УДК 004.8:81’322.2

ДОСЛІДЖЕННЯ МЕТОДІВ АВТОМАТИЧНОЇ ОБРОБКИ ТЕКСТОВОЇ ІНФОРМАЦІЇ ДЛЯ СТВОРЕННЯ КОРПУСІВ УКРАЇНСЬКОЇ МОВИ

Панченко К.А.
e-mail: kiril.panchenko@nure.ua
Науковий керівник – доц. Валенда Н.А.
Харківський національний університет радіоелектроніки, каф. ПІ
м. Харків, Україна

The current development of natural language processing (NLP) models requires large, high-quality, and verified text corpora. For the Ukrainian language, existing resources are mostly static and lack dynamic vocabulary from social networks. This work focuses on developing an automated system for creating a verified Ukrainian language corpus using Telegram data and official news websites. The methodology involves building a robust NLP pipeline for text preprocessing and applying a semantic verification algorithm. By utilizing TF-IDF vectorization and cosine similarity, the system compares unstructured messenger posts with reliable news sources. This approach filters out disinformation and noise, ensuring a high-quality dataset suitable for training modern AI models.

В епоху стрімкого розвитку штучного інтелекту та технологій обробки природної мови (NLP) ключовим ресурсом, що визначає якість кінцевих моделей, стають дані. Ефективність сучасних великих мовних моделей (LLM) безпосередньо залежить від наявності великих, репрезентативних та якісно розмічених текстових корпусів [1]. Для української мови проблема створення актуальних ресурсів стоїть особливо гостро [2]. Існуючі академічні корпуси (наприклад, ГРАК) часто є статичними, а масовий збір даних (web scraping) із соціальних мереж, зокрема з месенджера Telegram, без належної фільтрації призводить до накопичення «шуму» та дезінформації. Створення корпусу «сліпим» методом створює ризик навчання моделей на недостовірних фактах.

Метою роботи є розробка методів автоматичної обробки текстової інформації та проєктування архітектури програмного забезпечення для створення актуального, верифікованого корпусу української мови на основі даних із Telegram-каналів та авторитетних новинних ресурсів.

Основна ідея запропонованого підходу полягає в інтеграції етапу фактологічної верифікації (Fact-Checking) безпосередньо у процес збору даних [3]. Замість складних нейромережевих детекторів фейків, які потребують великих розмічених вибірок [4], пропонується використання методу порівняння з довіреними джерелами (Reference-based Verification) за допомогою оцінки семантичної подібності текстів (Semantic Textual Similarity).

Процес обробки та верифікації розділено на кілька алгоритмічних етапів. На першому етапі (Data Preprocessing) сирі дані з месенджера проходять через спеціалізований NLP-конвеєр. Він включає нормалізацію, очищення від специфічного шуму (URL-посилання, емодзі, хештеги), токенізацію та лематизацію за допомогою бібліотеки spaCy (модель *uk_core_news_sm*). Окремим важливим кроком є фільтрація доменно-специфічних стоп-слів, характерних для середовища месенджерів (наприклад, рекламних закликів «підпишись», «посилання», «канал»), що дозволяє суттєво зменшити розмір словника без втрати змістовної інформації. Лематизація є критично важливою для української мови як флективної, оскільки вона дозволяє зменшити розмірність векторного простору шляхом зведення різних словоформ до єдиної леми.

На другому етапі відбувається перетворення нормалізованих текстів у числовий вигляд за допомогою моделі «мішка слів» зі зважуванням TF-IDF. Цей метод ефективно фільтрує загальновживані слова та виділяє унікальні терміни, що описують конкретну подію.

Для визначення ступеня семантичної близькості між вектором новини з Telegram та векторами новин з перевірених джерел (еталонів) використовується метрика косинусної подібності:

$$Sim(\vec{A}, \vec{B}) = \cos(\theta) = \frac{\vec{A} \cdot \vec{B}}{\| \vec{A} \| \cdot \| \vec{B} \|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}, \quad (1)$$

де $\vec{A}$ – вектор новини з Telegram, $\vec{B}$ – вектор статті з офіційного новинного ресурсу. Значення подібності лежить у діапазоні $[0,1]$.

Для оптимізації обчислювальної складності пошук еталонної новини $\vec{B}$ здійснюється не по всій базі даних, а лише серед підмножини кандидатів, які потрапляють у задане часове вікно (Time Window Filtering): $[T_{tg} - \Delta t, T_{tg} + \Delta t]$, де T_{tg} – час публікації в месенджері. Якщо максимальне знайдене значення $Sim(\vec{A}, \vec{B})$ перевищує встановлений поріг (threshold τ), повідомлення отримує статус «Verified» і додається до фінального корпусу із посиланням на першоджерело.

Спроєктована архітектура системи реалізується мовою Python та має модульну структуру (збір, обробка, зберігання, аналіз). Взаємодія з джерелами забезпечується бібліотеками Telethon (для асинхронного парсингу MTProto) та BeautifulSoup. Для гнучкого зберігання неструктурованих BSON-документів використовується документо-орієнтована СУБД MongoDB. Особливістю архітектури є логічний розподіл бази даних на дві основні колекції: *telegram_posts* (зберігає оригінальні тексти, леми та метадані) та *trusted_news* (містить верифіковані

еталонні статті). Такий розподіл дозволяє алгоритму максимально швидко виконувати пошук та агрегацію даних.

Для об’єктивної оцінки ефективності запропонованого методу розроблено критерії експериментального дослідження.

Оцінка якості класифікації контенту здійснюватиметься за допомогою стандартних метрик інформаційного пошуку: Precision (точність), яка покаже частку дійсно правдивих новин серед підтверджених системою; Recall (повнота) для оцінки здатності знаходити всі релевантні факти; та загального балансу алгоритму через F1-Score. Окремим критичним параметром оцінювання стане середній час обробки одного повідомлення для підтвердження здатності системи працювати в режимі реального часу.

Висновки. Запропоновано метод автоматичної верифікації текстових даних під час формування корпусу української мови. Підхід базується на попередній обробці тексту, векторизації TF-IDF та визначенні семантичної подібності між повідомленнями з Telegram і новинами з авторитетних джерел за допомогою косинусної міри. Розроблено архітектуру програмної системи для збору, обробки та зберігання текстових даних. Запропонований метод дозволяє зменшити інформаційний шум і підвищити якість корпусу, що може бути використано для навчання сучасних NLP-моделей.

Список використаних джерел:

1. Chaplynskyi D. Introducing UberText 2.0: A Corpus of Modern Ukrainian at Scale. Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP), м. Dubrovnik, Croatia. Stroudsburg, PA, USA, 2023. DOI: https://doi.org/10.18653/v1/2023.unlp-1.1

2. Guo Z., Schlichtkrull M., Vlachos A. A Survey on Automated Fact-Checking. Transactions of the Association for Computational Linguistics. 2022. Т. 10. С. 178–206. DOI: https://doi.org/10.1162/tacl_a_00454

3. Haltiuk M., Smywiński-Pohl A. On the Path to Make Ukrainian a High-Resource Language. Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025), м. Vienna, Austria (online). Stroudsburg, PA, USA, 2025. С. 120–130. DOI: https://doi.org/10.18653/v1/2025.unlp-1.14

4. V. Vysotska, A. Chupryna, N. Valenda, O. Konduforov “Evaluating the in-context learning capabilities of large language models for misinformation detection for Ukrainian news”, CIAW-2025, Lviv. URL: https://ceur-ws.org/Vol-4110/paper15.pdf. (дата звернення: 25.02.2026).