



БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ



ИНСТИТУТ ПО ИНФОРМАЦИОННИ И КОМУНИКАЦИОННИ ТЕХНОЛОГИИ

Секция: „Информационни процеси и системи за подпомагане на вземането на решения“

Кристел Вула

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

АВТОРЕФЕРАТ

на дисертация

за придобиване на образователна и научна степен

„ДОКТОР“

професионално направление 4.6. „Информатика и компютърни науки“,

в докторска програма: Информатика

София, 2026 г.

Дисертацията е обсъдена и допусната до защита на разширено заседание на катедра „Информационни процеси и системи за подпомагане на вземането на решения“ към ИИКТ–БАН, проведено на 17.06.2026 г.

Дисертацията е структурирана в въведение, четири глави, общи заключения за постигнатите резултати, приноси към дисертацията, списък с публикации, свързани с дисертацията, декларация за оригиналност, библиография и приложения. Дисертацията има общо 210 страници, 47 фигури, 111 таблици и 191 библиографски източника.

Защитата на дисертацията ще се проведе на 2026 г. от..... часа в зала/стая на бл. 2 на ИИКТ–БАН на открито заседание на научно жури в състав:

Научно жури:

1.
2.
3.
4.
5.

Рецензиите и становищата на членовете на научното жури, заедно резюмето и дисертацията са публикувани на уебсайта на ИИКТ–БАН. Материалите за защитата са достъпни за всички заинтересовани в стая на ИИКТ-БАН, бл. 2, ул. „Акад. Г. Бончев“, София.

Автор: **Кристел Вула**

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

Структура на дисертацията

Вула, К. Базирано на трансформатори откриване и анализ на епидемичния модел на динамиката на фалшиви новини в социалните мрежи.

Научен ръководител: проф. д-р Васил Георгиев Гуляшки.

Секция: Информационни процеси и системи за подпомагане на вземането на решения.

Докторска програма: Информатика. София, 2026.

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.

Ключови думи

Откриване на фалшиви новини; Трансформаторни модели; Темпорални мрежи; SIR/SEIZ епидемично моделиране; Приблизително байесово изчисление; Интервенционен анализ.

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 [1]-[6].

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 [7], [8].

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], [5], [6].

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 полученият резултат се интерпретира като вероятностен сигнал за достоверност, а не като абсолютна преценка за проверка на фактите [1].

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 [1].

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

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.
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 [1].
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.
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 [6].
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.
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.

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 [1].

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 стилOMETрични 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) [1].

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 бустинг. Тяхната ценност е в интерпретируемостта и възпроизводимостта, докато ограничението им е слабото контекстуално разбиране и чувствителността към промяна на домейна [1].

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. [1];	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. [1];	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	Шу и др. [1];	Ефективен във високоразмерни разредени пространства.	Изисква внимателен избор на ядро и параметри за големи набори от данни.
Случайна гора	Стилометрични, лексикални, четимостни и метаданни характеристики	Шу и др. [1];	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	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 [2].

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 [2].

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.

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 [3].

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.

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 [2].
RoBERTa	Оптимизирано BERT обучение с повече данни, динамично маскиране и без предвиждане на следващото изречение.	Често подобрява устойчивостта и класификационните характеристики.
mBERT	Многоезичен BERT, обучен на 104 езика.	Полезно за откриване на фалшиви новини на различни езици или на много език.
BigBird	Трансформатор с разредено 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 [3].
SBERT	Вграждания на ниво изречения, базирани на сиамски/триплетни BERT мрежи.	Полезно за семантично сходство, сравнение на твърдения и класификация на ниво изречения.
XLM-RoBERTa	Междуетичната RoBERTa е обучена върху големи многоезични данни от CommonCrawl.	Полезно за многоезично откриване и трансфер между езици.
Големи езикови модели	Модели 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

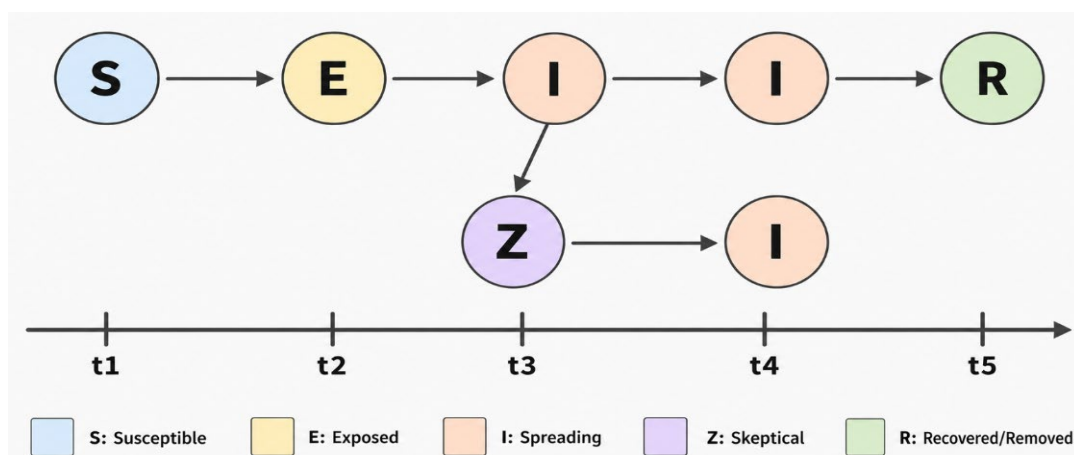
Трансформаторните детектори могат да бъдат изчислително скъпи, по-малко прозрачни от класическите базови линии и уязвими към състезателен парафраз или изместване на домейна. Следователно техните вероятности трябва да бъдат валидирани и интерпретирани като оценки за риск от съдържание, особено когато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

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 [5], [6].



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 епидемия 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 [9], [10].

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 [9].

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

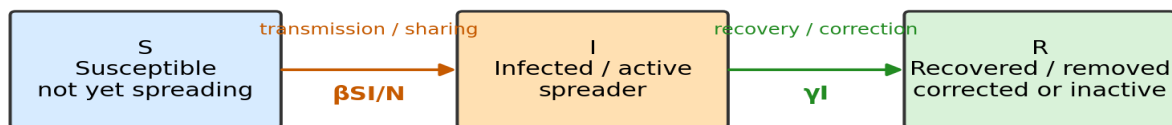
Модел	Основни състояния	Представителна литература	Тълкуване на фалшиви новини	Основна сила и ограничение
SI	$S \rightarrow I$	Кермак и Маккендрик [9]	Потребителите преминават от потенциално излагане към активно разпространение.	Полезно за ранна неконтролирана дифузия, но няма състояние на корекция, забравяне или възстановяване.
SIS	$S \rightarrow I \rightarrow S$	Кермак и Маккендрик [9]	Потребителите могат да спрат да разпространяват вируса и по-късно да станат отново податливи след многократно излагане.	Полезно за повтарящо се споделяне, но не представлява трайна корекция.

SIR	$S \rightarrow I \rightarrow R$	Кермак и Маккендрик [9]; Джин и др. [10];	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	$S \rightarrow E \rightarrow I \rightarrow R$	Кермак и Маккендрик [9]	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 [10].

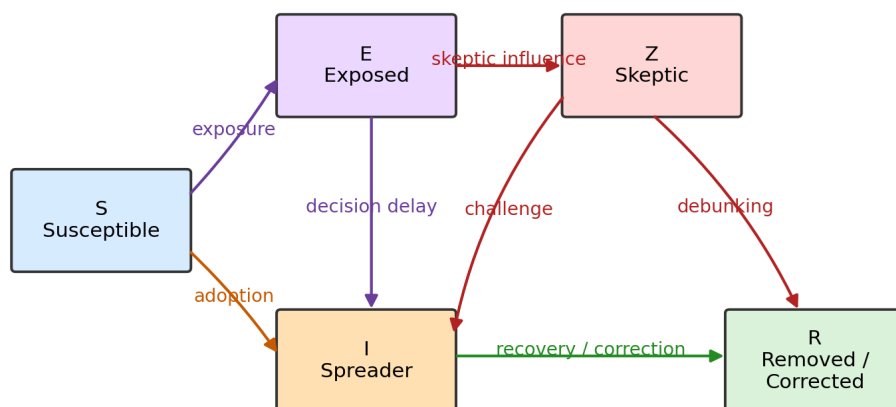
SIR state-transition graph for fake-news diffusion



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.

SEIZ state-transition graph for fake-news diffusion



Interpretation: the Exposed and Skeptic states represent delayed adoption, non-spreading, questioning, and debunking behaviours.

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; дискуссионни 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.

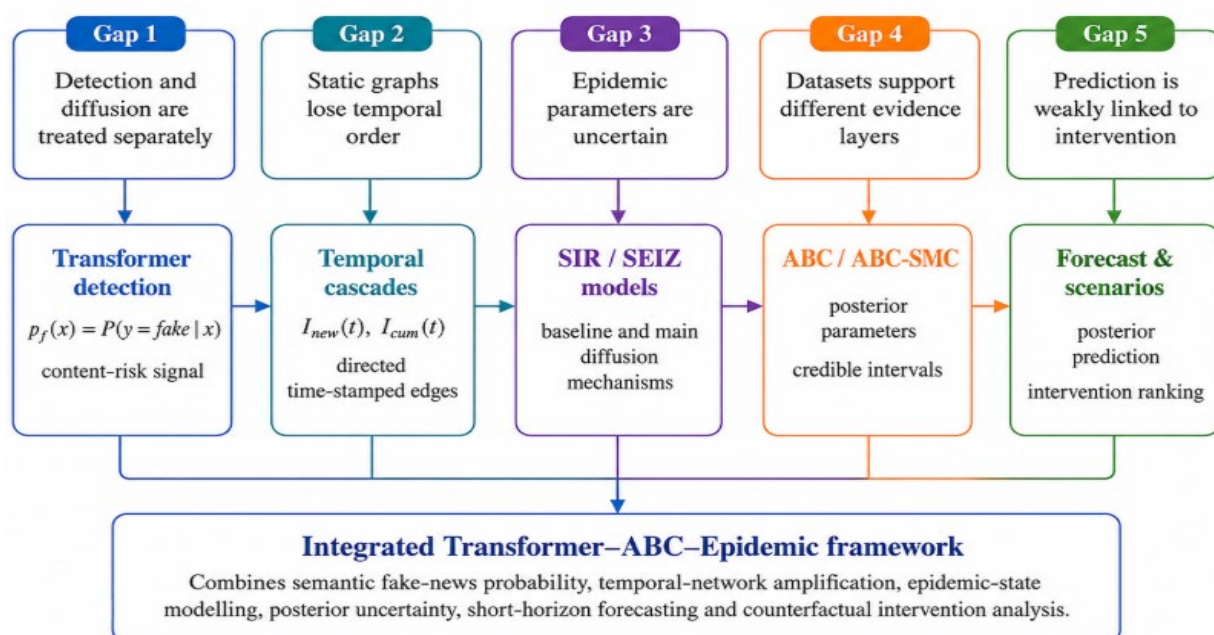
1.8 Приблизително байесово изчисление за дифузионен извод

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 [4].

1.9 Пропуски в изследванията и открити проблеми

Липсите в литературата са свързани: откриването често е отделено от разпространението, статичните графики скриват времеви 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 [7], [8].

Research-gap map and proposed Transformer-ABC-Epidemic integration



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 третираат 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 скриват пътища на дифузия, съобразяващи се с времето.	Представете твърдени-ята като насочени темпорални каскади с подредени взаимодействия и криви на раз-пространение, 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 инференция. Изследователските задачи, произтичащи от тази цел, обхващат преглед на литературата, разделяне на набори от данни, бинарно откриване, каскадно

конструирани, 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 методология.	Ограничено контекстуално разбиране и слаб трансфер при промяна на домейна.	Предоставя исторически контекст и обясними точки за 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	Представите 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 отпечатьци, насочени ребра $I_{new}(t)$ и $I_{cum}(t)$ 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.

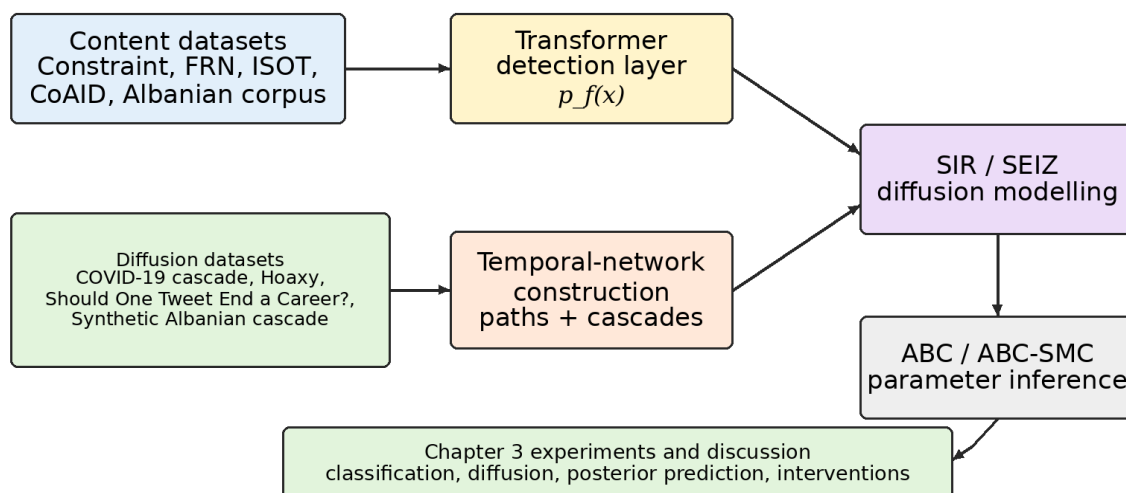
1. Почистете текстовите полета и хармонизирайте етикетите в проверяемо двоично представяне фалшив/реален код.
2. Токенизирайте съдържанието според избраното семейство Transformer и оценете слоя за двоично откриване.
3. Извлекете или запишете вероятността за фалшиви новини, оценена от модела, $p_f(x)$ 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 [7], [8].

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. Запазените 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 Noaxu / OSoMeNet, импортираното от Noaxu 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 [7], [8].



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 потребителите.	Може да се използва като метаданни, когато са налични етикети и текст.	Необходимо за темпорални мрежи, преходи между състояния и калибриране на SIR/SEIZ.
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 компактия набор от данни.

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	Мащабен 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-ориентиранияте 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 съдържание	Съпоставете хетерогенните етикети в общо фалшиво/реално представяне за експерименти 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 [7], [8].

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 ë и ç.	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.

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

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
Ноаху / 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

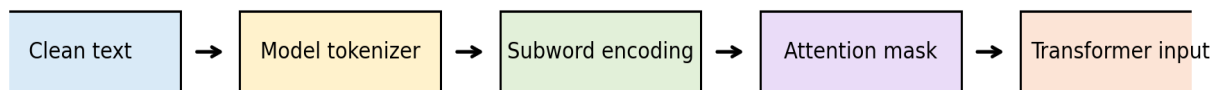
2.4.3 Токенизация и подготовка на входа на трансформатора

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.

Transformer input preparation workflow



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.

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

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 дисбаланс.

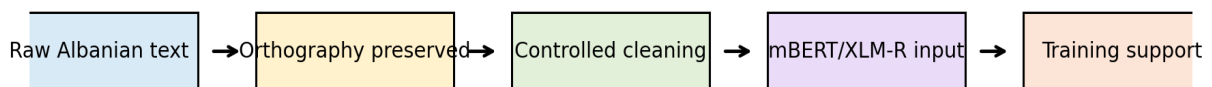
$$\mathcal{L}_{\text{WBCE}} = -[w_1 y \log(p) + w_0 (1 - y) \log(1 - p)] \quad (2.5)$$

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

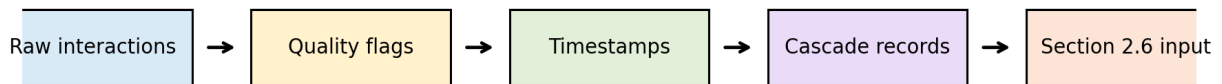
Таблица 2.9. Предизвикателства и отговори на предварителната обработка на Compact Albanian

Предизвикателство	Препоръчителен отговор	Причина
Специфични за езика знаци	Запазете $\ё$ и $\с$; 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 ненаатрапчива 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 Конструирание 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 [5], [6].

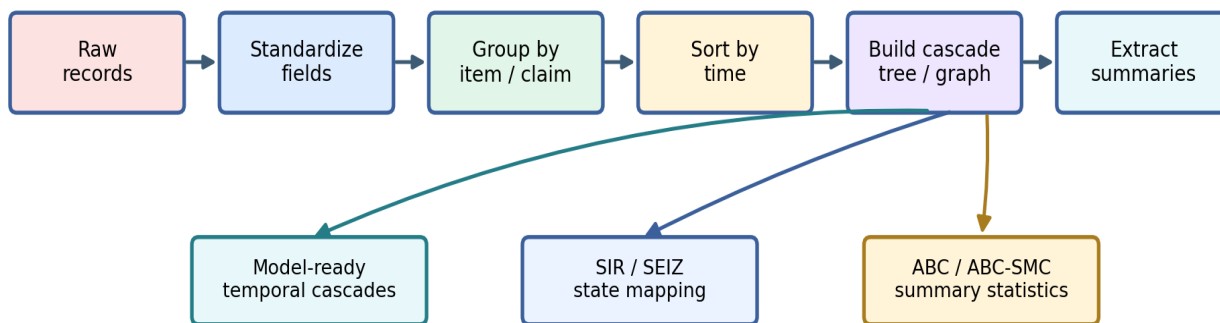
Вместо невронни архитектури с темпорални графи е избрано представяне на темпорална мрежа, тъй като дисертацията изисква интерпретируема каскадна реконструкция и извод за епидемични параметри, а не научени динамични вграждания на възли или модели за предсказване на връзки. Наличните дифузионно-ориентирани 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 третират 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.

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

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 зависимост от набора от данни.	Потребител на Twitter/X в Ноаху; туит от източника в каскада от ретуитове.	Определя единиците, които могат да станат податливи, изложени 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, туит на източника или етикет на външен елемент.	Групира събития за каскадна статистика и ABC обобщаващи функции.

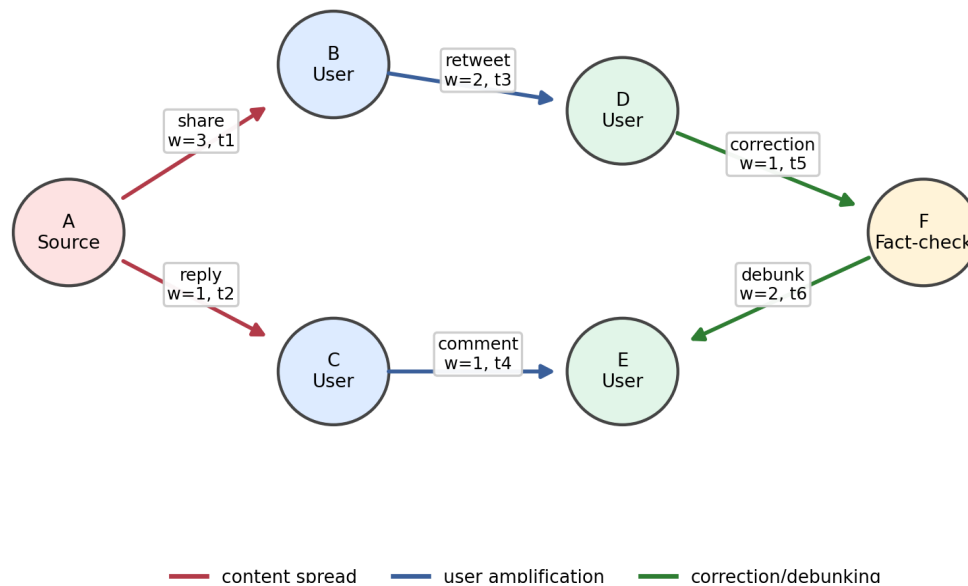
Таблица 2.12. Унифицирана схема, готова за моделиране, за дифузионно-ориентирани записи

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, като например ретуит, отговор, повторна публикация, твърдение или запис за проверка на фактите.	Поддържа интерпретация на показатели 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 по-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.

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

където $s_{out}(i)$ 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.

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

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 [6].

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 $I_{new}(t)$ и $I_{cum}(t)$. $I_{new}(t)$ е броят на новите уникални активни разпространители, наблюдавани в даден времеви прозорец, докато $I_{cum}(t)$ 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.

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.
$I_{new}(t)$	Брой новоактивни уникални разпространители във времеви прозорец t .	Използва се като наблюдавано увеличение на времето на заразяване за динамиката на дифузия.
$I_{cum}(t)$	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 $I_{new}(t)$.	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.

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

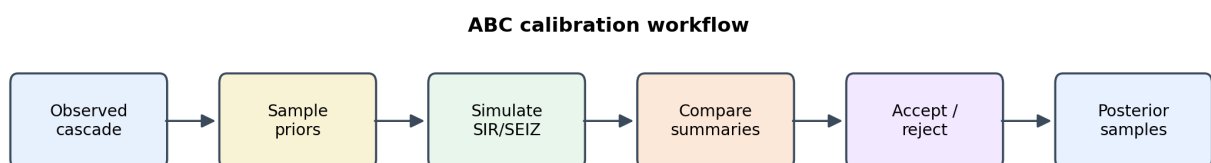
В уравнение (2.11), G_c^T означава насочената времева каскадна мрежа, свързана с твърдение за фалшиви новини с. Нейните възли са потребителите или обектите-източници в каскадата, докато нейните ръбове са насочени взаимодействия с времеви отпечатъци, по избор претеглени. Горният индекс Т показва времево подреждане. Този обект предоставя основа за характеристики на каскадно ниво и SIR/SEIZ входове [5], [6].

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 [5].

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:

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

Трудността в уравнение (2.14) е членът на вероятността $p(D_{\text{obs}} \mid \theta)$. 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 [4].

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

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 [4].

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

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.

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

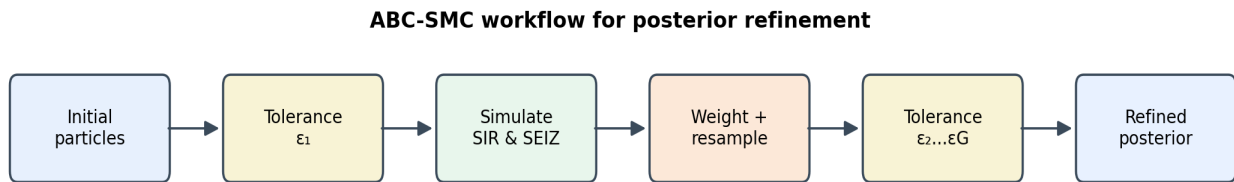
В уравнение (2.18), $p_f(x_i)$ означава вероятността, зададена от детектора на трансформатора, че текстовият елемент x_i принадлежи към класа на фалшивите новини. Резултатът не се използва като компонент на ансамбъла и не замества моделирането на времевата дифузия. Вместо това, той предоставя калибриран 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.

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

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 прогнозираният 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:

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

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 документирано разделяния. 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.	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 [3].	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.	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.	Използва се за откриване и синтетична валидация, а не като емпирично доказателство за Twitter/X или Nooxy.

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 Подходящост 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 $site_type = covid_fake_claim$.	Основна каскада от емпирични дезинформации в здравеопазването за темпорална мрежа, SEIZ/ABC-SMC и $I_{cum}(t)$ прогнозиране.
Политически измами / OSoMeNet	„Три милиона гласа на президентските избори бяха подадени от нелегални имигранти.“	Масщабна 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 [4], [7], [8].

2.7.3 Етични съображения, съображения за поверителност и възпроизводимост

Рамката използва отговорно наблюдаеми публични или документиранни 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 рискови 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.
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.
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.
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 Noаху / OSoMeNet.

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.

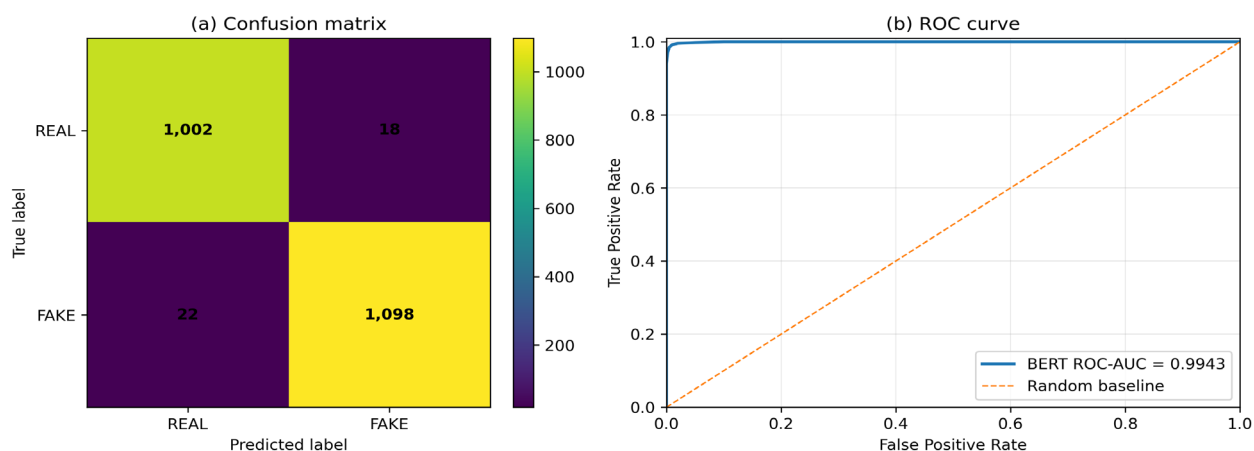
3.3.1 Резултати върху набори от данни, базирани на съдържание

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.

BERT final-run evaluation for Constraint@AAAI2021 Task-A



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 диагностички 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

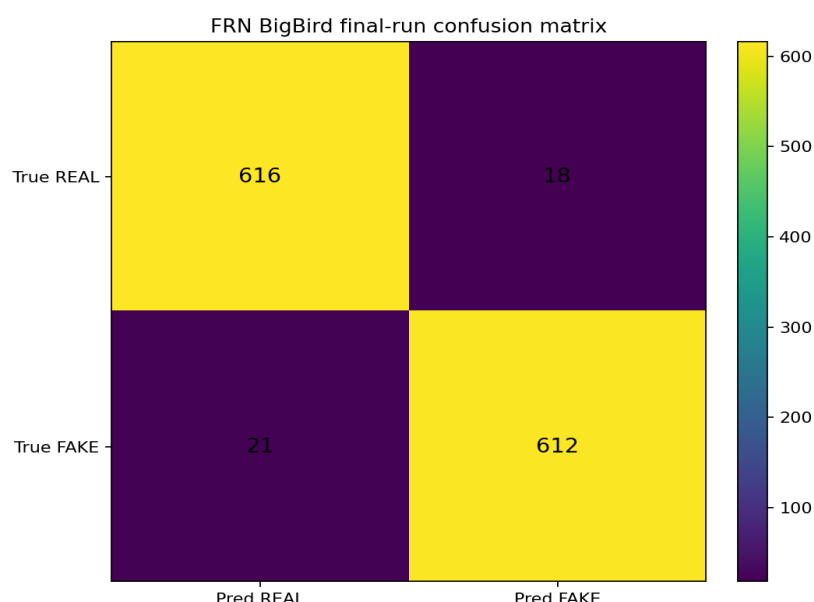
За дълги и специфични за дадена област набори от данни, производителността на Transformer зависи от дължината на текста и темата. BigBird-RoBERTa се представи най-добре при по-дългите новини на FRN и CoAID, RoBERTa се представи най-добре при ISOT и кратките твърдения за здравеопазване, а класическите 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

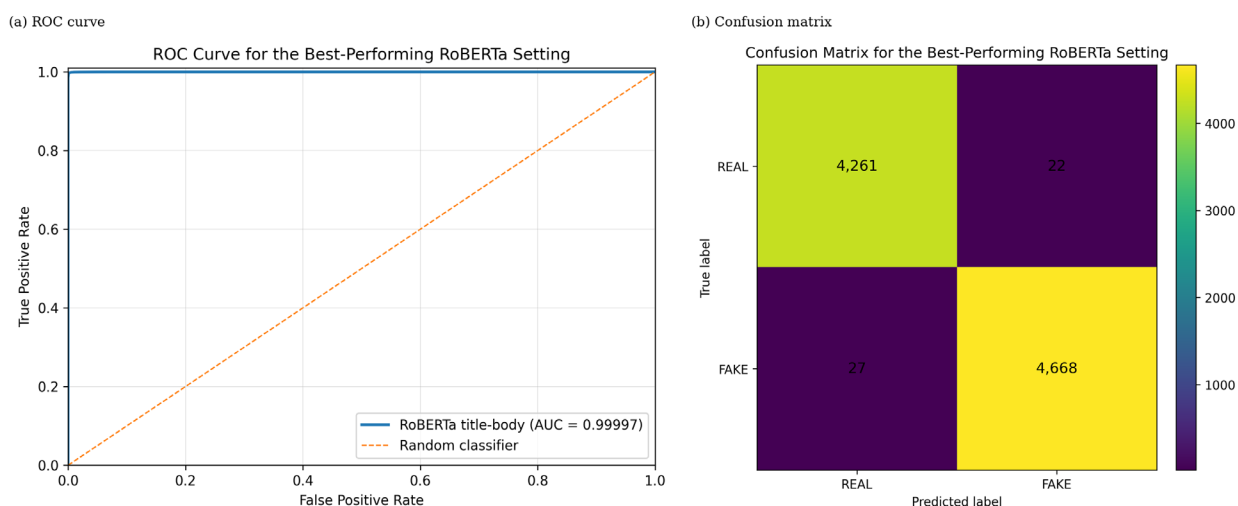
Таблица 3.4. Сравнителна производителност на откриване на трансформатори при FRN, ISOT и CoAID

Набор от данни / подмножество	Тип текст	Най-добра настройка на трансформатора	Фалшива Формула 1	Макро-F1	ROC-AUC
ФРН	Дълги политически/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
Публикации/искания за социални мрежи, свързани с CoAID	Кратки твърдения/публикации и 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.

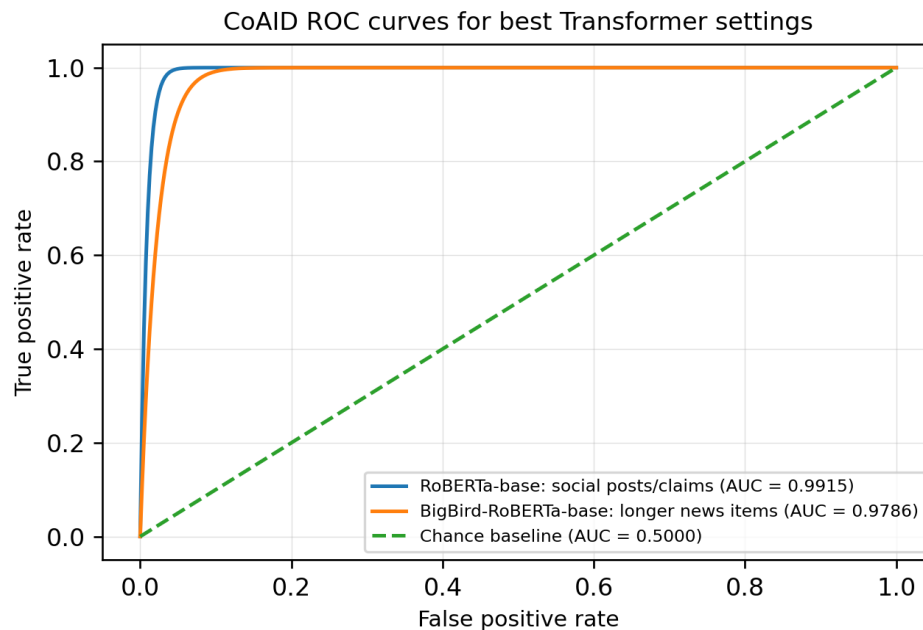
Фигура 3.2 потвърждава силно разделяне на класовете за модела FRN BigBird. Само 18 реални статии са неправилно маркирани като фалшиви, а 21 фалшиви статии са пропуснати. Този модел на грешки показва силно разделяне на класовете и балансиран компромис между отзоваването на фалшиви класове и фалшиво положителния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

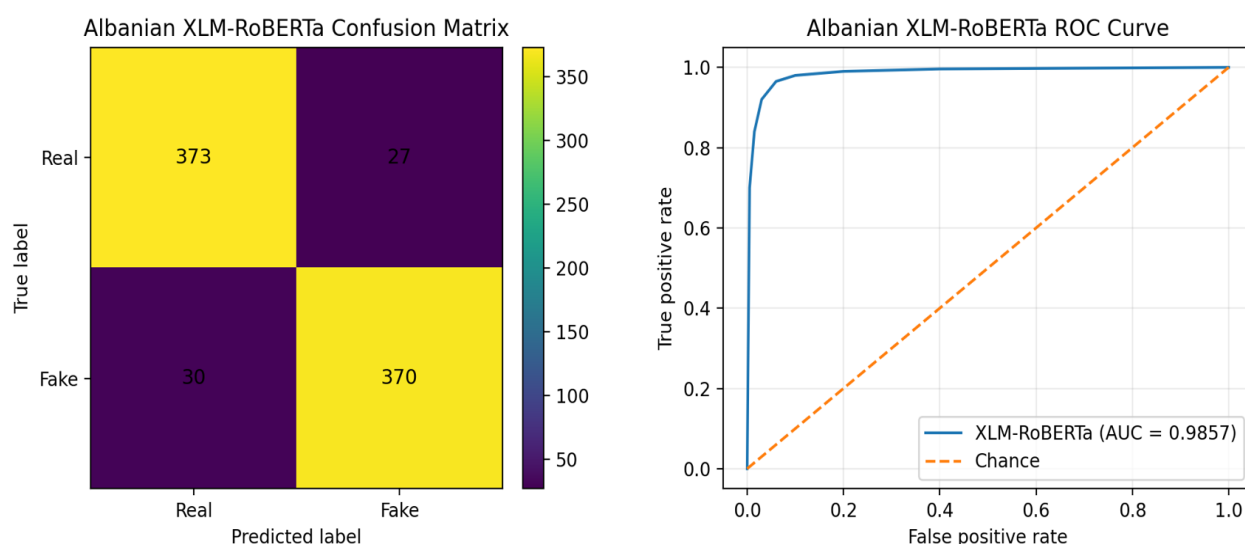
Фигура 3.4 обобщава най-силните резултати от класацията на CoAID Transformer за кратки публикации/твърдения в здравеопазването и по-дълги новинарски материали.

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 3.5. 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 3.6. 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/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
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 3.7. 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 избра две основни емпирични каскади за моделиране: твърдението за фалшивите новини в Twitter/X за COVID-19 и твърдението на Hoaxy / OSoMeNet за „три милиона гласа“. Случаите, внесени от Hoaxy, и синтетичните албански случаи са запазени като помощни примери за сравнение или валидиране, а не като равностойни доказателства за последващ извод.

Таблица 3.8. 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 твит?“ 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

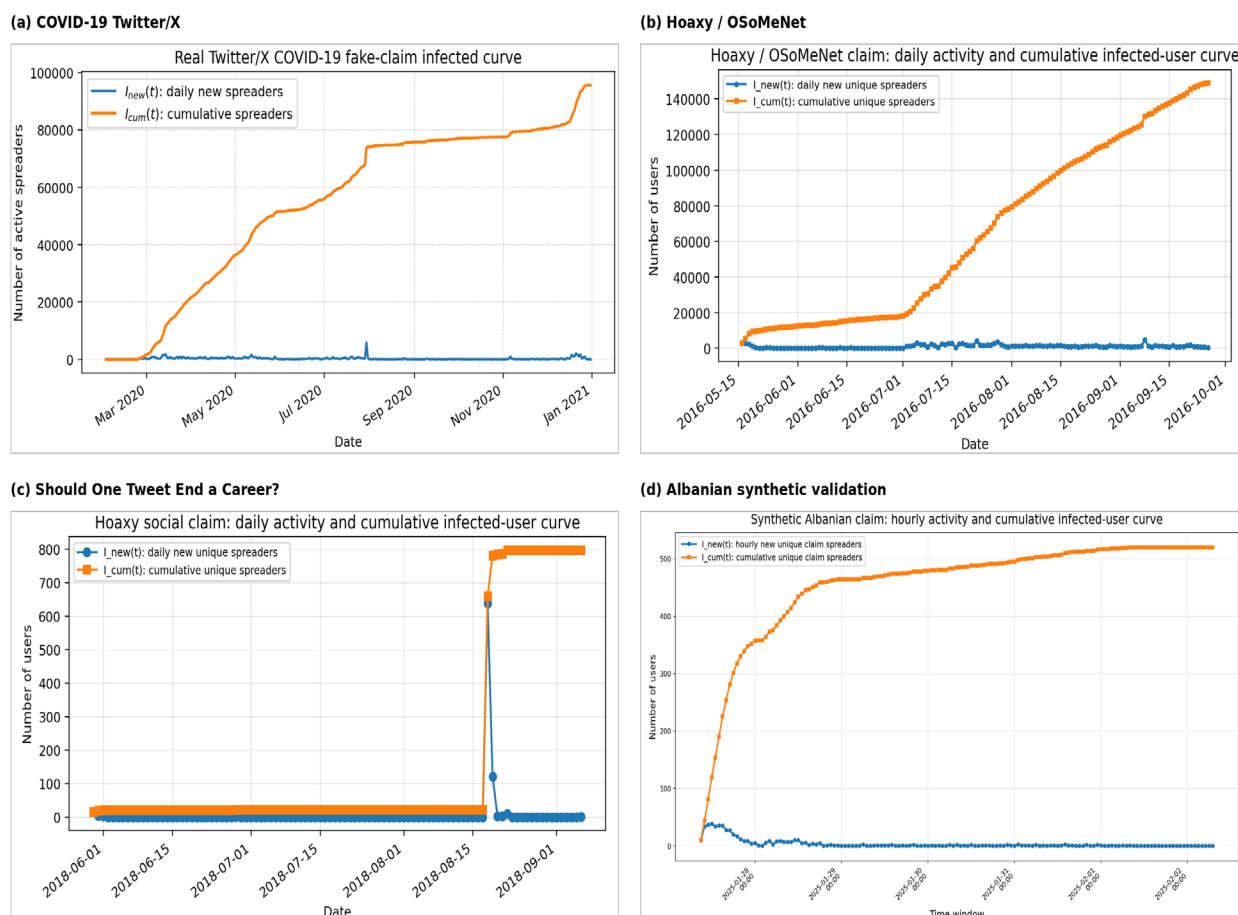
3.4.2 Наблюдавани траектории на заразени потребители

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 3.9. Каскадни обобщени статистики, използвани за прехода SIR/SEIZ.

Случай	Връх 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) Ноаху / OSoMeNet, (в) „Трябва ли един туит 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.

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 подобрение, достатъчно голямо, за да оправдае използването ѝ в последващия етап на калибриране на ABC и ABC-SMC.

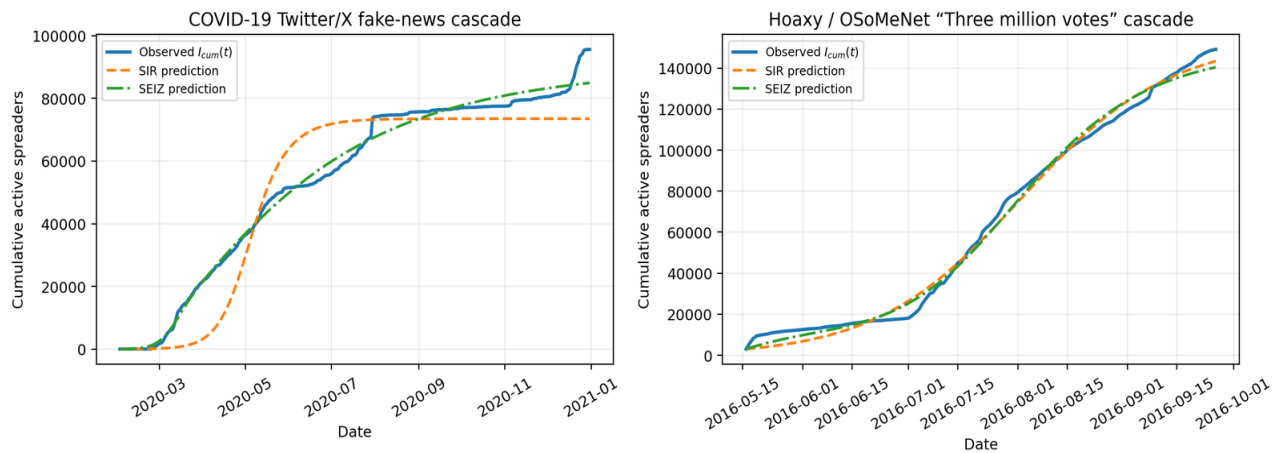
3.5.1 Резултати от моделното напасване

Таблица 3.10. 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	SIR	0.1626	9 958,4	7 793,3	95 599 / 73 510
Твърдение за фалшиви новини за COVID-19 в Twitter/X	SEIZ	0,0463	2 838,9	2 092,6	95 599 / 84 921
Измама / OSoMeNet Твърдение за „три милиона гласа“ – фалшива новина	SIR	0,0511	4 150,2	3 560,2	149 079 / 143 396
Измама / OSoMeNet Твърдение за „три милиона гласа“ – фалшива новина	SEIZ	0,0469	3 809,7	3 200,6	149 079 / 140 408

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.

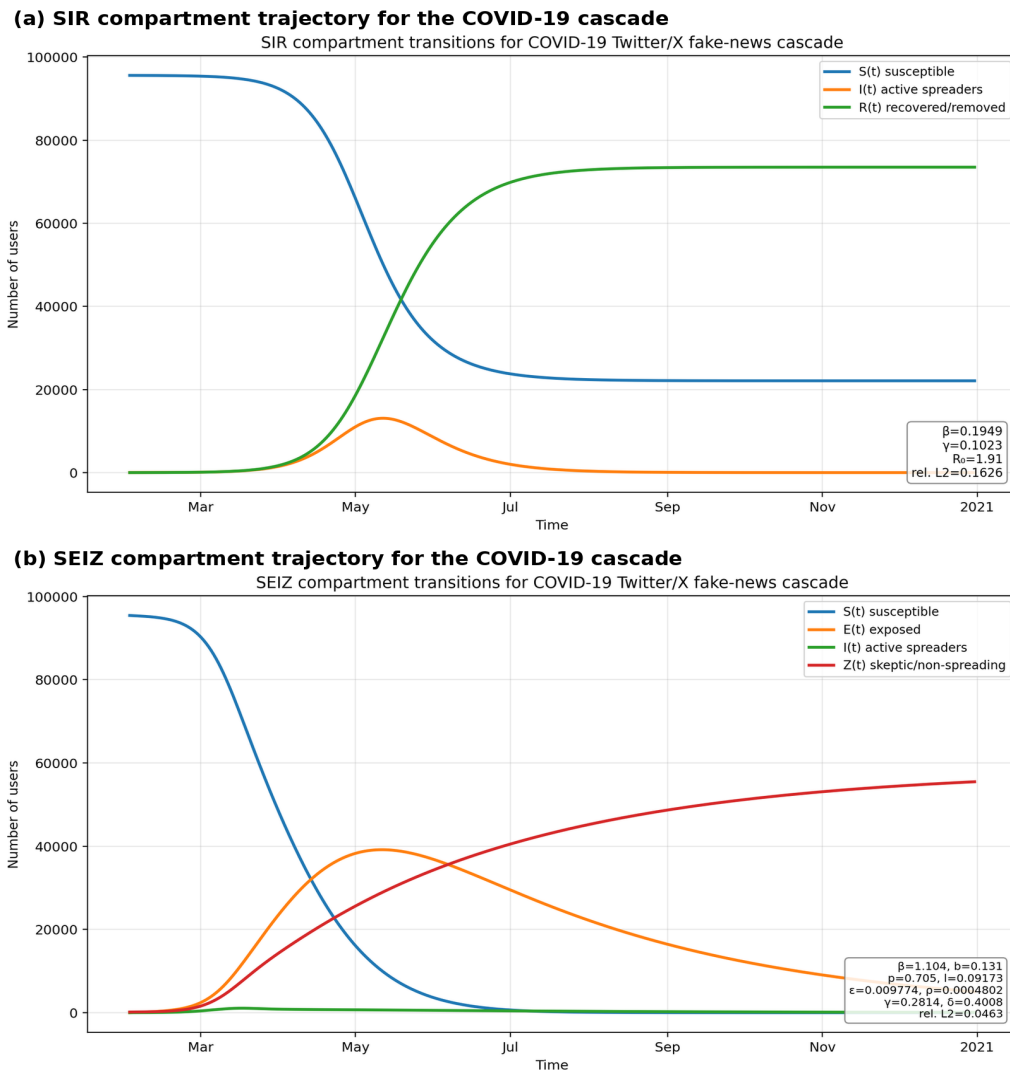
Фигура 3.8 сравнява наблюдаваните кумулативни траектории на разпространение с прогнозите на SIR и SEIZ за двете емпирични каскади. В случая на COVID-19, кривата на SIR достига насищане твърде рано, докато кривата на SEIZ следва наблюдаваната траектория по-точно в основния интервал на растеж. В случая 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 $I_{cum}(t)$ 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 3.11. Наблюдавани каскадни обобщени статистики, използвани за ABC извод.

Набор от данни	Краен размер	Пиков ден	Върх 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.

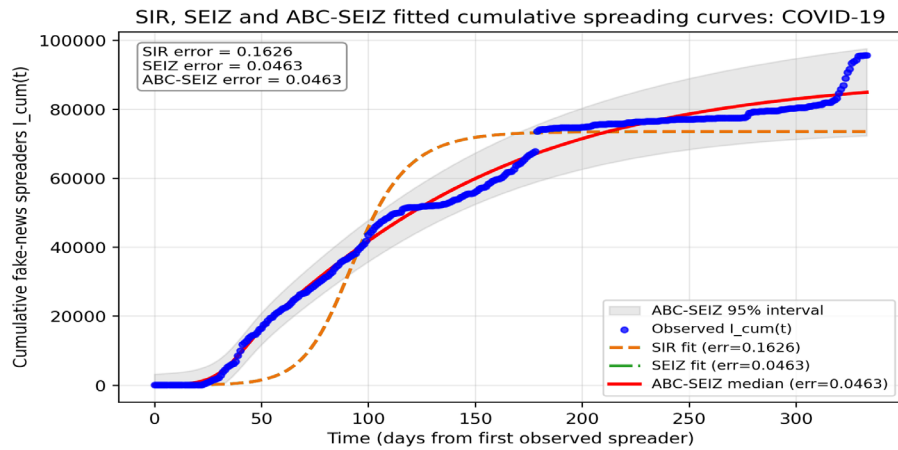
Таблица 3.12. Обобщение на последващите оценки за ABC-SIR и ABC-SEIZ

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	$\beta \approx 0$; $b=2.000$; $p=.0118$; $l=.0359$; $\epsilon=.0428$; $\gamma=.0086$

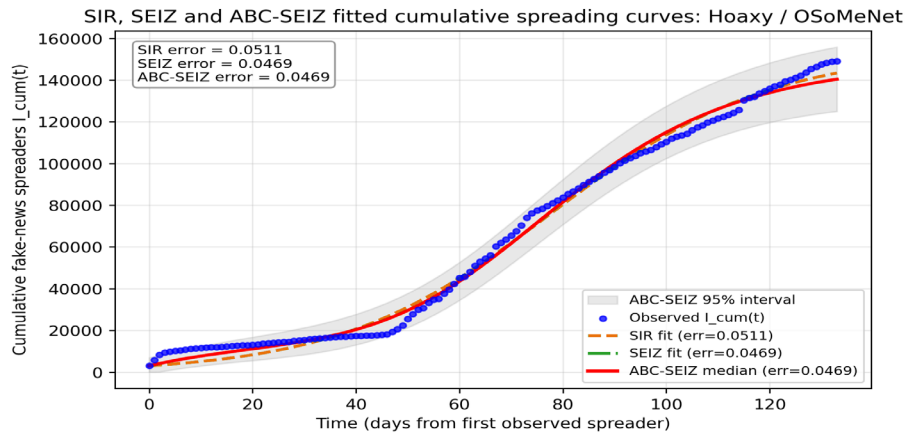
3.6.2 Постериорна предиктивна валидация

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.

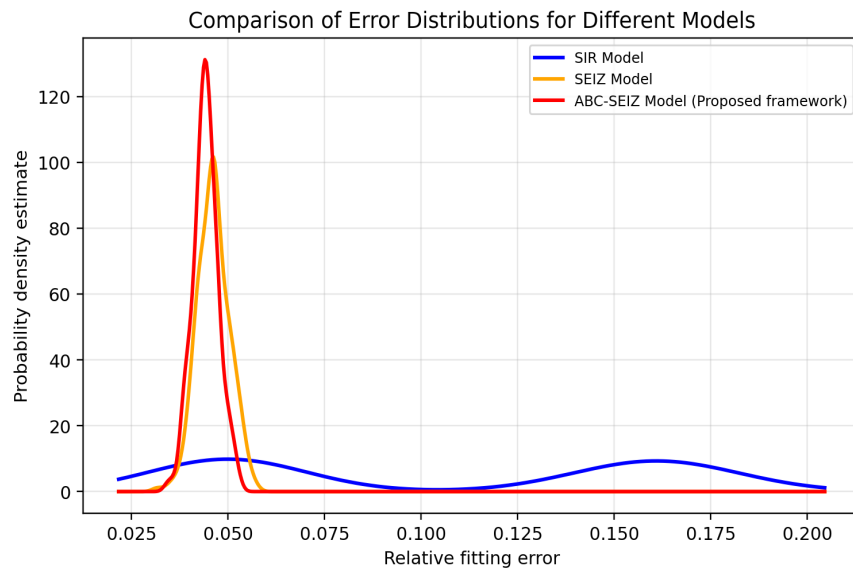
(a) COVID-19 Twitter/X cascade



(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.

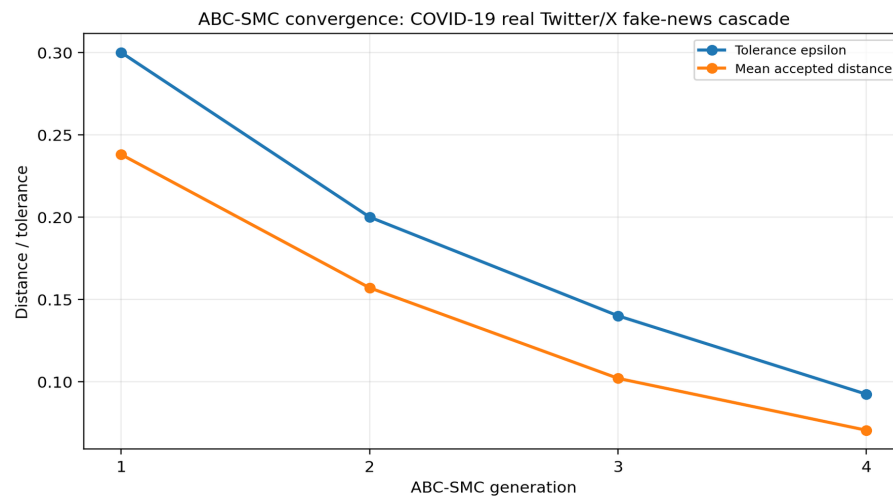
3.6.3 ABC-SMC конвергенция и калибриране на прогнозата

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.

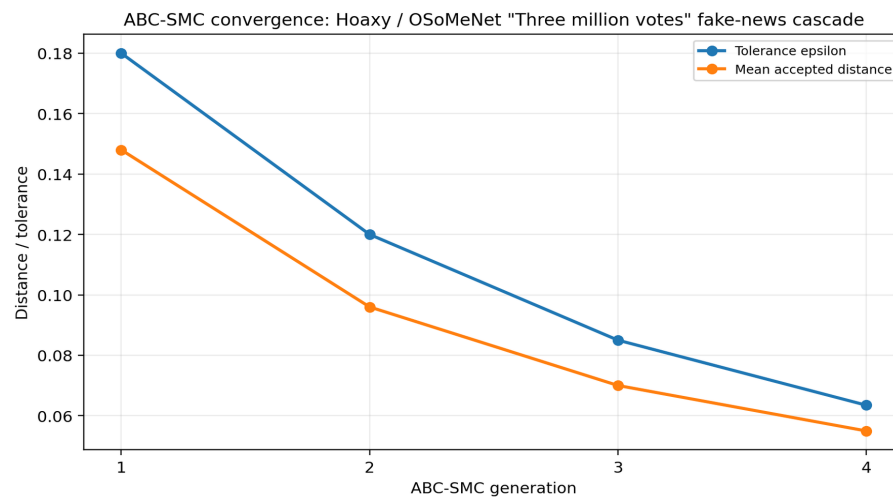
Таблица 3.13. Обобщение на конвергенцията ABC-SMC за модела за разпространение на фалшиви новини SEIZ.

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

(a) COVID-19 Twitter/X cascade

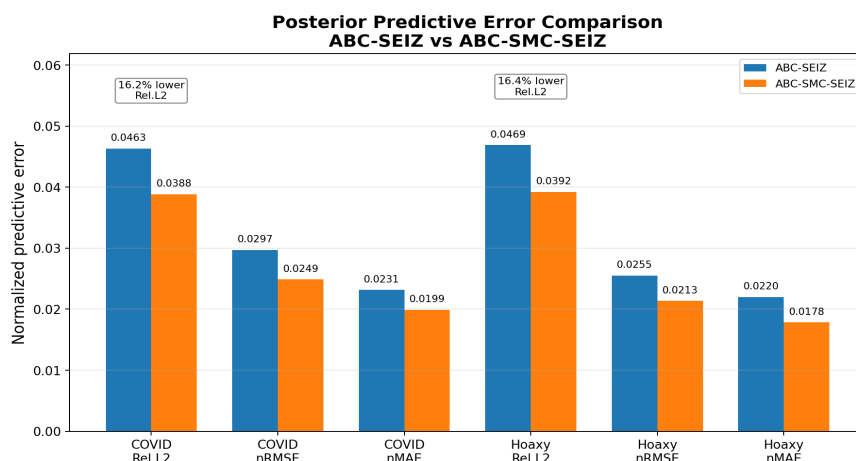


(b) Hoaxy / OSoMeNet cascade



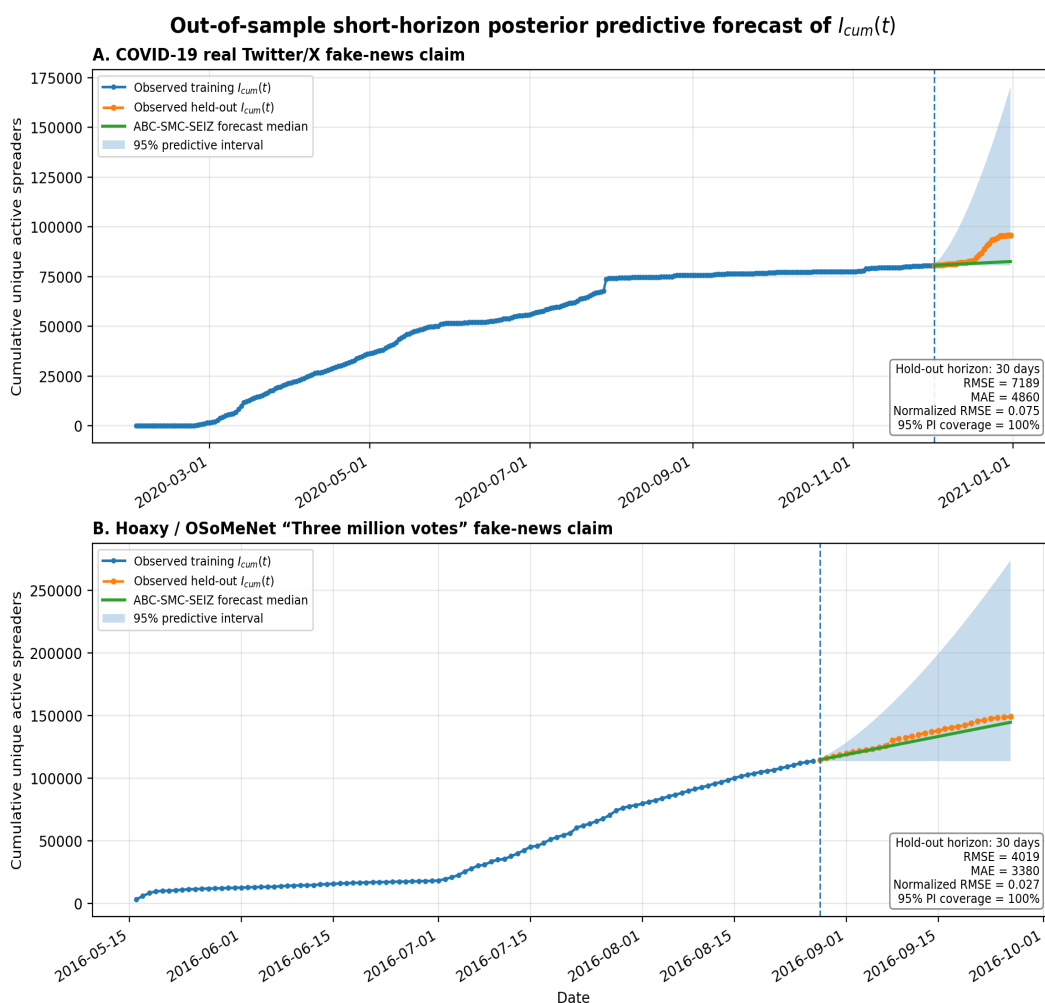
Фигура 3.12. Профили на конвергенция ABC-SMC за каскадите COVID-19 и Hoaxy / OSoMeNet.

Фигура 3.12.(a)(b) показва, че ABC-SMC прогресивно намалява приетите разстояния през поколенията, което показва стабилна постопериорна конвергенция. Профилът Hoaxy е постигнат 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 Хоаху 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 Хоаху / 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 3.14. 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.

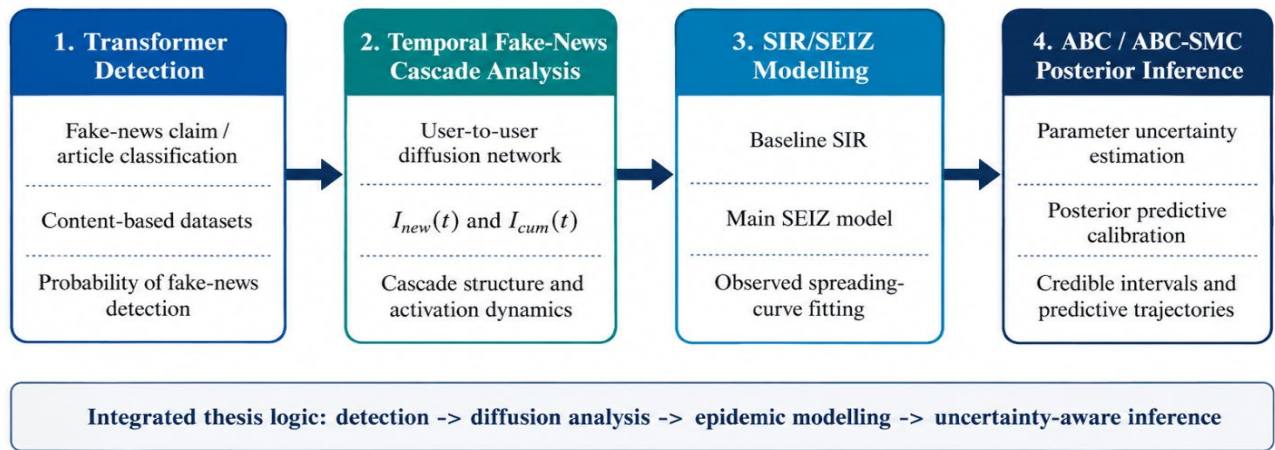
3.8 Интегрирана дискусия

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 $I_{\text{new}}(t)$, $I_{\text{cum}}(t)$, 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 Noaxy / OSoMeNet „Три милиона гласа“ са най-силните емпирични случаи, защото осигуряват достатъчна темпорална структура за епидемично моделиране и последващи изводи. Допълнителните случаи, импортирани от Noaxy 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 $I_{\text{cum}}(t)$ 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

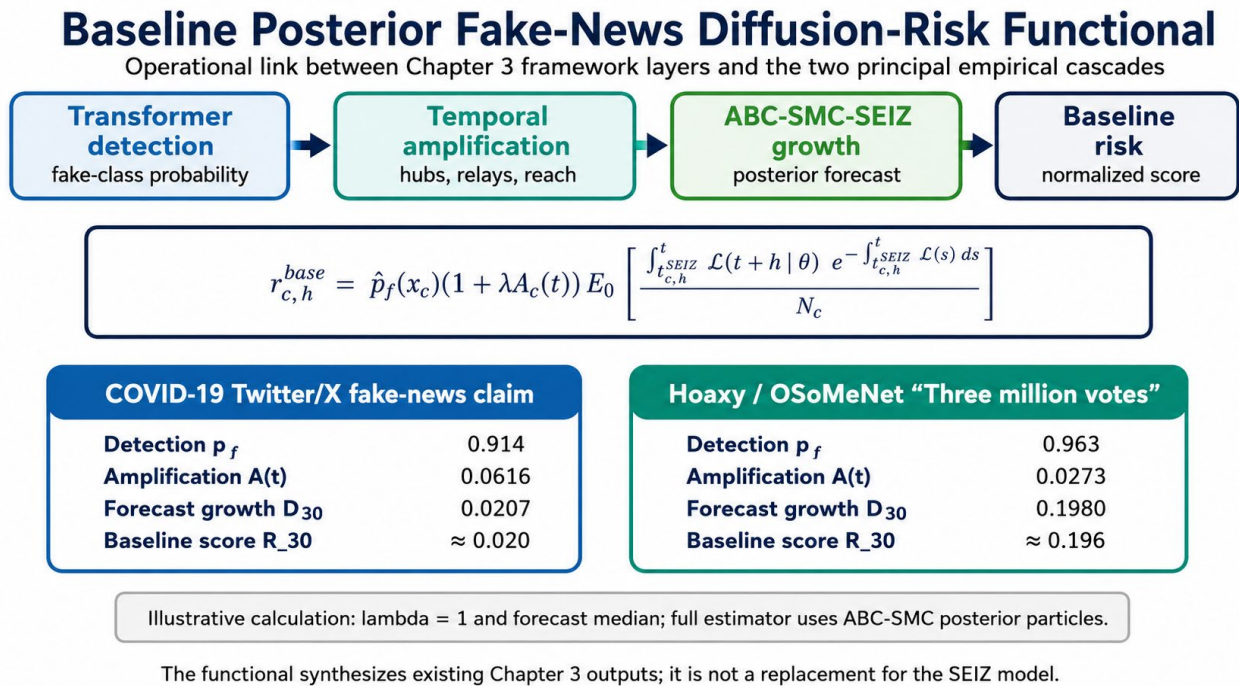


Фигура 3.15. Блок-схема на хибридната рамка на конвейера по Глава 3 от откриване на фалшиви новини, базирано на Transformer, до времеви каскаден анализ, дифузионно моделиране SIR/SEIZ и последващ извод ABC/ABC-SMC. Интегрираната рамка свързва откриването, времевия анализ, епидемичното моделиране и последващия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

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

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

Фигура 3.16 показва както концептуалната конструкция на оценката, така и двата емпирични примера. Случаите COVID-19 и Hoaxy / 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.



The functional synthesizes existing Chapter 3 outputs; it is not a replacement for the SEIZ model.

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 3.15. Обобщение на доказателствата по изследователските въпроси за Глава 3.

RQ	Доказателства в Глава 3	Заклучение
RQ1	Раздел 3.3; Таблицы 3.2-3.7 и Фигури 3.1-3.5	Производителност на трансформатора в различни набори от данни
RQ2	Раздел 3.4; Таблицы 3.8-3.9 и Фигури 3.6-3.7	Реконструирани времеви каскади
RQ3	Раздел 3.5; Таблица 3.10 и Фигури 3.8-3.9	SEIZ в сравнение със SIR
RQ4	Раздел 3.6; Таблицы 3.11-3.13 и Фигури 3.10-3.14	Оценена неопределеност на ABC/ABC-SMC
RQ5	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 модерирването; тя предоставя възпроизводим академичен метод за свързване на риска от съдържание с каскадно поведение и последваща предсказваща несигурност.

Заклучения

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, $I_{new}(t)$, $I_{cum}(t)$, 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 одитируема 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

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> [7]
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 [8]

Библиография

- [1] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, 19(1), 22–36. <https://doi.org/10.1145/3137597.3137600>
- [2] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- [3] Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontañón, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & Ahmed, A. (2020). Big Bird: Transformers for Longer Sequences. *Advances in Neural Information Processing Systems*, 33, 17283–17297. <https://proceedings.neurips.cc/paper/2020/hash/c8512d142a2d849725f31a9a7a361ab9-Abstract.html>
- [4] Kypraios, T., Neal, P., & Prangle, D. (2017). A tutorial introduction to Bayesian inference for stochastic epidemic models using Approximate Bayesian Computation. *Mathematical Biosciences*, 287, 42-53. <https://doi.org/10.1016/j.mbs.2016.07.001>
- [5] Holme, P., & Saramäki, J. (2012). Temporal networks. *Physics Reports*, 519(3), 97-125. <https://doi.org/10.1016/j.physrep.2012.03.001>
- [6] Vosoughi, S., Roy, D., & Aral, S. (2018). The Spread of True and False News Online. *Science*, 359(6380), 1146–1151. <https://doi.org/10.1126/science.aap9559>
- [7] Bozhiqi Vula, K., & Guliashki, V. G. (2025). BigSEIZ: Combining BigBird with epidemic models to analyse spreading fake news on Twitter temporal network. In L. Barolli, T. Enokido, & I. Woungang (Eds.), *Complex, Intelligent, and Software Intensive Systems: Proceedings of the 19th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS 2025)*, Fukuoka, Japan, July 2–4, 2025 (pp. 47–60). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-96096-3_5
- [8] Bozhiqi, K. (2025). A Hybrid NLP–Epidemiological Model for Fake News Network: BigBirdAssisted SEIZ Modelling with Approximate Bayesian Computation. *IEEE International Workshop on Technologies for Defense and Security (TechDefense 2025)*, Rome, Italy, November 5–7, 2025.
- [9] Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A*, 115(772), 700-721. <https://doi.org/10.1098/rspa.1927.0118>
- [10] Jin, F., Dougherty, E., Saraf, P., Cao, Y., & Ramakrishnan, N. (2013). Epidemiological modelling of news and rumors on Twitter. *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, 1-9. <https://doi.org/10.1145/2501025.2501027>