчитащи неопределеността, за частично наблюдавани каскади. Те оценяват последващите области за параметрите на предаване, възстановяване, експозиция и скептицизъм, вместо да разчитат на прекалено уверени детерминистични съпоставки.

RQ5. Интегрираната рамка „Трансформатор-времева-мрежа-ABC“ подкрепя предпазливо разсъждение, ориентирано към интервенция. Контрафактуалните резултати показват, че ранното комбинирано смекчаване е по-ефективно от забавените или само коригиращи действия, като същевременно остават контролирани симулации, а не оценки на причинно-следствената политика.

В синтез, базовите апостериорни функционални връзки между дифузията на фалшивите новини и риска от дифузия обединяват вероятността за трансформация, амплификацията на темпоралната мрежа и прогнозния растеж на ABC-SMC-SEIZ в една одитирана величина на риска. Основните ограничения са, че вероятностите на модела не са абсолютни преценки за проверка на фактите, споделянето на наблюдаеми данни не доказва достоверност, скритите състояния на SEIZ са моделни заместители и по-широката валидация изисква по-реални каскади, особено на албански.

Бъдеща работа

Бъдещата работа би трябвало да валидира рамката върху по-големи и по-разнообразни каскади от реалния свят, особено в различни езици, платформи и области на дезинформация. Допълнителни разширения включват по-богати данни от темпорални мрежи, модели за графово обучение, по-силна причинно-следствена оценка на интервенциите, невронна апостериорна оценка и реални албански набори от данни за темпорална дифузия.

Приноси на дисертацията

Тази дисертация има следните научни и методологични приноси при откриването и анализа на разпространението на фалшиви новини.

1. Предложена е и е оценена интегрирана рамка, която свързва базирано на трансформатори откриване на фалшиви новини, времева каскадна реконструкция, SIR/SEIZ епидемично моделиране и ABC/ABC-SMC апостериорна инференция. Приносът се състои в третирането на фалшивите

новини като процес от съдържание към каскада, а не като изолиран проблем с класификацията на текст.

2. Установена е ясна методологична граница между наборите от данни, използвани за откриване на бинарни фалшиви новини, и наборите от данни, използвани за моделиране на темпорална дифузия. Това разграничение предотвратява неправилното тълкуване на резултатите от класификацията само по съдържание като епидемични или каскадни доказателства.
3. В дисертацията са внедрени и оценени BERT, RoBERTa , BigBird , mBERT и XLM-RoBERTa в кратки текстове, дълги текстове, здравеопазване и ниско-ресурсни албански набори от данни за фалшиви новини. Полученият вероятностен резултат за фалшиви новини се използва като сигнал за риск от съдържание, който подкрепя прехода от откриване към времеви каскаден анализ.
4. Твърденията за фалшиви новини бяха дефинирани като насочени, времево маркирани темпорални каскади и извличат ключови дифузионни величини, като например нови активни разпространители, кумулативни активни разпространители, размер на каскадата, време до пик и обхват на мрежата. Този принос осигурява емпиричния мост между откриването на съдържание и моделирането в епидемичен стил.
5. Сравнението между моделите SIR и SEIZ върху основните емпирични каскади от фалшиви новини показва, че SEIZ предоставя по-подходящо представяне, когато са от значение забавено излагане, колебание, скептицизъм или неразпространяващо се поведение. SIR се запазва като прозрачна базова линия, докато SEIZ се използва като основен модел на разпространение.
6. Приложени са ABC и ABC-SMC за оценка на несигурни епидемични параметри от частично наблюдавани каскади на социалните мрежи. Този принос отвежда анализа отвъд детерминистичните точкови оценки, като се докладват апостериорни разпределения на параметрите, достоверни интервали, апостериорни предсказващи проверки и калибрационна диагностика.
7. Дисертацията въвежда базов функционал на риска, който свързва вероятността за фалшиви новини от типа трансформатор (Transformer), амплификацията на темпоралната мрежа и апостериорния прогнозен растеж на ABC-SMC-SEIZ. Функционалът не замества отделните модели; вместо това той предоставя проверим математически синтез на слоевете за откриване, дифузия и апостериорен инференциален извод.
8. За да се оценят сценариите за контрафактуална интервенция, в дисертацията се използва калибрираната дифузионна рамка, включително развенчаване,

намаляване на видимостта и ранно комбинирано смекчаване. Приносът не е причинно-следствено твърдение за реалната политика на платформата, а възпроизводимо, базирано на модел, класиране на механизмите за интервенция при изрични допускания.

Списък на публикациите, свързани с дисертацията

1. Bozhiki, 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. Bozhiki, 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. Bozhiki 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>
4. Bozhiki, 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, IEEE. pp. 1–7. <https://doi.org/10.23919/SoftCOM66362.2025.11197375>.
5. Bozhiki, 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 [89]

Abstracts of Dissertations

Number 10, 2026

INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGIES
BULGARIAN ACADEMY OF SCIENCES

БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ

ИНСТИТУТ ПО ИНФОРМАЦИОННИ И КОМУНИКАЦИОННИ ТЕХНОЛОГИИ

Брой 10, 2026

Автореферати на дисертации