V3.1 Výskumná správa o modeloch a metódach detekcie tvrdení a naratívov

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

Úvod

Práca mediálnych profesionálov, ako sú fact-checkeri a novinári, zahŕňa každodenné spracovanie veľkého objemu textu s cieľom identifikovať kľúčové informácie pre overenie. Obsah môže pochádzať z rôznych zdrojov – od dobre štruktúrovaných dokumentov (napr. novinové články, tlačové správy) až po vysoko neštruktúrované formáty (napr. príspevky na sociálnych sieťach, či prepisy rozhovorov). Táto úloha je zároveň komplikovaná jazykovou rozmanitosťou – mediálni profesionáli sa často stretávajú s obsahom v jazykoch, ktoré neovládajú. V tomto náročnom prostredí môžu automatizované systémy, ktoré identifikujú overenia-hodné informácie, výrazne znížiť manuálne úsilie a zlepšiť efektivitu ich práce. V rámci projektu sme sa zamerali na 3 navzájom nadväzujúce úlohy, úzko súvisiace s prácou mediálnych profesionálov.

V prvom rade sme riešili problém identifikácie overenia-hodných tvrdení (check-worthy claims, CW) – informačných jednotiek, ktoré sú dostatočne dôležité, overiteľné alebo impaktné na to, aby stáli za ďalšie preskúmanie. Napriek rastúcemu záujmu o túto úlohu zostávajú existujúce datasety pre detekciu CW tvrdení obmedzené: sú často obmedzené na angličtinu, úzke domény (napr. COVID-19) alebo špecifické štýly, ako sú novinové titulky. Tieto obmedzenia bránia schopnosti automatizovaných systémov generalizovať nad viacjazyčným, viac-doménovým alebo neformálnym obsahom. Na riešenie týchto nedostatkov predstavujeme Multi-Check-Worthy (MultiCW) dataset, na základe ktorého sme vytvorili niekoľko dolaďovaných modelov, ktoré, napriek výrazne menšiemu počtu parametrov, prekonávajú najnovšie veľké jazykové modely (LLM).

Následne sme sa v rámci výskumných úloh zamerali na mapovanie aktuálnych naratívov a identifikáciu dezinformačných kampaní, ktoré sú kľúčové pre pochopenie dynamiky šírenia informácií a detekciu manipulatívnych stratégií. Navrhli sme systém detekcie naratívov založený na analýze overenia-hodných tvrdení, ktorý zhlukuje tvrdenia podľa identifikovaných pomenovaných entít a ich sémantickej podobnosti. Takto vytvorená štruktúra umožňuje identifikovať koherentné naratívne rámce v analyzovanom korpuse.

Do budúcnosti plánujeme rozšíriť model o časovú dimenziu s cieľom sledovať vývoj naratívov a identifikovať koordinované online (dez)informačné kampane. Navrhnutý prístup slúži na podporu mediálnych profesionálov pri spracovaní veľkého objemu textu a zvyšuje ich schopnosť včas reagovať na škodlivé informačné aktivity.

Aby sme podporili úspešnosť klasifikačných metód, časť výskumného úsilia projektu DisTraceAI sme zamerali na dve oblasti low-resource NLP, v ktorých sme priniesli nové poznatky a metódy: augmentácia dát a metódy PEFT.

Augmentácia textových dát je technika v oblasti spracovania prirodzeného jazyka (NLP), pri ktorej sa z existujúcich textových údajov vytvárajú nové, mierne pozmenené verzie s cieľom rozšíriť tréningový korpus a zlepšiť schopnosť modelu generalizovať. V tejto téme sme priniesli dve štúdie, ktoré skúmajú potenciál generatívnych modelov obohacovať textové vzorky dát.

Ďalším naším cieľom prispieť k efektívnejšej a udržateľnej adaptácii LLM modelov prostredníctvom metód parameter-efficient fine-tuning (PEFT). Vďaka výskumu reprezentácie úloh v priestore soft promptov, návrhu jednotného benchmarku s metrikou zameranou na efektivitu a vývoja modulárneho open-source rámca sme vytvorili ucelený prístup, ktorý podporuje škálovateľný prenos znalostí a transparentné experimentovanie v oblasti PEFT metód.

2 Detekcia overenia-hodných tvrdení - MultiCW dataset

V rámci prvej úlohy sme vytvorili rozsiahly, vyvážený a viacjazyčný benchmarkový dataset MultiCW, na detekciu overenia-hodných tvrdení (check-worthy claims, CW), ktorý rieši nedostatky existujúcich datasetov – ich obmedzené jazykové a tematické pokrytie, nevyvážené triedy a neštruktúrované formálne štýly. Cieľom bolo vytvoriť dataset pre trénovanie modelov schopných generalizovať naprieč čo najväčším spektrom parametrov vstupného textu. Preto dataset obsahuje 16 jazykov, 7 tématických domén a 2 formálne štýly - formálny a neformálny. Súčasťou datasetu je aj out-of-distribution dataset (OOD), obsahujúci 4 ďalšie jazyky a veľmi podobnú kompozíciu tém a štýlov ako in-distribution dataset, ktorý slúži na overenie schopnosti dolaďovaných modelov generalizovať nad jazykmi mimo trénovacej množiny.

Na základe tohto datasetu sme dolaďovali 3 modely na báze Transformerov (XLM-R, mDeBERTa, LESA) a ich výkonnosť sme porovnali s 15 najmodernejšími veľkými jazykovými modelmi (LLMs) v náročnom benchmarkovom experimente, ktorý umožnil získať štandardizované a porovnateľné výsledky.

Dataset spolu s výsledkami benchmarkovej štúdie bude prezentovaný na konferencii EACL-2026. Preprint článku a zdrojový kód sú dostupné v práci *[1]*. Samotný dataset je verejne prístupný na Zenodo: *https://zenodo.org/records/17482958*.

2.1. Definícia overenia-hodných tvrdení

Pojem toho, čo predstavuje overenia-hodné tvrdenie (CW), nie je vždy jednoznačný. Pre účely konzistentnej anotácie a konštrukcie datasetu sme definovali na základe nasledovných pravidiel.

Tvrdenie je overenia-hodné, ak spĺňa jedno alebo viac z nasledujúcich kritérií:

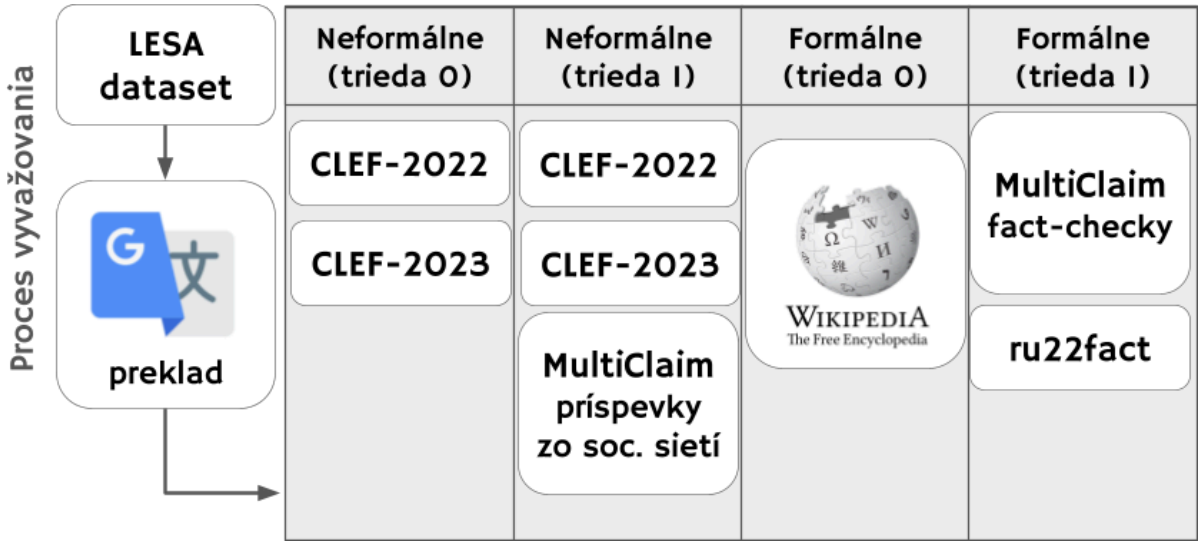
1. ***Významnosť*:** tvrdenie má dôsledky pre verejnú politiku, zdravie, bezpečnosť alebo spoločenský blahobyt.
2. ***Kontroverznosť*:** tvrdenie je sporné alebo pravdepodobne vyvolá verejnú alebo odbornú diskusiu.
3. ***Dopad*:** tvrdenie môže ovplyvniť verejnú mienku, formovať rozhodnutia alebo meniť správanie.
4. ***Spoľahlivosť zdroja*:** tvrdenie pochádza od verejnej osobnosti, orgánu alebo inštitúcie s veľkým verejným vplyvom.

Naopak, **tvrdenie sa nepovažuje za overenia-hodné (nevýznamné)**, ak patrí do aspoň jednej z nasledujúcich kategórií:

1. ***Subjektivita*:** čisto subjektívne vyjadrenia názoru alebo vkusu.
2. ***Trivialita*:** bezvýznamné tvrdenia bez širších dôsledkov.
3. ***Všeobecne známe informácie*:** všeobecne akceptované fakty, ktoré neprinášajú nové ani sporné informácie.
4. ***Nedostatočný dopad*:** tvrdenia s minimálnym vplyvom, alebo zanedbateľnými následkami, ak sú nepravdivé.

2.2 Konštrukcia datasetu MultiCW

Dataset **MultiCW** nie je jednoduchým agregátom existujúcich datasetov, ale starostlivo konštruovanou kolekciou určenou na trénovanie robustných viacjazyčných modelov detekcie overenia-hodných tvrdení (CW). Integruje viaceré existujúce datasety, medzi ktoré patria: *CLEF-2022* a *CLEF-2023* obsahujúce anotované neformálne tvrdenia v šiestich jazykoch; *MultiClaim* a jeho rozšírená verzia *MultiClaim v2* obsahujú formálne fact-checkované články, ako aj neformálne tvrdeniami zo sociálnych médií v 39 jazykoch; *Ru22Fact* – formálny viacjazyčný fact-checkingový dataset o rusko-ukrajinskej vojne z roku 2022. Kompiláciou týchto datasetov sme získali 63 936 tvrdení naprieč 7 tématickými doménami - zdravotníctvo, politika, životné prostredie, veda, šport, zábava a história.



Obrázok č. 1: Konštrukcia MultiCW a OOD datasetu.

Takto zostavený dataset však neobsahoval žiadne štrukturované, nevýznamné tvrdenia (trieda 0). Tieto tvrdenia sme získali z Wikipédie, pretože články z Wikipédie predstavujú formálne štruktúrovaný zdroj, ktorý spĺňa našu definíciu *všeobecne známych informácií*, a preto nie je potrebné ich overovať. Články z Wikipédie neboli vybrané náhodne, ale na základe menných entít, detegovaných z formálnej časti datasetu. Takýmto spôsobom sme získali 31 093 tvrdení, čím sme získali vyváženú formálnu časť datasetu.

Finálne vyváženie datasetu bolo následne vykonané s použitím *datasetu LESA*, ktorý obsahuje overenia-hodné tvrdenia výhradne v anglickom jazyku a v 3 formálnych štýloch - formálny, kvázi-formálny a neformálny. Prekladom vzoriek tohto datasetu do cieľového jazyka a použitím vzoriek s vhodným formálnym štýlom sme docielili vyvážený dataset, ktorý zahŕňa 16 jazykov - jazyky z nízkym a vysokým množstvom trénovacích dát, so širokého spektra jazykových skupín - afro-ázijská, indo-európska, turkická, sino-tibetská a austronézska.

	Jazyk	Neformálne (trieda 0)	Neformálne (trieda 1)	Formálne (trieda 0)	Formálne (trieda 1)
in-distribution	Arabic	2000	1993	2000	2000
	Bulgarian	2000	1993	1093	1093
	Brunei	1964	1953	2000	2000
	Czech	1961	1992	2000	2000
	German	1959	2000	2000	2000
	English	2000	2000	2000	2000
	Spanish	2000	1997	2000	2000
	French	1962	2000	2000	2000
	Hindi	1968	2000	2000	2000
	Polish	1963	2000	2000	2000
	Portugal	1961	2000	2000	2000
	Russian	1969	1997	2000	2000
	Slovak	1962	1986	2000	2000
	Turkish	2000	1898	2000	2000
	Ukrainian	1956	1986	2000	2000
	Chinese	1960	1984	2000	2000
Out-of-distribution	Italian	2000	1482	2000	2000
	Macedonian	1999	1385	1123	1123
	Malaysian	1999	1340	1297	1297
	Netherlands	1999	1910	1227	1227

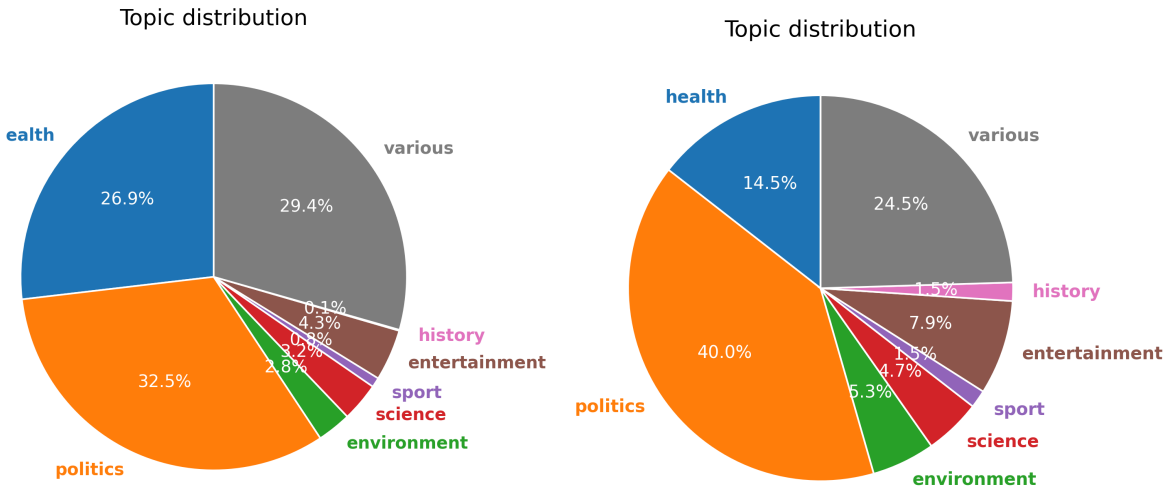
Tabuľka 1: Štatistiky pre in-distribution časť (hore) a out-of-distribution časť (dole) vyváženého datasetu MultiCW.

V rámci kompilácie sme brali sme do úvahy iba jazyky, ktoré v každej kategórii (obe neformálne triedy a formálna trieda 1) obsahovali aspoň 1500 vzoriek, výsledkom čoho bolo spomínaných 16 jazykov v MultiCW datasete. Znížením prahovej hodnoty na 1000 vzoriek na kategóriu sme získali 4 ďalšie jazyky, ktoré síce obsahovali menší počet vzoriek, avšak po ich doplnení a vyvážení rovnakým spôsobom, ako vyššie opísaný MultiCW dataset, predstavuje hodnotný out-of-distribution dataset (OOD). Takýto dataset je vhodný na overenie schopnosti dolaďovaných modelov generalizovať nad jazykmi mimo trénovacej množiny.

2.3 Tématické rozdelenie datasetu

Vo finálnej verzii oboch datasetov sme vykonali extrakciu tém pomocou modelu Llama 3 - 4b. Ako vidíme na grafoch na obrázku 2, témy *zdravie* a *politika* sú v oboch datasetoch dominantnými témami. Na druhej strane, témy história a šport nie sú v datasetoch dostatočne zastúpené, čo by veľmi sťažilo vyváženie tém. Navyše veľké množstvo vzoriek

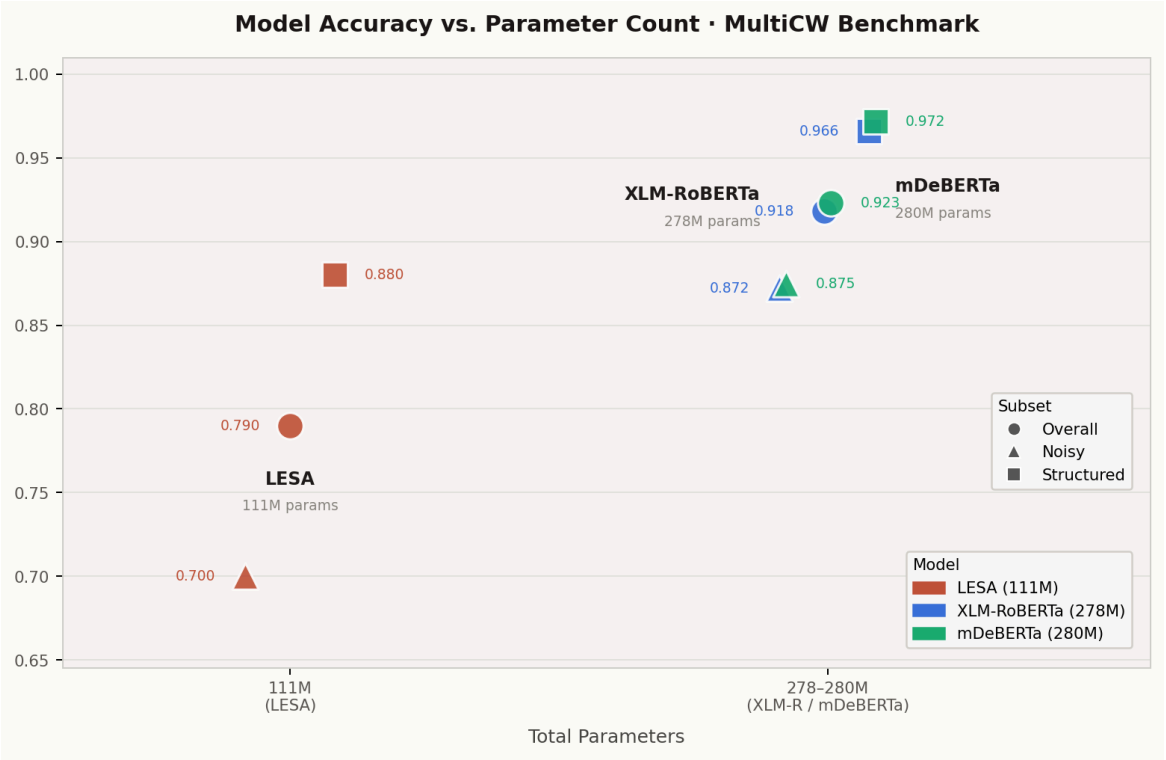
nebolo možné kategorizovať, keďže obsahovali buď nejednoznačné témy, alebo jedna vzorka obsahovala viac tém súčasne - tieto vzorky sme zahrnuli do kategórie “rôzne” (“various”). Toto všetko sú dôvody, prečo sme upustili od snahy náš dataset tématicky vyvážiť a považujeme to za možnosti na budúce vylepšenie datasetu.



Obrázok 2: Rozloženie tém v datasete MultiCW (vľavo) a v out-of-distribution datasete (OOD) (vpravo).

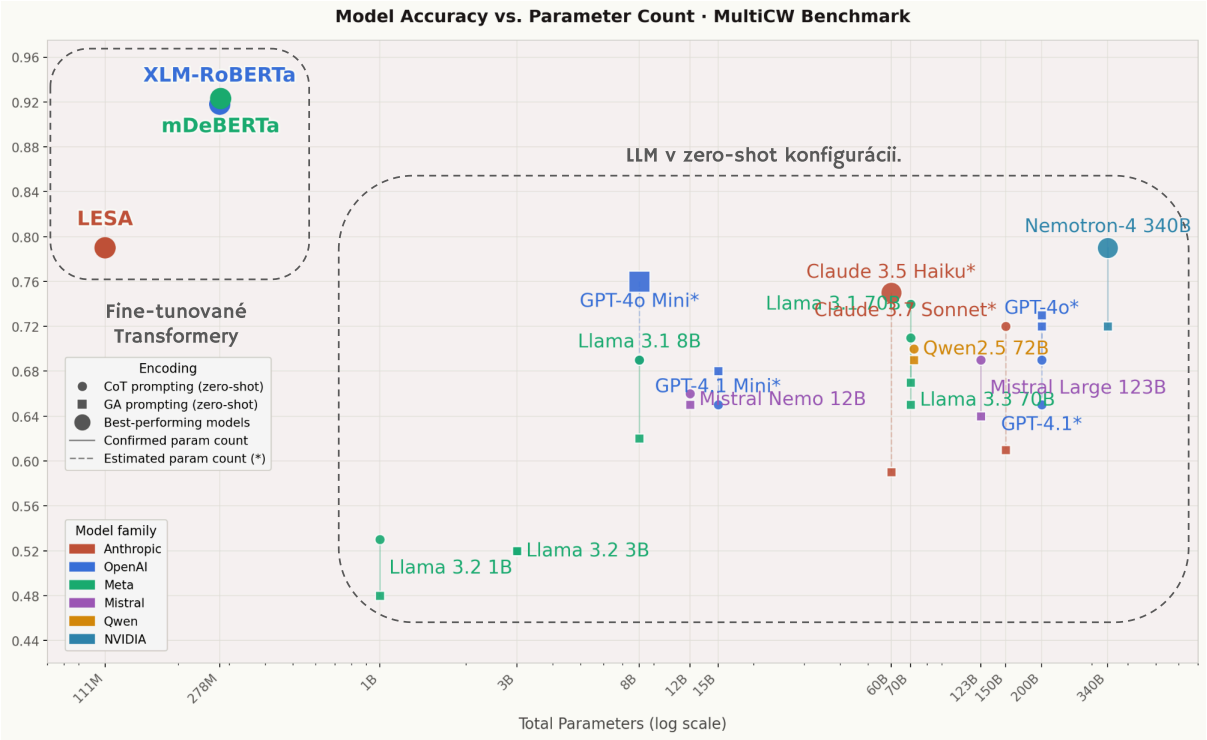
2.4 Komplexný benchmark modelov

V tejto časti popíšeme porovnanie troch dolaďovaných modelov, s komerčným a open-source LLM v zero-shot konfigurácii nad MultiCW datasetom. Na dolaďovanie sme vybrali tri moderné modely založené na báze transformerov, menovite mDeBERTA, XLM-RoBERTa a LESA. Najúspešnejším modelom bol mDeBERTA s celkovou presnosťou 92%, pričom XLM-R dosiahol podobné výsledky. LESA, aj napriek tomu, že sa jedná o staršiu architektúru, stále dobre generalizuje so 79% presnosťou. Ako vidíme na obrázku 3, všetky modely dosiahli podobné výsledky v rôznych štýloch písania – všetky tri mali problémy s tvrdeniami v neformálnom štýle. Vo všetkých modeloch sa chyby prejavujú najmä ako falošné negatívy pri neformálnych tvrdeniach s implicitným, alebo sarkastickým jazykom a ako falošné pozitívy pri štruktúrovaných tvrdeniach, ktoré uvádzajú triviálne fakty.



Obrázok 3: Presnosť verzus veľkosť modelov dolaďovaných na datasete MultiCW.

V porovnaní s 15 rôznymi, komerčnými aj open-source veľkými jazykovými modelmi, ktoré boli promptované rôznymi verziami promptov, môžeme konštatovať, že najlepší zero-shot LLM – konkrétne Nemotron-4, dosiahol približne rovnakú presnosť ako najmenej presný dolaďovaný transformer – konkrétne LESA, a to aj napriek tomu, že dolaďované modely boli výrazne menšie z hľadiska počtu parametrov. Dolaďované modely dosiahli presnosť približne 92%, čím prekonali všetky zero-shot LLM o 13 – 17 bodov.



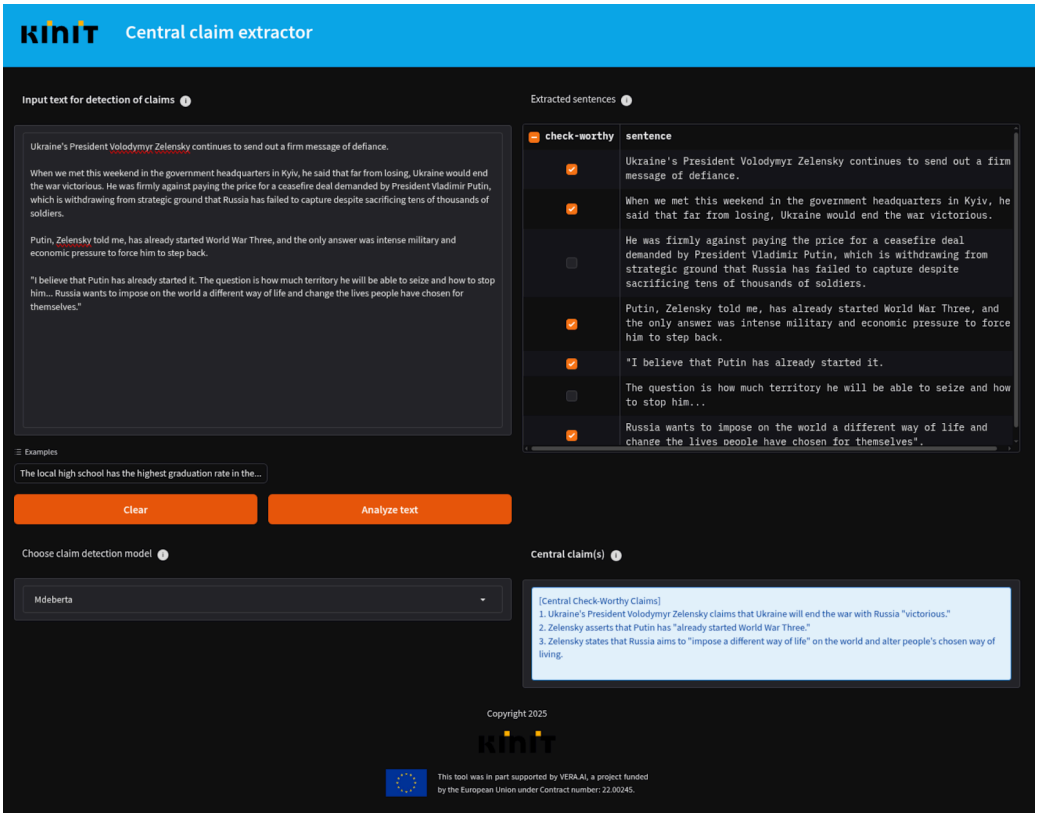
Obrázok 4: Presnosť verzus veľkosť modelov dolaďovaných na MultiCW datasete, v porovnaní s LLM v zero-shot konfigurácii.

Použilo sa aj viacero návrhov promptov, *chain-of-thought prompt* konzistentne prekonal ostatné návrhy promptov.

Dolaďované modely sme tiež vyhodnotili na našom out-of-distribution datasete (OOD), na ktorom sa preukázalo, že modely dobre generalizujú aj naprieč neznámymi jazykmi, čo potvrdzuje, že adaptácia modelov špecificky pre túto úlohu zostáva zásadná.

2.5 Online nástroje

Dva najpresnejšie modely z dolaďovaných modelov - konkrétne mDeBERTa a XLM-R, boli publikované, ako súčasť verejne dostupného online nástroja *Central Claim Extractor*, ktorý umožňuje detekciu overenia-hodných tvrdení, a na základe nich aj extrakciu centrálnych tvrdení z ľubovoľného vstupného textu. Jeho používaním a kontrolou výsledkov nášho nástroja nám používatelia poskytujú cennú spätnú väzbu, ktorá nám pomôže v budúcnosti vylepšiť náš dataset. Na obrázku 5 je snímka obrazovky zobrazujúca použitie tohto online nástroja v praxi.



Obrázok 5: Central Claim Extractor - online nástroj publikovaný KInIT-om.

3 Detekcia naratívov

V rámci prvej úlohy sme navrhli a implementovali viacstupňový hierarchický proces detekcie naratívov aplikovaný na viacjazyčný spravodajský korpus Media-Content Library a dataset fact-checkingových článkov MultiClaim. Navrhnutý proces je doménovo nezávislý a umožňuje spracovanie ľubovoľného textového vstupu.

Analytická pipeline pozostáva z nasledujúcich NLP komponentov:

1. Detekcia overenia-hodných tvrdení
2. Dekontextualizácia tvrdení
3. Rozpoznávanie pomenovaných entít
4. Tématické zhlukovanie tvrdení
5. Hierarchické zhlukovanie naratívov

Cieľom procesu je extrahovať a zoskupovať naratívy na základe ich sémantickej a štrukturálnej podobnosti.

Navrhnutý prístup využíva doménovo špecifické transformerové modely mDeBERTa, Qwen3 a GLiNER, ako aj moderné klastrovacie algoritmy BERTopic a DBSCAN. Výsledkom je hierarchická štruktúra naratívnych zhlukov extrahovaných z rozsiahlych textových kolekcií.

3.1 Definícia naratívu

Pojem naratív je v literatúre definovaný rôzne v závislosti od kontextu a analytického cieľa. V tejto práci vychádzame z definície podľa [2], podľa ktorej:

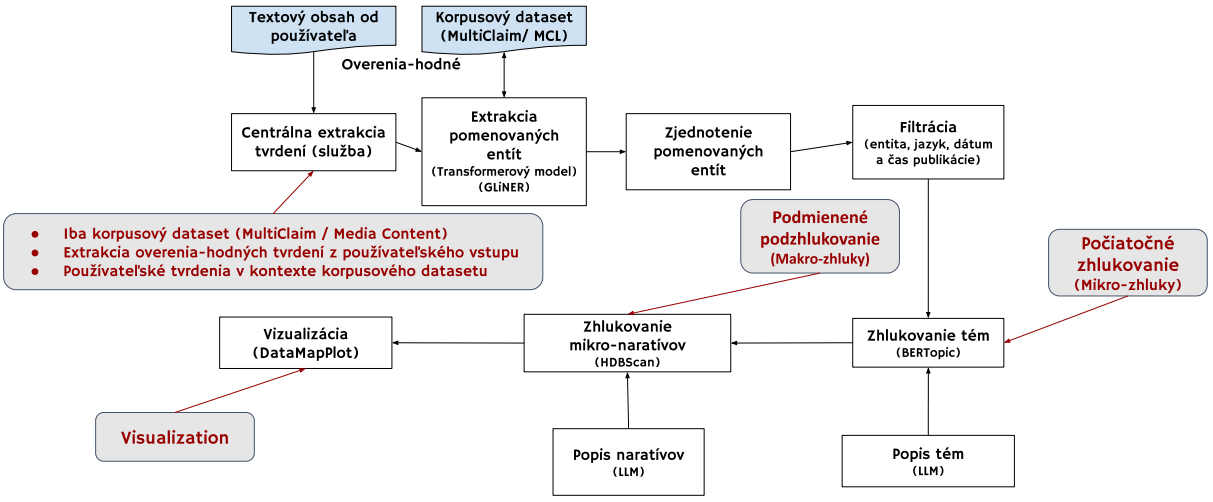
> *"Naratív možno chápať ako postupnosť nenáhodne prepojených udalostí, postupnosť vzájomne prepojených faktov, ktoré sú pozorované počas určitého časového obdobia a zahŕňajú základné prvky, ako sú organizácie, osoby, miesta alebo čas."*

Táto definícia umožňuje chápať naratív ako vyššiu informačnú jednotku, ktorá vzniká agregáciou tematicky a významovo previazaných tvrdení.

3.2 Metodológia

Navrhnutý proces využíva hierarchický, viacfázový prístup k extrakcii a zoskupovaniu naratívov na základe sémantickej a štrukturálnej podobnosti. Architektúra systému je modulárna a škálovateľná, pričom jednotlivé komponenty sú navrhnuté ako nezávislé moduly.

Na detekciu overenia-hodných tvrdení používame dolaďovaný model mDeBERTa (pozri kapitolu 2.4). Následne aplikujeme model GLiNER na rozpoznávanie pomenovaných entít. Tematická štruktúra je modelovaná pomocou BERTopic a finálna hierarchia naratívov je vytváraná algoritmom HDBSCAN.



Obrázok 6: Blokový diagram procesu hierarchickej detekcie naratívov.

3.2.1 Detekcia overenia-hodných tvrdení

Ako sme spomenuli v úvode, overiteľné tvrdenia tvoria základnú informačnú jednotku, na ktorej sa buduje vyššia sémantická štruktúra, ako sú témy a naratívy. Vstupné viacjazyčné spravodajské články sú najprv rozdelené na jednotlivé vety. Každá veta je následne spracovaná dolaďovaným modelom mDeBERTa (kapitola 2.4), ktorý vykonáva binárnu

klasifikáciu s cieľom identifikovať tvrdenia vhodné na ďalšie spracovanie. Táto fáza umožňuje redukciu šumu a filtrovanie nerelevantného obsahu, čím sa ďalšia analýza sústreďuje výhradne na štruktúrované a informačne významné jednotky.

3.2.2 Extrakcia pomenovaných entít

V druhej fáze prebieha extrakcia pomenovaných entít pomocou modelu *GLiNER*, moderného viacjazyčného transformerového modelu pre rozpoznávanie entít. Entity sú extrahované výlučne z identifikovaných overenia-hodných tvrdení, čo umožňuje jednoznačné prepojenie naratívnych zhlukov s konkrétnymi aktérmi (osoby, organizácie, lokality a pod.).

Následne prebieha zjednocovanie ekvivalentných entít (entity matching), napríklad mapovanie variantov *„Trump“* a *„Donald Trump“* na jednotnú kanonickú reprezentáciu. Na identifikáciu podobnosti používame Levenshteinovu vzdialenosť, pričom dve entity sú zlúčené, ak ich normalizovaná vzdialenosť prekročí prahovú hodnotu *0,2*.

Tento krok je kľúčový pre elimináciu fragmentácie tvrdení a prevenciu vzniku duplicitných naratívnych zhlukov.

3.2.3 Zhlukovanie do tématických klastrov

Tematické zhlukovanie je realizované pomocou modelu BERTopic, ktorý využíva BERT embeddingy na modelovanie sémantickej podobnosti medzi tvrdeniami. Výstupom je množina tematických klastrov obsahujúcich sémanticky príbuzné tvrdenia spolu s ich pridruženými entitami.

Na generovanie ľudsky čitateľných opisov tematických klastrov používame veľký jazykový model Llama 3 8B. Model ako vstup dostáva:

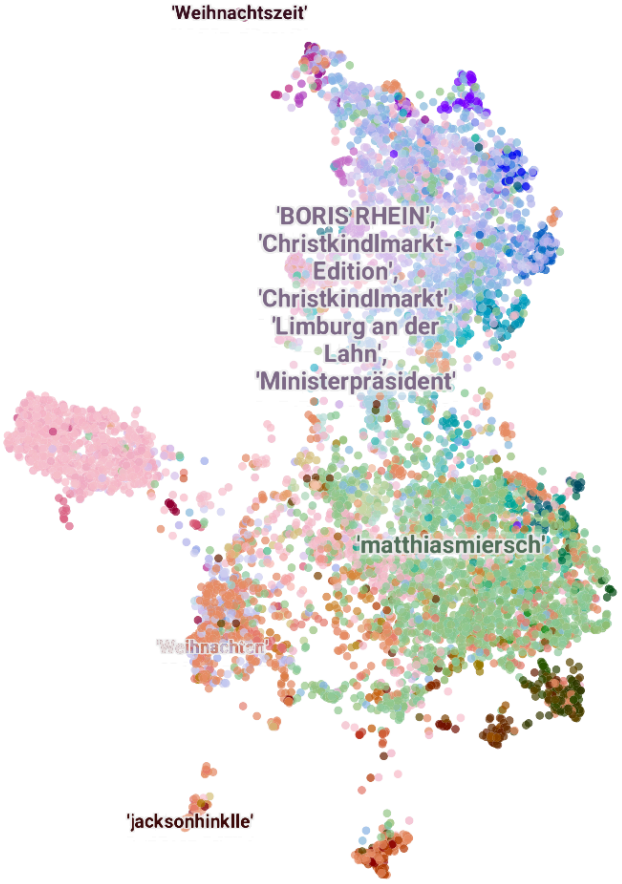
1. Množinu tvrdení patriacich do tematického zhluku,
2. Zoznam extrahovaných entít.

Na základe týchto údajov model generuje stručný a koherentný opis vystihujúci hlavný významový rámec danej témy.

3.2.4 Zhlukovanie do hierarchie naratívov

Finálna fáza spracovania spočíva v hierarchickom zoskupovaní tematických klastrov do naratívnych zhlukov pomocou algoritmu HDBSCAN, ktorý je založený na princípe hustoty dát a umožňuje identifikáciu zhlukov rôznej veľkosti a tvaru pri súčasnej robustnosti voči šumu.

Tematické klastre sú agregované na základe ich sémantickej podobnosti do koherentných naratívnych jednotiek.



Obrázok 7: Detegovaná hierarchia naratívov nad viacjazyčným spravodajským korpusom Media-Content Library.

Na generovanie sumarizačných opisov naratívnych zhlukov používame model *Gemma 3 12B*, ktorý ako vstup dostáva:

1. Popisy tematických klastrov
2. Agregovaný zoznam všetkých entít identifikovaných v danom naratíve

Model následne syntetizuje zastrešujúci opis vystihujúci hlavný významový rámec príslušného naratívu.

Výsledná hierarchická štruktúra je vizualizovaná pomocou knižnice *DataMapPlot*, ktorá umožňuje interaktívne prehliadanie dát naprieč rôznymi úrovňami abstrakcie – od globálnych naratívnych rámcov až po jednotlivé tvrdenia (pozri Obrázok 7).

3.3 Diskusia a budúce smerovanie

Navrhnutý prístup je explicitne orientovaný na spracovanie viacjazyčného spravodajského obsahu, čo predstavuje kľúčovú požiadavku pre reálne aplikačné scenáre.

Použitie viacjazyčných modelov mDeBERTa a GLiNER zabezpečuje jazykovú robustnosť systému. Spracovanie na úrovni viet umožňuje granulárnu extrakciu tvrdení nezávisle od jazyka, či štylistiky zdrojového textu.

Z hľadiska škálovateľnosti je riešenie vhodné pre:

- analýzu jednotlivých článkov
- spracovanie rozsiahlych korpusov
- priebežné monitorovanie informačného priestoru

Algoritmy *BERTopic* a *HDBSCAN* sú optimalizované pre prácu s rozsiahlymi dátovými kolekciami a umožňujú efektívne klastrovanie pri zachovaní vysokej kvality zhlukov. Modulárna architektúra systému umožňuje flexibilnú zmenu modelov, alebo algoritmických komponentov bez narušenia integrity celého systému.

V ďalšej fáze plánujeme rozšíriť systém o budovanie znalostnej bázy (knowledge base), ktorá bude okrem samotných naratívnych zhlukov obsahovať aj meta-informácie, ako napríklad:

- časová pečiatka
- kanál, z ktorého informácia pochádza
- autor článku
- typ média

Na základe týchto údajov bude možné modelovať:

- časový vývoj naratívov
- ich súbežný výskyt v rovnakých informačných kanáloch
- vzájomné prepojenia medzi naratívmi
- agregáciu súbežných naratívov do komplexných informačných kampaní

Takéto rozšírenie umožní systematické sledovanie dynamiky informačného priestoru a identifikáciu koordinovaných komunikačných vzorcov.

4 NLP metódy s malým množstvom zdrojov

Použiteľnosť metód klasifikácie tvrdení, naratívov a kampaní je potrebná aj v menších jazykoch. Takzvané *metódy spracovania prirodzeného jazyka s malým množstvom zdrojov* (low-resource NLP) sú nevyhnutné pre jazyky ako je slovenčina, pretože rozsah dostupných anotovaných dát, predtrénovaných modelov a výpočtových zdrojov je výrazne menší než pri vysoko zastúpených jazykoch, ako sú angličtina alebo čínština. Používateľská základňa a digitálna stopa malých jazykov sú obmedzené, čo vedie k menšiemu množstvu rozsiahlych korpusov, benchmarkov a doménovo špecifických dátových súborov. Táto situácia si vyžaduje použitie metód transferového učenia, viacjazyčné modelovanie, medzi-jazykové reprezentácie či augmentáciu dát, ktoré umožňujú vyvíjať robustné jazykové technológie aj pri vyššie uvedených obmedzeniach.

Aby sme podporili úspešnosť klasifikačných metód, časť výskumného úsilia projektu DisTraceAI sme zamerali na dve oblasti low-resource NLP, v ktorých sme priniesli nové poznatky a metódy:

1. **Augmentácia dát.** Augmentácia textových dát je technika v oblasti spracovania prirodzeného jazyka (NLP), pri ktorej sa z existujúcich textových údajov vytvárajú nové, mierne pozmenené verzie s cieľom rozšíriť tréningový korpus a zlepšiť schopnosť modelu generalizovať. Keďže jazykové modely, napríklad architektúry typu BERT alebo GPT, vyžadujú veľké množstvo dát, augmentácia pomáha najmä v prípade nedostatku dát. Medzi bežné metódy patrí nahrádzanie slov synonymami, parafrázovanie viet, náhodné vkladanie či vymazávanie slov, spätný preklad (preklad do iného jazyka a späť) alebo generovanie textu pomocou jazykového modelu. Cieľom je zachovať pôvodný význam textu, no zároveň zvýšiť variabilitu jazykových prejavov, čím sa znižuje riziko pretrénovania a zvyšuje robustnosť modelu voči rôznym formuláciám.
2. **Metódy PEFT** (Parameter-Efficient Fine-Tuning) je prístup k ladeniu veľkých predtrénovaných jazykových modelov, pri ktorom sa namiesto aktualizácie všetkých parametrov modelu trénuje iba malá, špecificky vybraná podmnožina parametrov alebo sa pridávajú nové, malé adaptačné vrstvy. Cieľom je výrazne znížiť pamäťové a výpočtové nároky pri zachovaní vysokej výkonnosti modelu na konkrétnej úlohe. Tento prístup sa často používa pri veľkých modeloch, kde by plné dolaďovanie všetkých parametrov bolo nákladné a neefektívne. Výhodou PEFT je, že umožňuje rýchlejšie experimentovanie, jednoduchšie nasadzovanie modelov a efektívnejšie prispôsobenie veľkých modelov pre špecifické domény či scenáre s nedostatkom dát.

4.1 Príspevky v oblasti augmentácie dát

Náš predchádzajúci výskum v oblasti augmentácie textových dát ukázal veľký potenciál využitia veľkých jazykových modelov pri úlohe parafrázovania a ďalších augmentačných technikách. Napriek prísľubu však stále zostávala otvorená otázka stability kvality výstupov a tiež ceny takejto augmentácie.

Preto sme sa v našom výskume [3], v rámci DisTraceAI, zamerali na porovnanie generatívnych veľkých jazykových modelov (LLM) s tradičnými metódami textovej augmentácie pri úlohách klasifikácie textu. Vychádzali sme z pozorovania, že LLM augmentácie, ktoré generujú nové tréningové vzorky napríklad parafrázovaním, prinášajú v literatúre nekonzistentné výsledky – niekedy zlepšujú presnosť klasifikátorov, inokedy len minimálne, alebo ju dokonca zhoršujú. Identifikovali sme preto potrebu systematicky preskúmať nielen ich výkon, ale aj nákladovosť, a odpovedať na otázku, kedy a či je použitie LLM augmentácie efektívnejšie než etablované techniky.

Realizovali sme rozsiahlu experimentálnu štúdiu na šiestich datasetoch pokrývajúcich úlohy ako analýza sentimentu, klasifikácia správ a klasifikácia zámeru. Testovali sme tri rôzne klasifikátory a dva režimy ladenia (plné ladenie a LoRA). Porovnávali sme LLM augmentácie s tradičnými postupmi, ako je spätný preklad, či úpravy slov pomocou embeddingov.

Experimenty zahŕňali rôzne počty seed vzoriek a augmentovaných dát, pričom sme analyzovali nielen presnosť modelov, ale aj pomer výkonu k finančným nákladom, výpočtovej záťaži a produkcii CO_2.

Zistili sme, že LLM augmentácie prinášajú výraznejší prínos najmä v situáciách s veľmi malým počtom seed vzoriek (napríklad 5–20 na triedu), kde dokážu zlepšiť presnosť klasifikácie. S rastúcim množstvom pôvodných dát sa ich výhoda postupne stráca a tradičné metódy často dosahujú porovnateľné alebo lepšie výsledky pri výrazne nižších nákladoch. Na základe týchto zistení odporúčame používať LLM augmentácie predovšetkým v scenároch s obmedzeným množstvom tréningových dát, kde ich prínos môže prevýšiť ich nákladovosť.

Augmentácia textových dát je principiálne metódou in-context learningu, teda generovania nových textov na základe poskytovania pôvodných príkladov. Správny výber príkladov je kľúčový, pretože veľký jazykový model sa pri generovaní odpovede neopiera o aktualizáciu váh, ale výlučne o informácie obsiahnuté v poskytnutom kontexte. Vybrané ukážky teda implicitne definujú formát úlohy, rozdelenie tried, jazykový štýl aj rozhodovacie hranice, ktoré si model počas inferencie internalizuje. Reprezentatívne, informatívne a dostatočne rôznorodé príklady môžu zlepšiť generalizáciu a stabilitu výstupu, zatiaľ čo nevhodne zvolené, alebo skreslené ukážky môžu viesť k systematickým chybám, zvýšenej variabilite odpovedí či k posilneniu nežiaduceho biasu. Výber príkladov preto priamo ovplyvňuje kvalitu generovaných dát a neskôr klasifikačných modelov na nich trénovaných.

V existujúcich prácach augmentácie textových dát sme sa stretli s viacerými technikami výberu príkladov, no so žiadnym systematickým porovnaním, ktorá z nich je najvhodnejšia (všeobecne či v danom kontexte).

Preto sme sa v ďalšom experimente [4] skúmali, do akej miery ovplyvňuje stratégia výberu ukážkových príkladov kvalitu generovaných dát a následný výkon modelov. Hoci sa v praxi najčastejšie používa náhodný výber príkladov, nebolo dopredu zrejmé, či nebudú lepšou voľbou stratégie informovanými metódami výberu.

Experimentálne sme porovnali osem rôznych stratégií výberu príkladov prevzatých z literatúry o few-shot (in-context) učení, spolu s dvoma referenčnými prístupmi: zero-shot generovaním a náhodným výberom. Hodnotenie sme realizovali naprieč viacerými veľkými jazykovými modelmi a rôznymi klasifikačnými úlohami, pričom sme generovali augmentované dáta a následne na nich trénovali downstream model. Výkon sme merali na dátach z rovnakého rozdelenia (in-distribution), ako aj na dátach z odlišného rozdelenia (out-of-distribution), aby sme posúdili robustnosť jednotlivých stratégií.

Naše výsledky ukazujú, že žiadna z pokročilých stratégií výberu príkladov nepreukázala konzistentnú prevahu nad náhodným výberom v in-distribution nastaveniach. Náhodný výber pritom často dosahoval porovnateľné, alebo najlepšie výsledky, bez dodatočných výpočtových nákladov. V prípade out-of-distribution scenárov sa určité stratégie založené na nepodobnosti ukázali ako čiastočne prínosné, avšak ich prínos nebol univerzálny. Na základe našich zistení preto odporúčame používať náhodný výber ako silnú a efektívnu východiskovú stratégiu pri LLM-based augmentácii dát.

4.2 Príspevky v oblasti PEFT

Dynamický rast veľkých jazykových modelov (LLM) priniesol zásadné zlepšenia vo výkone na širokom spektre úloh spracovania prirodzeného jazyka, no zároveň výrazne zvýšil nároky na výpočtové zdroje, pamäť a experimentálnu infraštruktúru. Plné doladenie všetkých parametrov modelu sa preto v mnohých scenároch stáva prakticky neudržateľným, najmä v akademickom prostredí, alebo pri opakovaných experimentoch vyžadujúcich reprodukovateľnosť. Tieto výzvy viedli k rozvoju PEFT (Parameter-Efficient Fine-Tuning) metód, ktorých cieľom je adaptovať veľké modely na nové úlohy prostredníctvom aktualizácie iba malej podmnožiny parametrov, prípadne zavedením nových malých vrstiev, alebo transformáciou parametrov do menších dimenzii.

V tejto sérii prác pristupujeme k problematike PEFT z troch komplementárnych perspektív: 1) skúmame reprezentáciu a prenos úlohovej informácie v priestore soft promptov prostredníctvom konceptu "task prompt vectors"; 2) navrhujeme jednotný a nákladovo orientovaný benchmark (PEFT-Bench s metrikou PSCP) pre systematické hodnotenie metód; 3) a zároveň vyvíjame modulárny open-source rámec PEFT-Factory, ktorý podporuje transparentné a reprodukovateľné experimentovanie s PEFT prístupmi.

V prvom experimente [5] sme sa zamerali na zvýšenie efektivity prompt tuningu pre veľké jazykové modely (LLM), pri ktorom sa namiesto úpravy všetkých parametrov modelu učí iba krátky soft prompt. Identifikovali sme obmedzenia existujúcich prístupov, najmä ich slabšiu prenositeľnosť medzi viacerými úlohami. Navrhli sme preto koncept „*task prompt vectors*", ktorý definujeme ako rozdiel medzi natrénovaným soft promptom a jeho pôvodnou náhodnou inicializáciou. Tento rozdielový vektor interpretujeme ako smer reprezentujúci špecifické charakteristiky danej úlohy v priestore parametrov, pričom ho môžeme využiť na efektívnejšiu inicializáciu prompt tuningu pre nové, príbuzné úlohy.

V rámci experimentálnej časti sme analyzovali vlastnosti task prompt vectors na 19 datasetoch pokrývajúcich rôzne NLP úlohy vrátane klasifikácie a generovania textu. Zistili sme, že tieto vektory sú do značnej miery nezávislé od konkrétnej náhodnej inicializácie, čo naznačuje, že stabilne zachytávajú informáciu o úlohe. Zároveň sme ukázali, že s nimi možno vykonávať aritmetické operácie, napríklad ich sčítanie, čím umožňujeme kombinovať znalosti z viacerých úloh bez potreby rozsiahleho dodatočného tréningu.

Naše experimenty ďalej preukázali, že využitie *task prompt vectors* ako inicializačného mechanizmu je obzvlášť prínosné v režimoch s obmedzeným množstvom anotovaných dát. Dosiahli sme porovnateľný alebo lepší výkon v porovnaní s existujúcimi metódami prenosu soft promptov, pričom sme zachovali modularitu a parametrickú úspornosť riešenia. Navrhovaný prístup tak predstavuje efektívny a škálovateľný spôsob multi-task prenosu znalostí pri adaptácii veľkých jazykových modelov na rôzne NLP úlohy.

V ďalšom experimente [6] sme sa venovali problematike toho, že existujúce hodnotenia PEFT metód sú fragmentované a často obmedzené na úzky výber datasetov, modelov alebo metrík, čo sťažuje ich objektívne porovnanie a reprodukovateľnosť výsledkov. Preto

navrhujeme systematický a jednotný benchmark zameraný na konzistentné hodnotenie PEFT prístupov pri autoregresívnych jazykových modeloch.

Predstavili sme benchmark PEFT-Bench, ktorý zahŕňa 27 datasetov pokrývajúcich rôznorodé úlohy (klasické NLP úlohy, matematické problémy aj generovanie zdrojového kódu) a 7 reprezentatívnych PEFT metód. Zaviedli sme novú metriku PEFT Soft Cost Penalties (PSCP), ktorá kombinuje kvalitu modelu s nákladovými faktormi, ako sú počet trénovateľných parametrov, pamäťová náročnosť a rýchlosť inferencie. Cieľom je poskytnúť komplexné hodnotenie, ktoré zohľadňuje nielen výkon, ale aj praktickú efektivitu jednotlivých prístupov.

Na základe rozsiahleho experimentálneho hodnotenia ukazujeme, že jednotlivé PEFT metódy sa významne líšia nielen vo výkone, ale aj v nárokoch na výpočtové zdroje. Prostredníctvom metriky PSCP dokážeme tieto rozdiely systematicky kvantifikovať a analyzovať kompromisy medzi efektivitou a presnosťou. Náš benchmark tak poskytuje jednotný a reprodukovateľný rámec, ktorý umožňuje spravodlivejšie porovnávanie PEFT metód a podporuje informované rozhodovanie pri ich praktickom nasadení.

Vo ďalšej práci [7] sme navrhli PEFT-Factory, jednotný a modulárny rámec pre doladenie pomocou PEFT pre autoregresívne jazykové modely. Tento rámec integroval 19 rôznych PEFT metód, podporoval 27 datasetov pre klasifikáciu a generovanie textu a obsahoval metriky pre štandardné aj PEFT-špecifické hodnotenia, čo významne zlepšilo experimentálne porovnávanie a reprodukovateľnosť výsledkov v rámci výskumu PEFT metód.

Pri implementácii sme ako základ využili existujúci open-source projekt LLaMA-Factory, ktorý sme rozšírili o podporu širšej skupiny PEFT metód vrátane reparametrizácie, soft prompt-based a adapter-based prístupov. Okrem toho sme sprístupnili dynamický systém pre načítanie vlastných používateľských PEFT metód, čím sme umožnili ich jednoduchú integráciu a rozširovanie. Výsledkom bol stabilný a ucelený softvérový ekosystém, ktorý uľahčuje adopciu a experimentovanie výskumníkov, bez potreby budovania infraštruktúry od nuly.

Naše experimentálne prípadové štúdie ilustrovali, že PEFT-Factory dokázal efektívne reprodukovať a porovnávať výkon rôznych PEFT metód na viacerých úlohách a modeloch. Tým sme ukázali, že rámec nielen znižuje bariéry pri používaní týchto techník, ale zároveň podporuje transparentnejšie a systematickejšie hodnotenie PEFT metód.

5 Záver

Tento dokument prezentuje výsledky systematicky orientovaného výskumu zameraného na zlepšenie automatizovaného spracovania informácií pre mediálnych profesionálov. Sústreďujeme sa na tri kľúčové oblasti: detekciu overenia-hodných tvrdení, identifikáciu naratívov a detekciu dezinformačných kampaní, pričom osobitnú pozornosť venujeme jazykom z nízkym množstvom trénovacích dát a metódam parameter-efficient fine-tuning (PEFT).

Medzi hlavné príspevky patrí vytvorenie viacjazyčného a doménovo diverzifikovaného datasetu MultiCW s out-of-distribution časťou na testovanie generalizácie modelov. Experimentálne výsledky ukazujú, že dolaďované menšie modely dokážu prekonávať najnovšie veľké jazykové modely v zero-shot nastavení, čo poukazuje na efektívnejšie a udržateľnejšie prístupy k adaptácii modelov.

Navrhnutý hierarchický systém detekcie naratívov umožňuje extrahovať koherentné informačné rámce a poskytuje interaktívne vizualizácie vhodné pre analýzu šírenia informácií. V oblasti low-resource NLP demonštrujeme efektívnosť LLM-based augmentácie najmä pri obmedzených dátach a ukazujeme, že náhodný výber príkladov je často najstabilnejšou stratégiou.

V oblasti PEFT predstavujeme koncept „task prompt vectors“, benchmark PEFT-Bench s metrikou PSCP zohľadňujúcou aj výpočtové náklady a open-source rámec PEFT-Factory podporujúci reprodukovateľné experimentovanie.

Navrhnuté riešenia kladú dôraz na škálovateľnosť, efektívnosť a praktickú využiteľnosť. Do budúcnosti plánujeme rozšíriť systém o časové modelovanie, znalostné bázy a pokročilé analytické nástroje na včasnú identifikáciu nebezpečných informačných aktivít. Projekt tak prispieva k vyššej efektívnosti, transparentnosti a udržateľnosti AI systémov v mediálnej analýze.

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*.
2. Santana, B., Campos, R., Amorim, E. et al. A survey on narrative extraction from textual data. Artif Intell Rev 56, 8393–8435 (2023). https://doi.org/10.1007/s10462-022-10338-7
3. Cegin, J., Simko, J. and Brusilovsky, P., 2025, April. LLMs vs Established Text Augmentation Techniques for Classification: When do the Benefits Outweight the Costs?. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 10476-10496).
4. Cegin, J., Pecher, B., Simko, J., Srba, I., Bielikova, M. and Brusilovsky, P., 2025, November. Use Random Selection for Now: Investigation of Few-Shot Selection Strategies in LLM-based Text Augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2025* (pp. 5533-5550).
5. Belanec, R., Ostermann, S., Srba, I. and Bielikova, M., 2025, September. Task prompt vectors: Effective initialization through multi-task soft prompt transfer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 77-94). Berlin, Heidelberg: Springer Berlin Heidelberg.
6. Belanec, R., Pecher, B., Srba, I. and Bielikova, M., 2025. PEFT-Bench: A Parameter-Efficient Fine-Tuning Methods Benchmark. *arXiv preprint arXiv:2511.21285*. [ACCEPTED EACL 2026]
7. Belanec, R., Srba, I. and Bielikova, M., 2025. PEFT-Factory: Unified Parameter-Efficient Fine-Tuning of Autoregressive Large Language Models. *arXiv preprint arXiv:2512.02764*. [ACCEPTED EACL 2026]