Rezumat
„Advancements in Fake News Automatic Detection”,

Ștefan Emil Repede

Teza de doctorat reprezintă un studiu aprofundat asupra problematicii știrilor false în mediul digital, cu accent pe dezvoltarea unor soluții automate de identificare și clasificare a acestora. Lucrarea abordează fenomenul dintr-o perspectivă largă, combinând analiza conceptuală cu contribuții tehnice aplicate.

Autorul propune o definiție clară și operațională a ceea ce înseamnă „știre falsă”, ancorată în verificabilitatea conținutului și intenția deliberată de inducere în eroare. În cadrul lucrării este realizată o distincție între diverse tipuri de informații înșelătoare, precum dezinformarea, misinformarea, conținutul manipulat sau umoristic, dar și forme mai nuanțate de prezentare distorsionată a realității. Aceste clasificări sunt analizate din punctul de vedere al intenției, al stilului de redactare și al impactului, oferind o structură teoretică utilă pentru dezvoltarea ulterioară a metodelor automate.

Un aspect important al cercetării constă în crearea și adnotarea a două seturi de date bilingve, în limbile română și engleză, menite să sprijine atât clasificarea binară (adevărat versus fals), cât și clasificarea multi-clasă, care reflectă mai fidel diversitatea gradelor de veridicitate a conținutului. Etichetarea include categorii intermediare precum „parțial adevărat” sau „înșelător”, ceea ce contribuie la o abordare mai realistă a fenomenului.

Teza introduce două modele de detecție bazate pe arhitectura transformer, considerate contribuții tehnice centrale. Primul model, numit FakeLuke, este dezvoltat pornind de la arhitectura LUKE și este specializat în clasificarea binară, având integrat un mecanism de auto-atenție care ia în considerare entitățile semantice din text. Cel de-al doilea model este o versiune ajustată a LLaMA 3, adaptată pentru clasificare multi-clasă în contexte multilingve. Reglajul fin al acestuia a fost realizat prin tehnici de optimizare eficiente din punct de vedere computațional, cum ar fi Low-Rank Adaptation (LoRA), care permit scalabilitate și reutilizare în alte limbi sau domenii.

Ambele modele au fost supuse unor teste comparative, fiind analizate în raport cu metode tradiționale de învățare automată și modele avansate de procesare a limbajului natural. Rezultatele indică performanțe ridicate în ceea ce privește acuratețea, precizia și robustețea clasificării, atât pentru FakeLuke, în cazul sarcinilor binare, cât și pentru modelul bazat pe LLaMA 3, în sarcinile mai complexe, cu etichetare detaliată. Evaluările au fost realizate folosind metrice standard, iar rezultatele obținute au fost comparate cu cele ale unor modele cunoscute, precum BERT, GPT-4 sau Gemini.

Autorul analizează și limitele abordărilor propuse, subliniind că modelele se concentrează în principal pe conținut textual, în timp ce dezinformarea contemporană tinde să fie multimodală, incluzând imagini, videoclipuri și sunet sintetic. În acest sens, sunt propuse direcții viitoare de cercetare care vizează extinderea

spre detecția de conținut mixt. De asemenea, este abordată problema interpretabilității modelelor mari de limbaj și se sugerează integrarea unor metode de inteligență artificială explicabilă, care să ofere transparență în deciziile algoritmilor.

Lucrarea se încheie cu o discuție privind implicațiile etice ale detectării automate a știrilor false, insistând asupra necesității ca aceste tehnologii să fie utilizate responsabil, în conformitate cu principiile libertății de exprimare și ale supravegherii umane. Se subliniază importanța colaborării interdisciplinare între experți în inteligență artificială, jurnaliști și specialiști în științe sociale pentru ca soluțiile propuse să răspundă nevoilor reale ale societății.

În ansamblu, teza oferă un cadru teoretic și tehnic solid pentru înțelegerea și abordarea automată a fenomenului știrilor false, contribuind la consolidarea integrității ecosistemului informațional digital și la dezvoltarea unor instrumente mai eficiente și adaptabile pentru identificarea conținutului înșelător.

Student-doctorand,