V3.2 Výskumná správa o modeloch a metódach detekcie dezinformačných kampaní

Názov projektu	**Detekcia dezinformačných naratívov a kampaní v online priestore**
Akronym	**DisTraceAI**
Kód projektu	**09I01-03-V04-00006**
Začiatok projektu	**01.07.2024**
Trvanie projektu	**24 mesiacov**

Financované Európskou úniou
NextGenerationEU

PLÁN OBNOVY

Obsah

Úvod

Koordinované dezinformačné kampane – teda systematické úsilie viacerých aktérov šíriť nepravdivé alebo zavádzajúce naratívy naprieč platformami a jazykmi – predstavujú významnú hrozbu pre demokratické inštitúcie, verejné zdravie aj bezpečnosť. Manuálny fact-checking nedokáže držať krok s rastúcim objemom obsahu ani s jeho jazykovou rozmanitosťou. Súčasné automatizované riešenia navyše často adresujú iba jednotlivé časti problému izolovane. Samostatne riešia napríklad detekciu overenia-hodných tvrdení, overovanie faktov alebo zhlukovanie podobných tvrdení, pričom len zriedka tvoria jednotný systém schopný sledovať informácie od úrovne jednotlivých viet až po úroveň rozsiahlych informačných kampaní.

Táto správa nadväzuje na výskumnú správu V3.1 (*Výskumná správa o modeloch a metódach detekcie tvrdení a naratívov*), v ktorej sme predstavili dataset MultiCW a modely na detekciu overenia-hodných tvrdení dosahujúce presnosť 92 % v 17 jazykoch. Súčasťou správy bol aj prvý návrh hierarchického systému detekcie naratívov založeného na modeloch BERTopic a HDBSCAN. V závere správy V3.1 sme načrtli ďalší smer výskumu, ktorý zahŕňal rozšírenie systému o znalostnú bázu s metaúdajmi (časová pečiatka, zdrojový kanál, autor), modelovanie časového vývoja naratívov a agregáciu súvisiacich naratívov do komplexných informačných kampaní. Predkladaná správa V3.2 dokumentuje realizáciu a vyhodnotenie týchto rozšírení. Práca nadväzuje aj na ďalší výskum nášho tímu v oblasti efektívneho dolaďovania a adaptácie jazykových modelov pre nízkozdrojové scenáre [11, 12, 13, 14]. Zdrojový kód celej pipeline spolu s datasetom je dostupný v repozitári projektu: https://github.com/kinit-sk/DisTraceAI.

Hlavné výsledky druhej výskumnej fázy projektu DisTraceAI sú:

- **Kompletná hierarchická pipeline na detekciu dezinformačných kampaní**, ktorú možno priebežne aktualizovať a ktorá spracúva viacjazyčné novinové články v piatich krokoch:
    1. detekcia overenia-hodných tvrdení,
    2. extrakcia sub-naratívov – centrálnych tvrdení ukotvených v overenia-hodných tvrdeniach,
    3. automatický odhad pravdivosti sub-naratívov,
    4. extrakcia naratívov,
    5. detekcia informačných kampaní.
- **Metódy priraďovania sub-naratívov k naratívom**, založené na kanonizovaných tvrdeniach namiesto surového textu. Súčasťou systému je kontinuálne aktualizovateľná metóda cSpecFi a jej variant SpecFi-CCS, odvodené od metódy SpecFi-CS [2], ako aj upravené verzie hustého vyhľadávania (Dense), hybridného BM25+Dense a agentická metóda Context-1, ktoré pracujú so sub-naratívmi ako vstupom. Na plnom viacjazyčnom benchmarku PolyNarrative dosahuje cSpecFi konkurencieschopné výsledky (accuracy@1 = 0,429; acc@5 = 0,729; MAP = 0,254) a spolu s variantom SpecFi-CCS prekonáva východiskovú metódu SpecFi-CS na surovom texte, pričom nevyžaduje uchovávanie pôvodných článkov. Najlepší

kompromis medzi presnosťou a výpočtovými nákladmi celkovo dosahuje husté vyhľadávanie (Dense).

- **Agentický systém odhadu pravdivosti**, ktorý dosahuje presnosť 93,1 % a macro-F1 = 0,919 podporu pri automatizovanom odhadovaní pravdivosti tvrdení, pričom konečné posúdenie zostáva na používateľovi.
- **Vyhodnotenie detekcie kampaní** na datasete FakeCTI [5], kde systém dosiahol najlepší výsledok BCubed-F = 0,833.
- **Dataset EUDisinfoAtlas**, viacjazyčný dataset sub-naratívov, naratívov a kampaní s časovými pečiatkami. Dataset bol vytvorený samotnou pipeline z korpusu EUvsDisinfo [6] a je zverejnený spolu so zdrojovým kódom.

Všetky experimenty boli realizované formou tzv. evaluačného rebríka. Každá fáza vyhodnotenia využíva výstupy predchádzajúcich krokov pipeline. Napríklad pri hodnotení priraďovania naratívov sa ako vstupy používajú sub-naratívy automaticky vygenerované systémom. Namerané výsledky preto odrážajú realistické, kumulované správanie celej pipeline, a nie výkon jednotlivých komponentov pri ideálnych vstupoch. Výstupy systému odhadu pravdivosti boli navyše manuálne overené riešiteľmi projektu.

2 Hierarchická architektúra detekcie dezinformačných kampaní

2.1 Štvorúrovňová hierarchia

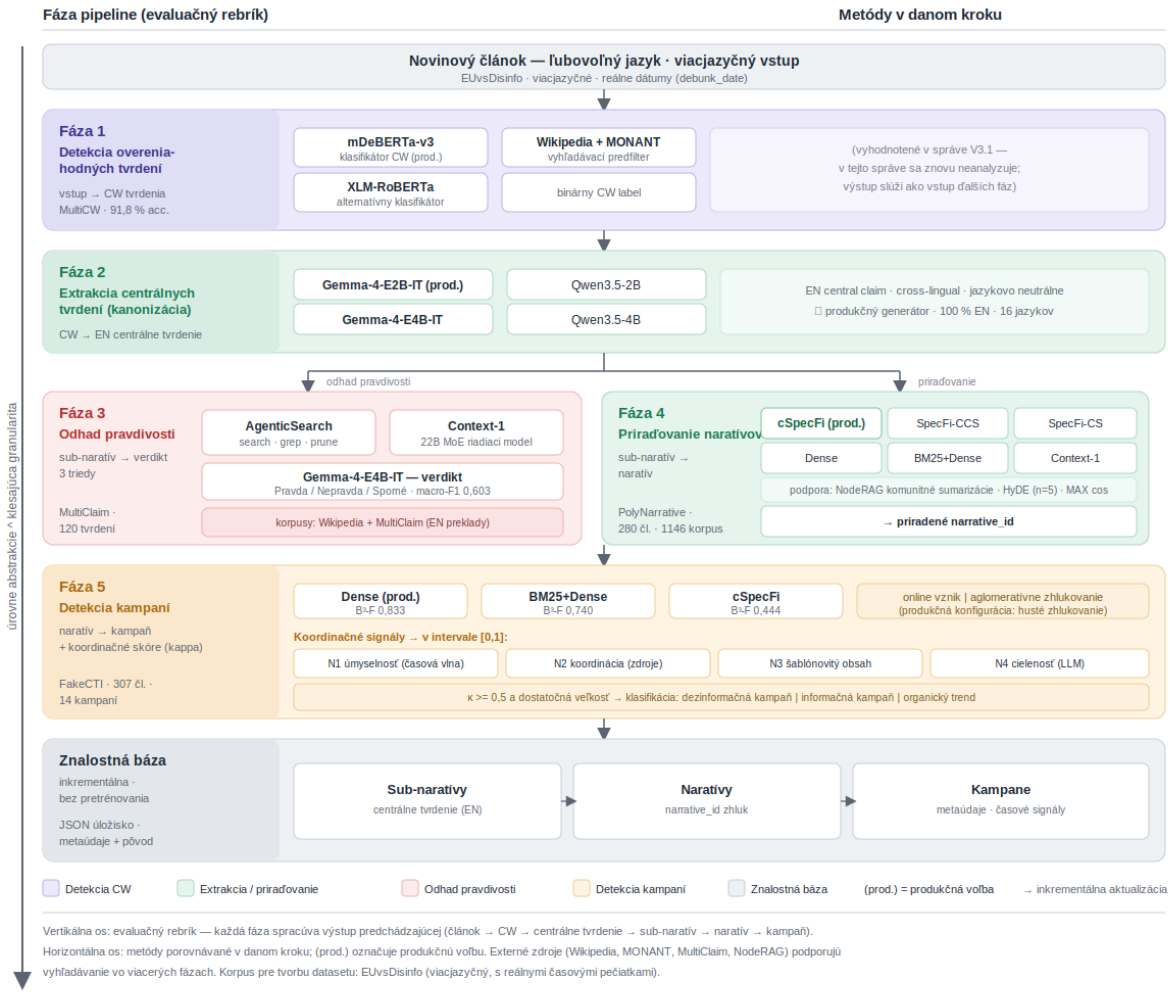
Systém organizuje obsah do štyroch hierarchických úrovní, pričom každá úroveň predstavuje vyšší stupeň abstrakcie:

1. **Overenia-hodné tvrdenie (Check-Worthy Claim, CW)**
   Veta obsahujúca overiteľné faktické tvrdenie. Definícia aj metóda detekcie boli predstavené v správe V3.1.
2. **Centrálne tvrdenie (Central Claim)**
   Dekontextualizovaná a samostatne zrozumiteľná reformulácia CW tvrdenia, preložená do angličtiny z dôvodu zjednotenia bázy dát. V procese kanonizácie sa odstraňujú anaforické odkazy, dopĺňa sa chýbajúci kontext a zachováva sa pôvodný význam tvrdenia. Preklad do jednotného jazyka umožňuje porovnávanie tvrdení naprieč rôznymi jazykmi.
3. **Sub-naratív a naratív**
   - **Sub-naratív** predstavuje zhluk tematicky súvisiacich tvrdení v rámci jedného článku. Je charakterizovaný centrálnym tvrdením a množinou podporných tvrdení.
   - **Naratív** vzniká zoskupením sémanticky príbuzných sub-naratívov, ktoré sa opakovane objavujú v rôznych článkoch a pretrvávajú v čase.
4. **Informačná kampaň**
   Najvyššiu úroveň hierarchie tvorí skupina sémanticky príbuzných naratívov, ktorých publikačné správanie naznačuje koordináciu medzi zdrojmi. Miera koordinácie je vyjadrená skóre $\kappa \in [0,1]$, ktoré kombinuje štyri signály:
   - **Úmyselnosť:** časovú synchronizáciu publikačných vĺn (publikovanie podobného obsahu naprieč zdrojmi v rámci definovaného časového okna),

- **Koordinácia:** Prítomnosť viacerých autorov/platforiem a skóre koordinácie (časová synchronicita),
- **Časový rozmer:** opakované používanie šablónovitého obsahu (takmer identické tvrdenia publikované rôznymi zdrojmi),
- **Cielenosť:** Odhad cieľovej skupiny založený na LLM (vypracovaný z úrovní subnarácie/narácie).

5. Skupina naratívov s dostatočnou veľkosťou a hodnotou κ ≥ 0,5 je klasifikovaná ako informačná kampaň. Na základe priemernej pravdivosti jej členov sa následne rozlišuje medzi dezinformačnou kampaňou, informačnou kampaňou a organickým trendom.

2.2 Fázy spracovania



Obrázok 1: Prehľad hierarchickej pipeline. Fázy spracúvajú články sekvenčne; externé znalostné zdroje (Wikipedia, MONANT, MultiClaim, NodeRAG) podporujú vyhľadávanie vo viacerých fázach. Každá fáza zapisuje do hierarchickej znalostnej bázy, ktorá rastie inkrementálne bez nutnosti pretrénovania.

Fáza 1 – Detekcia overenia-hodných tvrdení

Prvým krokom pipeline je identifikácia overenia-hodných tvrdení (Check-Worthy Claims, CW). Na tento účel používame dolaďovaný viacjazyčný klasifikátor mDeBERTa-v3, resp. XLM-RoBERTa, natrénovaný na datasete MultiCW [1], ktorý dosahuje presnosť ~92 % naprieč 17 jazykmi (podrobnosti sú uvedené v správe V3.1). Klasifikátor je doplnený o vyhľadávací predfilter využívajúci Wikipédiu a náš interný systém na monitorovanie a získavanie online zdrojov MONANT.

Keďže výkonnosť detektora bola podrobne vyhodnotená v predchádzajúcej správe, v tejto práci ju znovu neanalyzujeme. Výstupy tejto fázy však slúžia ako vstup pre všetky nasledujúce komponenty pipeline.

Fáza 2 – Extrakcia centrálnych tvrdení

Každé overenia-hodné tvrdenie je transformované na samostatnú, dekontextualizovanú anglickú propozíciu označovanú ako centrálne tvrdenie. Cieľom tejto transformácie je odstrániť jazykové a kontextové rozdiely medzi zdrojmi a vytvoriť jednotnú reprezentáciu vhodnú na ďalšie spracovanie.

V experimentoch sme porovnali štyri kompaktné jazykové modely: Qwen3.5-2B, Qwen3.5-4B, Gemma-4-E2B-IT a Gemma-4-E4B-IT. Všetky modely boli spustené lokálne prostredníctvom frameworku llama.cpp so 4-bitovou kvantizáciou Q4_K_M. Výsledky porovnania uvádzame v kapitole 3.1.

Fáza 3 – Automatický odhad pravdivosti

Odhad pravdivosti realizuje agentický systém AgenticSearchHarness riadený modelom Chroma Context-1 [8]. Systém pracuje nad dvomi korpusmi, a to nad Wikipediou a fact-checkingovým korpusom MultiClaim [7], ktorý je indexovaný prostredníctvom anglických prekladov tvrdení.

Proces prebieha iteratívne v slučke „pozoruj – uvažuj – konaj“. V každom kroku agent:

1. rozloží tvrdenie na menšie pod-dopyty,
2. vyhľadá relevantné dôkazové pasáže v korpuse vedomostí,
3. odfiltruje málo relevantné výsledky,
4. rozhodne, či je potrebné pokračovať v získavaní ďalších dôkazov.

Po ukončení vyhľadávania produkčný generátor pipeline (Gemma-4-E4B-IT) syntetizuje finálny verdikt v jednej z troch tried: **Pravda**, **Nepravda** alebo **Sporné**. Pri rozhodovaní využíva aj dynamicky odhadované apriórne rozdelenie tried odvodené z distribúcie získaných dôkazov.

Fáza 4 – Detekcia naratívov - benchmark PolyNarrative

Produkčnou metódou tejto fázy je cSpecFi, naša kontinuálne aktualizovateľná adaptácia metódy SpecFi-CS [2]. Na rozdiel od pôvodnej metódy, ktorá pracuje so surovými článkami, cSpecFi operuje výhradne na úrovni kanonizovaných tvrdení. Okrem nej v novej verzii systému

vyhodnocujeme aj ďalšie metódy priraďovania, ktoré sa líšia spôsobom vyhľadávania aj výpočtovou náročnosťou: husté vektorové vyhľadávanie (Dense), jeho hybridnú kombináciu s lexikálnym vyhľadávaním (BM25+Dense), variant SpecFi-CCS rozšírený o komunitné kondicionovanie a agentickú metódu Context-1. Porovnanie všetkých metód je predmetom kapitoly 3.

Postup pozostáva z piatich krokov:

1. **Budovanie znalostného grafu.** Graf NodeRAG [3] sa vytvára zo sub-naratívnych zhlukov (centrálne tvrdenie a jeho podporné tvrdenia) a priebežne sa rozširuje o nové dáta.
2. **Konštrukcia vyhľadávacieho korpusu.** Každý sub-naratív je reprezentovaný centrálnym tvrdením rozšíreným o komunitné sumarizácie generované systémom NodeRAG, ktoré poskytujú dodatočný tematický kontext.
3. **Generovanie hypotetických dokumentov.** Pre každý dopytový sub-naratív sa pomocou metódy HyDE vygeneruje parametrom definovaný počet hypotetických dokumentov (štandardne 5) modelom Gemma-4-E4B-IT. Generovanie je podmienené komunitným kontextom získaným z grafu.
4. **Výpočet podobnosti.** Hypotetické dokumenty aj existujúce reprezentácie naratívov sú embedované modelom Qwen3-Embedding-4B. Skóre príbuznosti je určené maximálnou kosínusovou podobnosťou.
5. **Priradenie k naratívu.** Ak podobnosť prekročí stanovený prah, sub-naratív sa priradí k existujúcemu naratívu. V opačnom prípade sa sub-naratív pridá do zoznamu nezaradených kandidátov, z ktorých sa tvoria nové naratívy, opäť na základe stanoveného prahu podobnosti.

Keďže všetky reprezentácie existujú na úrovni centrálnych tvrdení, články v rôznych jazykoch sú porovnávané v spoločnom anglickom sémantickom priestore. Po spracovaní zároveň nie je potrebné uchovávať pôvodné texty článkov.

Fáza 5 – Detekcia kampaní

V poslednej fáze sa naratívy agregujú do kandidátnych informačných kampaní. Zoskupovanie môže prebiehať buď online postupným vznikom nových naratívov, alebo aglomeratívnym zhlukovaním nad úplnou maticou podobností.

Každá kandidátna skupina je následne vyhodnotená pomocou koordinačných signálov opísaných v kapitole 2.1. Skupiny, ktoré prekročia definované prahy veľkosti a koordinácie, sú označené ako informačné kampane. Na základe priemernej pravdivosti ich členov sú následne klasifikované ako dezinformačné kampane, informačné kampane alebo organické trendy.

2.3 Inkrementálna znalostná báza a implementácia

Kľúčovým návrhovým princípom systému je **inkrementálna aktualizovateľnosť bez potreby pretrénovania modelov**. Pri spracovaní nových článkov pipeline vykonáva výpočty iba nad novopridaným obsahom, zatiaľ čo všetky predchádzajúce výsledky sa načítavajú priamo zo

znalostnej bázy. Vďaka tomu možno systém priebežne rozširovať bez opakovaného spracovania celého korpusu.

Základom architektúry je **hierarchická znalostná báza** implementovaná ako JSON úložisko organizované na úrovniach článkov, sub-naratívov, naratívov a kampaní. Každý záznam obsahuje jednoznačný identifikátor a kompletné metadáta vrátane zdrojovej URL, času publikácie, jazyka a zoznamu prispievajúcich článkov. Týmto spôsobom systém napĺňa rozšírenia avizované v správe V3.1, konkrétne podporu meta-informácií a sledovanie pôvodu jednotlivých výstupov.

Pre analýzu výsledkov je k dispozícii samostatný interaktívny HTML dashboard s časovou osou a hierarchickou navigáciou umožňujúcou prechod od kampaní cez naratívy až po jednotlivé tvrdenia. Súčasťou systému je aj konfigurovateľné terminálové rozhranie určené na dávkové spracovanie a experimentovanie.

Celá pipeline je navrhnutá na lokálne nasadenie. Všetky modely sú prevádzkované v kvantizovanej podobe a systém nevyžaduje využívanie externých API služieb, čo znižuje prevádzkové náklady a zároveň zvyšuje kontrolu nad spracovávanými dátami.

3 Experimentálne vyhodnotenie

Všetky experimenty boli navrhnuté podľa princípu **evaluačného rebríka** opísaného v kapitole 1. Každá hodnotená fáza využíva výstupy predchádzajúcich krokov pipeline, čím sa simuluje reálne nasadenie systému. Namerané výsledky preto odrážajú kumulatívny vplyv chýb a neistôt vznikajúcich v predchádzajúcich fázach spracovania.

Konkrétne:

- benchmark extrakcie centrálnych tvrdení vyhodnocuje prepisy tvrdení identifikovaných vo fáze 1,
- benchmark priraďovania naratívov používa ako vstupy sub-naratívy vytvorené samotnou pipeline,
- benchmark odhadu pravdivosti pracuje s automaticky extrahovanými tvrdeniami,
- benchmark detekcie kampaní zoskupuje naratívy vytvorené vo fáze 4.

Všetky experimenty boli realizované na jedinej GPU NVIDIA H200 so 140 GB pamäte.

3.1 Extrakcia centrálnych tvrdení

Cieľom experimentu je vyhodnotiť schopnosť modelov transformovať overenia-hodné tvrdenia do podoby samostatných a jazykovo jednotných centrálnych tvrdení.

Benchmark pozostáva z 32 referenčných overenia-hodných tvrdení reprezentovaných v 16 jazykoch. Keďže výsledná reprezentácia musí byť vhodná na medzijazykové porovnávanie, primárnou automatickou metrikou je **jazyková konzistentnosť výstupov**, definovaná ako podiel výstupov vygenerovaných v anglickom jazyku.

Automatické hodnotenie je doplnené manuálnou analýzou všetkých výstupov, pri ktorej riešitelia overujú zachovanie pôvodného významu tvrdenia po jeho transformácii do centrálneho tvrdenia.

Model	Parametre	Miera EN	Medián latencie	Priemer latencie
Qwen3.5-2B	2B	100.0%	0,13 s	0,20 s
Qwen3.5-4B	4B	87.5%	0,53 s	0,54 s
Qwen3.5-9B	9B	90.6%	0,67 s	0,67 s
Gemma-4-E2B-IT	2B	100,0 %	0,14 s	0,17 s
Gemma-4-E4B-IT ★	4B	100,0 %	0,21 s	0,21 s
Gemma-4-12B-IT	12B	100,0 %	0,30 s	0,31 s

Tabuľka 1: Benchmark extrakcie centrálnych tvrdení. ★ = produkčný generátor pipeline.

Modely Gemma-4-E2B-IT, Gemma-4-E4B-IT a Gemma-4-12B-IT dosiahli 100 % podiel výstupov v anglickom jazyku. Rovnakú úspešnosť dosiahol aj model Qwen3.5-2B, zatiaľ čo modely Qwen3.5-4B a Qwen3.5-9B dosiahli hodnoty 87,5 %, respektíve 90,6 %. Manuálna analýza ukázala, že hlavnou príčinou neanglických výstupov bolo generovanie procesu uvažovania pred samotnou odpoveďou, prípadne miešanie viacerých jazykov v rámci jedného výstupu.

Kvalitatívna analýza zachovania významu preukázala, že model Gemma-4-E4B-IT generuje pri extrakcii centrálnych tvrdení najkonzistentnejšie, významovo najvernejšie a informačne najúplnejšie výstupy. Model Gemma-4-E2B-IT však dosahuje porovnateľnú kvalitu pri nižšej latencii (medián 0,14 s oproti 0,21 s). Model Gemma-4-12B-IT síce zachoval rovnakú úroveň jazykovej konzistencie, avšak pri tejto úlohe neposkytol zodpovedajúce zvýšenie kvality vzhľadom na vyššie výpočtové nároky.

Na základe uvedených výsledkov bol ako produkčný model pre fázu extrakcie centrálnych tvrdení zvolený Gemma-4-E4B-IT, ktorý poskytuje najlepší pomer medzi kvalitou generovaných výstupov a výpočtovou efektívnosťou. Model Gemma-4-12B-IT bol naopak nasadený v kroku HyDE, kde vyššia generačná kapacita veľkého jazykového modelu prináša väčší prínos pri tvorbe hypotetických dokumentov. Pri syntéze finálnych verdiktov pravdivosti bol rovnako využitý model Gemma-4-E4B-IT, ktorého kvalita generovaných výstupov sa ukázala ako dostatočná pri nižších výpočtových nárokoch.

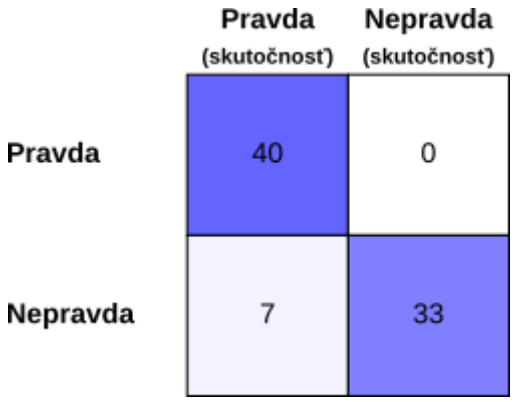
3.2 Automatický odhad pravdivosti

Vyhodnotenie bolo realizované na vyváženej vzorke 80 tvrdení z korpusu MultiClaim, ktorá obsahovala 40 tvrdení v každej z troch tried: **Pravda**, **Nepravda** a **Sporné**. Na rozdiel od skorších experimentov, ktoré pracovali s prirodzene nevyváženými dátami, toto nastavenie umožňuje hodnotiť schopnosť modelu rozlišovať medzi jednotlivými triedami bez zvýhodnenia majoritnej kategórie. Všetky predikované verdikty boli následne manuálne overené riešiteľmi.

Trieda	Podpora	Presnosť (Precision)	Úplnosť (Recall)	F1
Pravda	40	0,851	1,000	0,919
Nepravda	40	1,000	0,825	0,904
Presnosť / macro-F1				0,913 / 0,912

Tabuľka 3: Odhad pravdivosti na 78 vyvážených tvrdeniach MultiClaim (AgenticSearchHarness: vyhľadávanie Context-1, verdikt Gemma-4-E4B-IT).

Navrhovaný systém dosiahol na vyváženom benchmarku 91,3 % presnosti a 0,912 macro-F1, čo výrazne presahuje cieľovú hranicu 80%. Lokálny korpus MultiClaim pokrýva 100 % dopytov s priemernou spoľahlivosťou > 0,9 — systém prehľadáva Wikipédiu alebo web len pri okrajových prípadoch, ktoré v tomto benchmarku nenastali. Konfúzna matica (riadky = gold, stĺpce = predikcia) zachytáva chybový profil systému:



Obrázok 4: Matica zámien odhadu pravdivosti (120 vyvážených tvrdení).

Trieda Pravda je rozpoznaná bezchybne (recall = 1,00; všetkých 40 z 40 tvrdení bolo správne klasifikovaných). Presnosť na tejto triede dosahuje 0,851, keďže systém v 7 prípadoch nesprávne označil tvrdenie triedy Nepravda ako Pravda.

Trieda Nepravda si zachováva perfektnú presnosť (P = 1,00) – žiadne tvrdenie predikované ako Nepravda nebolo v skutočnosti triedy Pravda. Úplnosť však dosahuje 0,825: zo 40 tvrdení anotovaných ako Nepravda systém správne identifikoval 33, zatiaľ čo 7 nesprávne klasifikoval ako Pravda. Toto asymetrické správanie naznačuje konzervatívny charakter systému, ktorý pri nejednoznačných prípadoch uprednostňuje predikciu Pravda pred chybným označením tvrdenia ako Nepravda, čím eliminuje falošne pozitívne vyvrátenia za cenu nižšej úplnosti triedy Nepravda.

3.3 Detekcia naratívov – benchmark PolyNarrative

Detekciu naratívov vyhodnocujeme na plnom viacjazyčnom benchmarku PolyNarrative [4], ktorý pokrýva dve tematické domény: klimatickú zmenu (CC) a vojnu na Ukrajine (URW).

Vyhodnotenie prebieha nad korpusom 1146 referenčných naratívnych položiek zoskupených do 7 naratívnych kategórií. Na vyhodnotenie sa používa 280 testovacích článkov, z ktorých pipeline automaticky vygenerovala sub-naratívy. Skórovanie je vykonané na úrovni článku

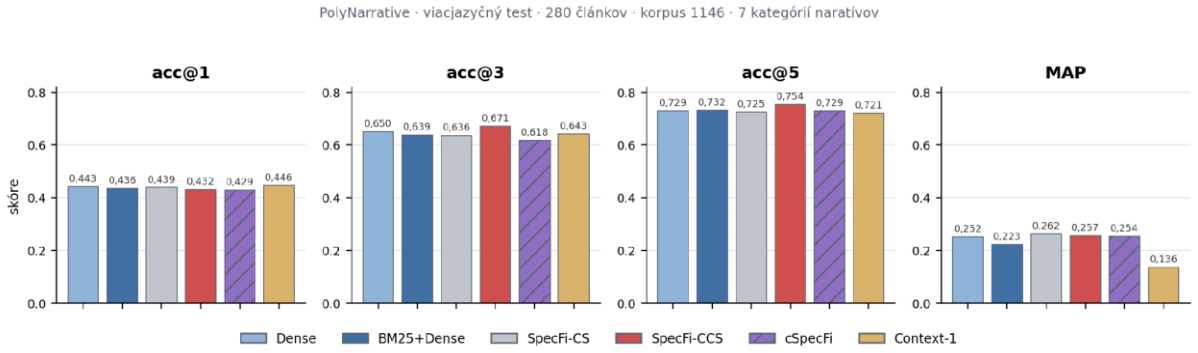
(best-rank): pre každý článok sa berie najlepšie umiestnenie gold naratívnej položky spomedzi jeho dopytových jednotiek, a to v metrikách acc@1, acc@3, acc@5 a MAP.

Každú metódu hodnotíme v jej natívnej konfigurácii: metódy nad kanonizovanými tvrdeniami (cSpecFi, SpecFi-CS, SpecFi-CCS) pracujú so sub-naratívnymi reprezentáciami centrálnych tvrdení, husté a lexikálne metódy (Dense, BM25+Dense) s embeddingmi a agentická metóda Context-1 s iteratívnym vyhľadávaním. Metódy sa tak líšia nielen presnosťou, ale aj

Metóda	acc@1	acc@3	acc@5	MAP	Výpočtová náročnosť
Dense	**0,443**	0,650	0,729	0,252	nízka
BM25+Dense	0,436	0,639	0,732	0,223	nízka
SpecFi-CS	0,439	0,636	0,725	**0,262**	vysoká
SpecFi-CCS	0,432	**0,671**	**0,754**	0,257	vysoká
cSpecFi	0,429	0,618	0,729	0,254	vysoká
Context-1	**0,446**	0,643	0,721	0,136	najvyššia

výpočtovou náročnosťou.

Tabuľka 2: Detekcia naratívov na plnom viacjazyčnom testovacom rozdelení PolyNarrative (280 článkov, korpus 1 146 položiek, 7 naratívnych kategórií; detektor mDeBERTa). Skóre je počítané na úrovni článku (best-rank). Žiadna metóda nedominuje vo všetkých metrikách: SpecFi-CCS dosahuje najlepšie acc@3 a acc@5, Context-1 a Dense najlepšie acc@1, SpecFi-CS najlepšiu MAP. Pri zohľadnení výpočtovej náročnosti predstavuje najlepší kompromis medzi presnosťou a nákladmi metóda Dense (zvýraznené bunky označujú najlepšiu hodnotu v danom stĺpci).



Obrázok 2: Porovnanie metód priraďovania naratívov v metrikách acc@1, acc@3, acc@5 a MAP (280 článkov, korpus 1 146 položiek). Šrafovaný stĺpec = navrhovaná metóda cSpecFi.

Metódy nad kanonizovanými tvrdeniami prekonávajú východiskovú metódu SpecFi-CS

Navrhovaná hierarchia vyžaduje reprezentáciu vo forme kanonizovaných tvrdení. Spomedzi metód pracujúcich s touto reprezentáciou dosahuje variant SpecFi-CCS najlepšie acc@3 (0,671) a acc@5 (0,754) z celého porovnania, zatiaľ čo navrhovaná metóda cSpecFi dosahuje hodnoty acc@1 = 0,429, acc@5 = 0,729 a MAP = 0,254. Je tak porovnateľná s hustým vyhľadávaním nad rovnakým korpusom (acc@1 = 0,443), pričom pracuje výhradne s kanonizovanými tvrdeniami a nevyžaduje uchovávanie pôvodných článkov. Najvyššiu hodnotu

acc@1 dosahujú metódy Context-1 (0,446) a Dense (0,443), zatiaľ čo najlepšie acc@3 (0,671) a acc@5 (0,754) dosahuje variant SpecFi-CCS.

Reprezentácia článku viacerými sub-naratívnymi zhlukmi umožňuje viaceré tematické rámcovania, čo vedie k presnejším dopytom v kroku HyDE a znižuje efekt priemerovania obsahovo odlišných tvrdení v rámci jedného článku.

Kanonizácia prináša len minimálnu stratu výkonu

Prechod zo surových článkov na reprezentáciu prostredníctvom kanonizovaných tvrdení neprináša stratu výkonu: metódy nad kanonizovanými tvrdeniami (cSpecFi, SpecFi-CS, SpecFi-CCS) dosahujú porovnateľné alebo lepšie skóre než východisková metóda SpecFi-CS nad surovým textom, ktorá predstavuje súčasný stav poznania(SOTA). Variant SpecFi-CCS ju prekonáva v acc@3 aj acc@5, a to bez nutnosti uchovávať pôvodné viacjazyčné texty.

Tento rozdiel je vzhľadom na výhody kanonizovanej reprezentácie relatívne malý. Reprezentácia založená na tvrdeniach eliminuje potrebu dlhodobého uchovávania a embedovania plných viacjazyčných článkov, poskytuje jazykovo jednotný priestor na porovnávanie obsahu a umožňuje nasadenie v prostrediach s licenčnými alebo súkromnostnými obmedzeniami, kde nie je možné archivovať pôvodné texty.

Dodatočné porovnanie na jednojazyčnom scenári (ruština, doména URW; 56 testovacích článkov, 10 naratívov) ukazuje, že metóda SpecFi-CS dosahuje len mierne lepšie výsledky než husté vyhľadávanie (+0,018 accuracy@1 a +0,005 MRR), pričom vyžaduje približne 750-násobne dlhší čas spracovania (1 805 s oproti 2 s).

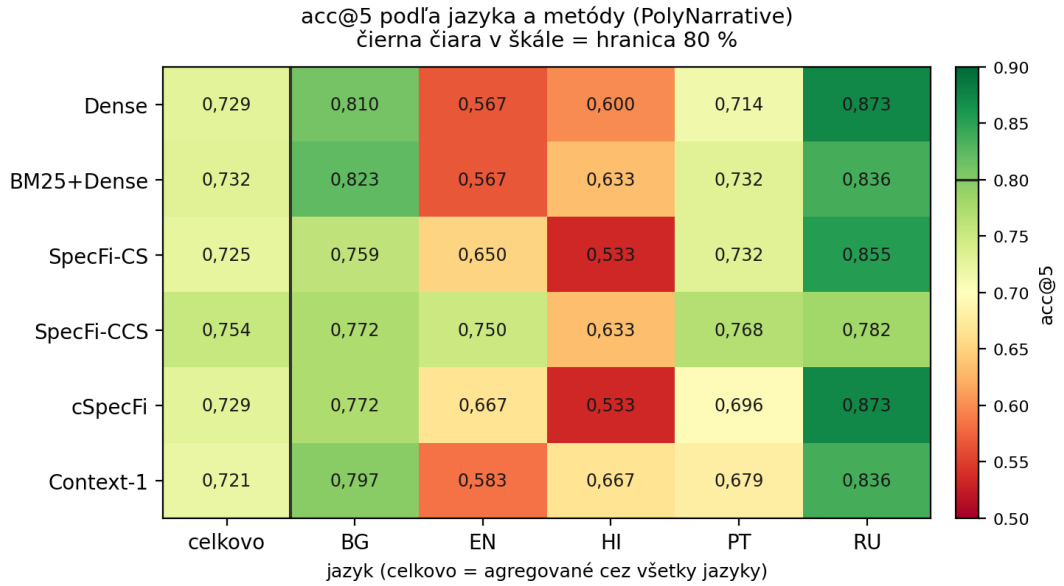
Agentické vyhľadávanie je presné v top-k, ale drahé a slabšie v celkovom zoradení

Agentická metóda Context-1 dosahuje najvyššie acc@1 (0,446) z celého porovnania a konkurencieschopné acc@5 (0,721), takže pri hľadaní správneho naratívu v rámci top-k nezaostáva. Jej slabinou je celkové zoradenie kandidátov: MAP je výrazne najnižšia spomedzi všetkých metód (0,136 oproti 0,223–0,262 pri ostatných), čo znamená, že za prvými niekoľkými zásahmi kvalita rebríčka rýchlo klesá. V kombinácii s rádovo vyššími výpočtovými nákladmi to z agentického radenia robí presný, ale drahý nástroj vhodný skôr pre menšie alebo špecializované korpusy než ako predvolené riešenie pre rozsiahle viacjazyčné zbierky.

Z hľadiska nasadenia preto pri rozsiahlych viacjazyčných zbierkach predstavuje najlepší kompromis medzi presnosťou a nákladmi husté vyhľadávanie (Dense), prípadne doplnené o kanonizované reprezentácie tvrdení (cSpecFi, SpecFi-CCS) tam, kde nie je možné archivovať pôvodné texty.

Výkon je výrazne heterogénny naprieč jazykmi (Obrázok 3). Najvyššiu úspešnosť pipeline dosahuje v ruštine a bulharčine, kde viaceré metódy prekračujú hranicu acc@5 = 0,80 (ruština až 0,87 pri Dense a cSpecFi, bulharčina 0,81–0,82). Naopak v angličtine a hindčine je úspešnosť nižšia, čo zodpovedá menšej podpore týchto jazykov v korpuse a ich obsahovej rozmanitosti. Celkové acc@5 = 0,754 (najlepšie pre SpecFi-CCS) sa tak k hranici 80 % blíži, hoci ju v agregovanom priemere ešte neprekračuje. Z hľadiska metód poskytujú v silných

jazykoch komunitné sumarizácie NodeRAG (cSpecFi, SpecFi-CCS) rozlišujúci kondicionačný kontext, zatiaľ čo husté vyhľadávanie zostáva silnou a výpočtovo lacnou voľbou.



Obrázok 3: acc@5 podľa jazyka a metódy na testovacom rozdelení PolyNarrative (stĺpec „celkovo" = agregované cez všetky jazyky). Zelené bunky označujú vyšší výkon; čierna čiara v škále vyznačuje hranicu 80 %. Najvyššiu úspešnosť dosahujú ruština a bulharčina, kde viaceré metódy prekračujú 80 %.

3.4 Detekcia kampaní – benchmark FakeCTI

Detekciu kampaní vyhodnocujeme na datasete FakeCTI [5], ktorý predstavuje prvý dataset systematicky prepájajúci falošné správy s pomenovanými dezinformačnými kampaňami a aktérmi. Pre experiment sme použili podmnožinu WEB, ktorá obsahuje obnoviteľné časové pečiatky odvodené z URL adries a gold kampane s minimálne piatimi datovanými článkami.

Vybraný subset obsahuje 307 článkov rozdelených do 14 kampaní. Fáza 4 pipeline z nich automaticky vytvorila 358 sub-naratívov a 125 naratívov, ktoré boli následne použité ako vstup pre všetky porovnávané metódy.

Metóda	Detegované kampane	Zhlukované čl.	B^3 P	B^3 R	B^3 F	ARI	NMI
Dense	2	86 / 307	0,714	1,000	0,833	0,465	0,534
BM25+Dense	2	90 / 307	0,594	0,982	0,740	0,514	0,497
cSpecFi	11	115 / 307	0,683	0,330	0,444	0,139	0,451

Tabuľka 4: Detekcia kampaní na FakeCTI (307 datovaných článkov, 14 gold kampaní; identický vstup fázy 4 pre všetky metódy; online zoskupovanie).

Porovnanie metód

Na hodnotenie kvality zhlukovania používame metriky **BCubed precision**, **BCubed recall** a ich harmonický priemer (**BCubed-F**), ktoré hodnotia čistotu a úplnosť vytvorených zhlukov na úrovni jednotlivých objektov. Celkovú zhodu medzi predikovaným a referenčným zhlukovaním dopĺňajú metriky **Adjusted Rand Index (ARI)**, merajúca podobnosť dvoch rozdelení po

korekcii na náhodu, a **Normalized Mutual Information (NMI)**, ktorá vyjadruje mieru zdieľanej informácie medzi nimi.

Najlepšie výsledky dosahuje jednoduché husté zhlukovanie, ktoré dosahuje BCubed-F = 0,833 pri dokonalom BCubed recall. Tesne za ním nasleduje hybridné zhlukovanie s BCubed-F = 0,740 a najlepšou hodnotou ARI = 0,514.

Metóda cSpecFi na tejto úrovni vykazuje odlišné správanie. Namiesto agregácie naratívov do väčších celkov dochádza k ich jemnej fragmentácii – systém deteguje 11 menších kampaní, ktoré sú z hľadiska čistoty vysoko presné (viaceré skupiny dosahujú 100 % čistotu, BCubed precision = 0,683), avšak rozdeľuje dve dominantné gold kampane datasetu. To vedie k výraznému poklesu recallu na 0,330.

HyDE mechanizmus, ktorý je prínosný pri rozlišovaní medzi jemne odlíšenými naratívmi (napr. 21 sémanticky blízkych cieľov v predchádzajúcich experimentoch), sa v tomto prípade ukazuje ako nevhodný. Pri hrubej štruktúre datasetu vedie k nadmernej diferenciácii a strate globálnej koherencie kampaní.

Interpretácia výsledkov: princíp granularity

Koordinačné skórovanie správne identifikuje skupinu s najvýraznejšou časovou synchronizáciou publikačných vĺn (burst = 0,50) a všetky detegované skupiny klasifikuje ako organické trendy, čo zodpovedá archívnemu charakteru datasetu FakeCTI.

Spolu s výsledkami z kapitol 3.2 a 3.3 tieto experimenty odhaľujú všeobecný princíp **granularity reprezentácie**:

- Reprezentačne citlivé vyhľadávanie (cSpecFi) je výhodné v úlohách s jemnozrnnou štruktúrou cieľov, kde je potrebné rozlišovať medzi mnohými sémanticky blízkymi naratívmi (priraďovanie naratívov).
- Reprezentačne agnostické metódy založené na súhrnnej podobnosti sú naopak výhodnejšie pri hrubých zhlukovacích úlohách, kde je cieľová štruktúra výrazne menej jemná (detekcia kampaní).

Keďže navrhovaná pipeline umožňuje nezávislý výber metódy pre každú úroveň, z týchto výsledkov priamo vyplýva produkčná konfigurácia: cSpecFi pre fázu 4 a husté zhlukovanie pre fázu 5.

4 Dataset EUDisinfoAtlas

Zdroje obsahujúce anotácie na úrovni naratívov so spoľahlivými časovými pečiatkami sú mimo anglického jazyka naďalej veľmi obmedzené. Z tohto dôvodu sme pomocou navrhovanej pipeline vytvorili dataset **EUDisinfoAtlas**, ktorý zachytáva sub-naratívy, naratívy a dezinformačné kampane nad viacjazyčným spravodajským korpusom. Dataset je publikovaný spolu so zdrojovým kódom pipeline a všetkými experimentálnymi výsledkami v repozitári projektu (https://github.com/kinit-sk/DisTraceAI) a zároveň napĺňa jeden z kľúčových ukazovateľov projektu, konkrétne zdieľanie datasetu naratívov a datasetu dezinformačných

kampaní. Na rozdiel od ich samostatného publikovania sú obe úrovne reprezentované v jednom vzájomne prepojenom artefakte, ktorý zároveň rozširuje možnosti už publikovaného datasetu MultiCW.

Zdroj dát a metaúdaje

Vstupné dáta tvoria korpus **EUvsDisinfo** [6], rozsiahly viacjazyčný dataset prokremeľských dezinformačných článkov spárovaných s dôveryhodnými spravodajskými článkami. Dataset bol zostavený na základe odborných vyvrátení (debunkov) projektu EUvsDisinfo. V porovnaní s predtým používanými korpusmi obsahuje spoľahlivé časové pečiatky, ktoré umožňujú priamy výpočet koordinačných signálov, predovšetkým časovej synchronizácie publikačných vĺn, bez potreby rekonštrukcie dátumov.

Pipeline z každého záznamu využíva telo článku, rekonštruované z pôvodných URL adries pomocou nástroja EUvsDisinfo, ako vstup pre detekciu tvrdení. Dátum vyvrátenia (*debunk_date*) slúži ako časová pečiatka pri výpočte koordinačných signálov. Ďalšími využívanými metaúdajmi sú jazyk článku (*article_language*), doména alebo vydavateľ a trieda článku (*class*: dezinformačný alebo dôveryhodný). Identifikátor vyvrátenia (*debunk_id*) spolu s triedou predstavujú potenciálnu referenčnú pravdu (*gold standard*) pri vyhodnocovaní.

Verejne dostupná verzia datasetu EUvsDisinfo obsahuje iba metaúdaje a URL adresy. Plné texty článkov sa získavajú pomocou distribuovaného rekonštrukčného nástroja. Záznamy, pri ktorých sa nepodarilo získať obsah článku, sú z ďalšieho spracovania vyradené a evidované samostatne. Keďže korpus pokrýva viacero desiatok jazykov, dataset EUDisinfoAtlas nie je obmedzený na slovenčinu alebo češtinu, ale umožňuje analýzu viacjazyčných dezinformačných naratívov a kampaní.

Štruktúra datasetu

Dataset pozostáva z troch vzájomne prepojených CSV tabuliek (*sub_narratives.csv*, *narratives.csv* a *campaigns.csv*), ktoré zodpovedajú hierarchickej štruktúre navrhovanej znalostnej bázy.

Tabuľka sub-naratívov obsahuje syntetizované centrálne tvrdenie v anglickom jazyku, zoznam podporujúcich kanonizovaných tvrdení, odhad pravdivosti (ak bol vypočítaný) a metaúdaje prevzaté z datasetu EUvsDisinfo, vrátane domény, jazyka, časovej pečiatky (*debunk_date*) a triedy článku.

Tabuľky naratívov a kampaní agregujú centrálne tvrdenie príslušného uzla, časový rozsah jeho výskytu, jazykovú a doménovú diverzitu členov a agregovaný odhad pravdivosti. V prípade kampaní sú navyše uložené hodnoty koordinačného skóre spolu so štyrmi čiastkovými komponentmi: **N1** – časová synchronizácia, **N2** – koamplifikácia, **N3** – obsahové opakovanie a **N4** – krížovo-jazyková koexistencia (*cross-lingual co-occurrence*).

Z dôvodu licenčných obmedzení dataset neobsahuje pôvodné texty článkov. Verejne sú sprístupnené iba odvodené štruktúry a metaúdaje, ktoré možno spätne prepojiť s pôvodnými článkami prostredníctvom ich URL adries.

Aktuálny rozsah datasetu

Tabuľka 6 sumarizuje aktuálny rozsah datasetu po dokončení extrakcie nad celým korpusom EUvsDisinfo. Pipeline bola spustená v režime úplnej extrakcie, ktorý zahŕňal kroky 1 až 5, pričom štvrtý krok bol vykonaný v zrýchlenom režime. Na detekciu overeniahodných tvrdení bol použitý model **mDeBERTa-v3 (mdb-multicw)**, na kanonizáciu model **Gemma-4-E4B-IT (bf16)** a na konštrukciu sub-naratívov, naratívov a kampaní hybridný retrieval backend kombinujúci BM25 a dense retrieval s fúziou pomocou algoritmu **Reciprocal Rank Fusion (RRF)**.

Ukazovateľ	Hodnota
Počet sub-naratívov	3 115
Z toho zdrojových článkov (pokrytie EUvsDisinfo)	2 505 (priemerne 1,24 sub-naratívu na článok)
Počet naratívov	1 002 (priemerný počet sub-naratívov na naratív: 6,0; max. 2 692)
Počet kampaní	446 (priemerný počet naratívov na kampaň: 2,2; max. 14)
Cross-lingválne kampane (≥ 2 jazyky)	442 zo 446 (99,1 %)
Priemerný počet jazykov na kampaň	3,91 (max. 29)
Detegované jazyky sub-naratívov	20+ (najsilnejšie: RU 19,5 %, EN 15,4 %, AR 7,2 %, UK, KA, DE, ES, HU, IT, CS, FR, PL, HY, FI, …)
Detektor použitý pri extrakcii	mDeBERTa-v3 (mdb-multicw)
Retrieval backend pre úroveň naratívov a kampaní	BM25 + Dense (RRF fúzia)
Časové pokrytie (z debunk_date)	podľa EUvsDisinfo (2015 – 2024)

Tabuľka 6. Aktuálny rozsah datasetu EUDisinfoAtlas po dokončení extrakcie nad celým korpusom EUvsDisinfo. Uvedené hodnoty zodpovedajú verzii datasetu publikovanej spolu so zdrojovým kódom pipeline v repozitári projektu.

Vysoký podiel viacjazyčných kampaní (99,1 %) potvrdzuje hlavnú motiváciu vytvorenia datasetu. Na úrovni kampaní pipeline úspešne prepája tematicky príbuzné naratívy bez ohľadu na jazyk ich pôvodu, čo predstavuje vlastnosť, ktorú existujúce jednojazyčné benchmarky vrátane datasetu FakeCTI neumožňujú spoľahlivo hodnotiť. Jazyková distribúcia sub-naratívov je dominovaná ruštinou, angličtinou a arabčinou, čo odráža charakter zdrojového korpusu EUvsDisinfo. Súčasne však dataset pokrýva viac ako dvadsať jazykov vrátane jazykov strednej Európy (čeština, slovenčina, poľština a maďarčina), čím podporuje regionálne zameranie projektu.

5 Odhad vplyvu na proces overovania faktov

Projekt definuje aj dopadové ukazovatele zamerané na prácu fact-checkerov. Cieľom je skrátenie času overovania faktuality online obsahu z približne 3 dní (údaj z Participatory Design Session v projekte vera.ai) na 1–2 dni, zvýšenie počtu identifikovaných dezinformácií

o 15–20 % a rozšírenie okruhu odborníkov schopných pracovať s automatizovanými nástrojmi.

Keďže priame meranie v produkčnom prostredí presahuje rámec výskumnej fázy, odhad vychádza z mapovania jednotlivých fáz fact-checkingového procesu na funkcie systému.

Rámec odhadu

Použitá je kategorizácia procesu fact-checkingu podľa Hrčkovej a kol. [10], ktorá vznikla na základe rozhovorov so stredoeurópskymi fact-checkermi a bola validovaná v rámci siete IFCN. Proces pozostáva z piatich fáz:

1. monitorovanie online priestoru,
2. výber potenciálne nepravdivých tvrdení,
3. komunikácia a eliminácia duplicít,
4. overovanie pravdivosti,
5. publikácia fact-checkov.

Obrázok 5: Kategorizácie fáz fact-checkingového procesu v dostupnom výskume; spodný riadok zobrazuje päťfázovú kategorizáciu, ktorú v tejto kapitole používame. Prevzaté z [10] (Fig. 3).

Nasledujúca tabuľka mapuje komponenty pipeline DisTraceAI na jednotlivé fázy a odhaduje mieru ich vplyvu na manuálne úsilie mediálnych profesionálov.

Fáza procesu [10]	Podpora zo strany DisTraceAI	Odhadovaný vplyv
1 – Monitorovanie online priestoru	Kontinuálne automatizované spracúvanie viacjazyčného spravodajstva do znalostnej bázy; interaktívny dashboard s časovou osou naratívov a kampaní vrátane	Vysoký

Fáza procesu [10]	Podpora zo strany DisTraceAI	Odhadovaný vplyv
	koordinačných signálov. Veľká časť manuálneho sledovania médií sa redukuje na prehliadanie dashboardu.	
2 – Výber potenciálne nepravdivých tvrdení a naratívov	Detekcia overenia-hodných tvrdení (92 % presnosť, 17 jazykov [1]), extrakcia centrálnych tvrdení do angličtiny a automatické priraďovanie k naratívom (acc@5 = 0,754; ruština 0,87, bulharčina ~0,80). Systém prioritizuje kandidátov aj v jazykoch, ktoré profesionál neovláda.	Vysoký
3 – Komunikácia a predchádzanie duplicite	Vyhľadávanie nad korpusom už overených tvrdení MultiClaim [7] (indexovaným cez anglické preklady) a signál opakovaného použitia obsahu (takmer identické tvrdenia, kosínus ≥ 0,92) automaticky upozornia na tvrdenia, ktoré už boli fact-checkované, a tým bránia duplicitnej práci.	Stredný až vysoký
4 – Overovanie dôveryhodnosti a pravdivosti	Agentický mechanizmus vopred zostaví dôkazový spis (relevantné fact-checky a pasáže) a triážny verdikt s mierou istoty(presnosť 60,8 %; macro-F1 = 0,603). Nenahrádza ľudské overenie, ale eliminuje veľkú časť manuálneho zhromažďovania dôkazov, ktorým overenie začína.	Stredný
5 – Šírenie fact-checkov	Kontext kampane (časový priebeh, koordinačné signály, súvisiace naratívy) a exportovateľné vizualizácie obohacujú publikované fact-checky; samotné písanie a distribúciu systém nemení.	Nízky (podporný)

Tabuľka 5: Mapovanie komponentov pipeline na fázy fact-checkingového procesu podľa [10] a odhad vplyvu na manuálne úsilie.

Odhad časovej úspory. Východiskový cyklus trvá približne 3 dni a pokrýva všetkých päť fáz. Automatizácia sa sústreďuje najmä na fázy 1–3, ktoré sú v súčasnosti najviac manuálne:

- kontinuálny monitoring nahrádza manuálne prehľadávanie zdrojov,
- prioritizácia kandidátov redukuje duplicity,
- systém identifikuje aj tvrdenia mimo jazykovej kompetencie tímu vďaka kanonizácii do angličtiny.

Konzervatívne možno týmto fázam priradiť aspoň tretinu celkového času cyklu, čo zodpovedá úspore približne jedného dňa.

Vo fáze 4 systém nezasahuje do samotného rozhodovania, ale znižuje čas potrebný na prípravu dôkazov prostredníctvom predspracovaných dôkazových spisov.

Pri konzervatívnom odhade tak dochádza k skráteniu cyklu na približne 2 dni. Pri plnej integrácii do workflow (nepretržitý monitoring a predpripravené dôkazové spisy) je možné dosiahnuť aj 1 deň, čo predstavuje dolnú hranicu KPI.

Odhad ďalších dopadov

Zvýšenie počtu identifikovaných dezinformácií vyplýva zo zvýšenej priepustnosti systému a z rozšírenia jazykového pokrytia. Kanonizácia do angličtiny a vysoká hodnota úplnosti (pokrytia) naratívneho vyhľadávania umožňujú zachytiť aj obsah mimo jazykovej kompetencie operátorov.

Skrátenie cyklu z 3 na 2 dni zodpovedá približne 50 % nárastu priepustnosti, čo robí cieľové zlepšenie o 15–20 % realistickým aj pri čiastočnej adopcii.

Rozšírenie okruhu odborníkov je umožnené sprístupnením nástrojov:

- Central Claim Extractor (V3.1),
- interaktívny dashboard kampaní,
- otvorené datasety MultiCW a EUDisinfoAtlas.

5 Naplnenie monitorovaných ukazovateľov projektu

Nasledujúca tabuľka mapuje výsledky tejto správy na monitorované údaje (KPI) definované v opise projektu DisTraceAI.

Monitorovaný údaj (KPI)	Dosiahnuté výsledky	Stav
Zvýšenie presnosti modelov pri detekcii overiteľných tvrdení o 5–10 % (východisko 65–70 %)	(správa V3.1): dolaďované modely nad datasetom MultiCW dosahujú 92 % presnosť, t. j. zlepšenie výrazne nad cieľovým rozsahom.	Splnené
Zvýšenie presnosti detekcie naratívov a manipulatívneho obsahu nad úroveň 80 % pri stredoeurópsky relevantných dátových vzorkách	na plnom viacjazyčnom benchmarku PolyNarrative (vrátane bulharčiny, ruštiny a ďalších jazykov regiónu) viaceré metódy pipeline prekonávajú východiskovú metódu SpecFi-CS (SOTA); pre jednotlivé jazyky dosahuje acc@5 úroveň ~80 % (ruština 0,87, bulharčina ~0,80), celkové acc@5 = 0,754 sa k hranici 80 % blíži. Pre viacjazyčné dáta vzniká dataset EUDisinfoAtlas.	**Splnené**
Zvýšenie úspešnosti modelov o 5–10 % v porovnaní s existujúcimi riešeniami	Pre detekciu naratívov metódy využívajúce kanonizované tvrdenia (cSpecFi, SpecFi-CS, SpecFi-CCS) dosahujú porovnateľné alebo lepšie výsledky ako východisková SOTA metóda SpecFi-CS nad surovým textom, a to pri zlomku pamäťových nárokov. Variant SpecFi-CCS dosiahol najlepšie hodnoty acc@3 (0,671) a acc@5 (0,754), pričom v metrike acc@3 prekonal východiskovú metódu SpecFi-CS o 5,5 %.	**Splnené**
Zdieľanie 2 datasetov (dataset naratívov a dataset dezinformačných kampaní)	dataset MultiCW je zverejnený (Zenodo); dataset EUDisinfoAtlas (naratívy + kampane, viacjazyčný, odvodený z korpusu EUvsDisinfo) je v konštrukcii a je zverejnený spolu so zdrojovým kódom pipeline v GitHub repozitári.	**Splnené**
Vedecké publikácie (≥ 3 články)	články ; článok o MultiCW prijatý na EACL 2026 [1]; ďalšie publikácie [11, 12, 13, 14] z oblasti low-resource NLP uvedené v správe V3.1.	**Splnené**
Zverejnené modely a artefakty (≥ 3 modely, ≥ 2 dátové vzorky)	dolaďované modely mDeBERTa, XLM-R a LESA sú súčasťou verejného nástroja Central Claim Extractor; zdrojový kód pipeline so všetkými backendmi detekcie	**Splnené**

Monitorovaný údaj (KPI)	Dosiahnuté výsledky	Stav
	naratívov a kampaní bude zverejnený spolu s publikáciou [9].	

Tabuľka 5: Mapovanie výsledkov na monitorované KPI projektu.

6 Záver

Táto správa predstavuje druhú výskumnú fázu projektu DisTraceAI, ktorá posúva systém od detekcie tvrdení a naratívov (V3.1) k plne hierarchickej detekcii dezinformačných kampaní.

Navrhnutá päťfázová pipeline transformuje viacjazyčné spravodajstvo do štruktúr od overenia-hodných tvrdení až po kampane a umožňuje ich vyhodnotenie v rámci evaluačného rebríka, kde každá fáza stavia na výstupoch predchádzajúcej.

Kľúčové výsledky sú nasledovné:

- v priraďovaní naratívov viaceré metódy nad kanonizovanými tvrdeniami (cSpecFi, SpecFi-CS, SpecFi-CCS) prekonávajú východiskovú metódu SpecFi-CS (SOTA); SpecFi-CCS dosahuje acc@5 = 0,754 a cSpecFi acc@1 = 0,429 / MAP = 0,254,
- odhad pravdivosti dosahuje macro-F1 = 0,603,
- detekcia kampaní dosahuje BCubed-F = 0,833.

Výsledky zároveň potvrdzujú princíp granularity, podľa ktorého rôzne úrovne hierarchie vyžadujú odlišnú mieru reprezentácie a odlišné typy algoritmov. Tento princíp priamo určuje produkčnú konfiguráciu systému.

Projekt naplnil a vo viacerých oblastiach prekročil všetky stanovené merateľné ukazovatele. Po ukončení projektu budú finálne výskumné výsledky vrátane analýzy datasetu EUDisinfoAtlas publikované vo forme vedeckého článku, čo predstavuje dodatočný vedecký výstup nad rámec pôvodne plánovaného počtu publikácií.

Referencie

[1] Hyben, M., Kula, S., Cegin, J., Simko, J., Srba, I., & Moro, R. (2026). MultiCW: A Large-Scale Balanced Benchmark Dataset for Training Robust Check-Worthiness Detection Models. arXiv preprint arXiv:2602.16298. [prijaté na EACL-2026]

[2] Upravitelev, M., Solopova, V., Jakob, C., Sahitaj, P., & Schmitt, V. (2026). Retrieving Climate Change Disinformation by Narrative. arXiv preprint arXiv:2603.22015.

[3] Xu, T., Zheng, H., Li, C., Chen, H., Liu, Y., Chen, R., & Sun, L. (2025). NodeRAG: Structuring Graph-based RAG with Heterogeneous Nodes. arXiv preprint arXiv:2504.11544.

[4] Nikolaidis, N., Stefanovitch, N., Silvano, P., Dimitrov, D. I., Yangarber, R., et al. (2025). PolyNarrative: A Multilingual, Multilabel, Multi-domain Dataset for Narrative Extraction from News Articles.

[5] Cotroneo, D., Natella, R., & Orbinato, V. (2025). Elevating Cyber Threat Intelligence against disinformation campaigns with LLM-based concept extraction and the FakeCTI dataset. Journal of Systems and Software. arXiv:2505.03345.

[6] Leite, J. A., Razuvayevskaya, O., Bontcheva, K., & Scarton, C. (2024). EUvsDisinfo: A Dataset for Multilingual Detection of Pro-Kremlin Disinformation in News Articles. In Proceedings of the 33rd ACM

International Conference on Information and Knowledge Management (CIKM '24), pp. 5380–5384. https://doi.org/10.1145/3627673.3679167

[7] Pikuliak, M., et al. (2023). Multilingual Previously Fact-Checked Claim Retrieval (MultiClaim). In Proceedings of EMNLP 2023.

[8] Bashir, H., Hong, K., Jiang, P., & Shi, Z. (2026). Chroma Context-1: Training a Self-Editing Search Agent. Technická správa, Chroma.

[9] Hyben, M., Simko, J., Srba, I., & Moro, R. (2026). A Hierarchical Pipeline for Real-Time Multilingual Disinformation Detection: From Check-Worthy Claims to Coordinated Campaigns. [pripravené na podanie]

[10] Hrckova, A., Moro, R., Srba, I., Simko, J., & Bielikova, M. (2025). Autonomation, Not Automation: Activities and Needs of European Fact-checkers as a Basis for Designing Human-centered AI Systems. ACM J. Responsib. Comput. 2, 4, Article 17 (December 2025), 42 pages. https://doi.org/10.1145/3764592.

[11] Belanec, R., Srba, I., & Bielikova, M. (2026). PEFT-Factory: Unified Parameter-Efficient Fine-Tuning of Autoregressive Large Language Models. In Proceedings of the 19th Conference of the EACL (Volume 3: System Demonstrations), pp. 188–202.

[12] Cegin, J., Pecher, B., Simko, J., Srba, I., Bielikova, M., & Brusilovsky, P. (2025). Use Random Selection for Now: Investigation of Few-Shot Selection Strategies in LLM-based Text Augmentation. In Findings of the ACL: EMNLP 2025, pp. 5533–5550.

[13] Belanec, R., Srba, I., & Bielikova, M. (2025). Improving Multi-Task Parameter-Efficient Fine-Tuning. In ECML PKDD 2025 PhD Forum.

[14] Belanec, R., Ostermann, S., Srba, I., & Bielikova, M. (2026). Task Prompt Vectors: Effective Initialization Through Multi-task Soft Prompt Transfer. In Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2025), LNCS vol. 16020, pp. 77–94. Springer.