УДК: 004.85 :004.912:316.77

Комаров В. Ф., канд. техн. наук,
старший викладач кафедри прикладної математики та кібербезпеки,
Стеблина Н. О., д-р політ. наук, професор, професор кафедри журналістики,
Донецький національний університет імені Василя Стуса

АНАЛІЗ ЕФЕКТИВНОСТІ КЛАСИФІКАЦІЇ УКРАЇНОМОВНИХ ДОПИСІВ У ПАБЛІКАХ TELEGRAM ЗАСОБАМИ ШІ

Анотація. Досліджено ефективність п'яти моделей ШІ для виявлення маніпуляцій в україномовних Telegram-дописах. На власному датасеті порівняно логістичну регресію, нейронну мережу на TF-IDF, LSTM та Transformers. Донавчена модель Transformers (youscan/ukr-roberta-base) показала найкращу здатність розпізнавати міноритарний негативний клас. Головним обмеженням визначено сильний дисбаланс класів у даних.

Ключові слова: соціальні мережі, медіагігієна, класифікація текстів, машинне навчання.

Вступ. У сучасному інформаційному просторі пабліки в соціальних мережах та месенджерах (як-от Telegram) стали потужним інструментом для поширення новин та думок. Водночас це створює ризики поширення прихованої реклами («джинси») та маніпулятивного контенту. Так, дослідження низки українських професійних медіаорганізацій показали наявність матеріалів з ознаками замовності у Telegram [1–5]. Автоматизація виявлення таких дописів [6–9] є важливим завданням у протидії інформаційним загрозам та забезпеченні медіагігієни. Українські дослідники медіа вивчають маніпулятивні дописи в медіа вручну: зазвичай експерти відбирають повідомлення за короткий проміжок часу та виявляють такий контент [10]. Метою цього дослідження є порівняльний аналіз ефективності п'яти різних моделей машинного навчання для класифікації тональності та релевантності україномовних дописів у Telegram на власному датасеті.

Основний текст. Дослідження проводилося на основі датасету, що містить текстові дописи з Telegram-пабліків, розмічені як позитивні (pos) та негативні (neg), де «негативні» дописи, ймовірно, містять маніпуляції чи «джинсу» (рис. 1). Датасет укладений на основі повідомлень одного з найбільш популярних телеграм-каналів в Україні – «Украина сейчас». Попередні дослідження показали, що це медіа розміщує «джинсу», зокрема просуваючи одного з українських олігархів [11].

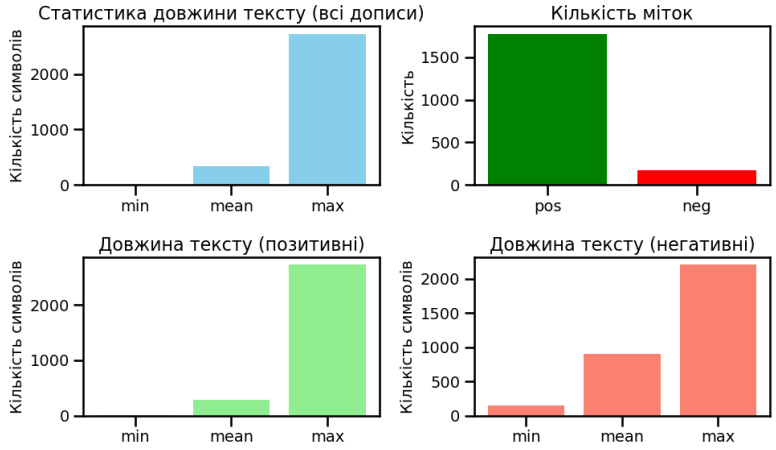

Рисунок 1 – Статистичні показники набору даних

Розмір повідомлень варіювався від 6 до 2 725 знаків (із середньою довжиною – 337.65 знаків). Кількість позитивних міток у датасеті – 1 777, негативних – 171 (за загальної кількості дописів – 1 948). Маніпулятивні (негативні) дописи та позитивні (звичайні новини) відбиралися вручну у період – серпень 2024 року. До уваги бралися критерії, що були встановлені для матеріалів з ознаками замовності Інститутом демократії імені Пилипа Орлика [10].

Для аналізу було протестовано п’ять моделей з різними підходами до обробки тексту: **1** – базова модель, що використовує логістичну регресію на основі векторизації тексту за методом TF-IDF; **2** – проста нейронна мережа з одним лінійним шаром, яка працює на тих самих TF-IDF векторах; **3** – модифікація базової моделі з попередньою лематизацією тексту за допомогою бібліотеки Stanza; **4** – рекурентна нейронна мережа, здатна аналізувати послідовності та враховувати контекст слів; **5** – сучасна модель на архітектурі трансформера (попередньо навчена модель youscan/ukr-roberta-base [12]), донавчена на зібраному досліджуваному датасеті для задачі класифікації.

Порівняння ключових метрик (зважених середніх) демонструє загальну картину ефективності моделей (рис. 2, табл. 1).

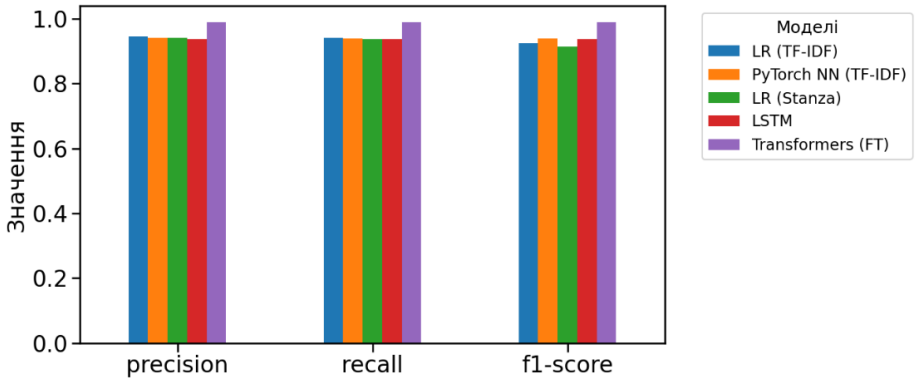

Рисунок 2 – Загальна продуктивність моделей

Таблиця 1 – Ключові метрики (зважені середні) досліджуваних моделей

Метрика	LR (TF-IDF)	PyTorch NN (TF-IDF)	LR (Stanza)	LSTM	Transformers (FT)
Точність	0.945	0.940	0.940	0.937	**0.990**
Повнота	0.941	0.939	0.936	0.936	**0.990**
Міра F1	0.924	0.939	0.914	0.936	**0.989**

На перший погляд, усі моделі показують високу точність (>93 %). Проте детальний аналіз матриць невідповідностей, які представлені на рис. 3, для всіх розглянутих моделей та метрик для окремих класів виявляє суттєві відмінності.

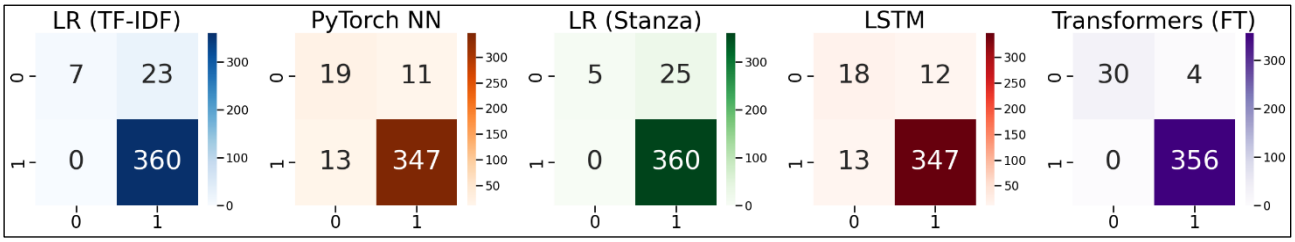

Рисунок 3 – Порівняння матриць невідповідностей

Моделі на основі TF-IDF (LR, PyTorch NN, LR+Stanza), хоч і показали високу загальну точність, мають критично низький показник повноти для негативного класу (від 17 % до 63 %). Це означає, що вони не здатні ефективно виявляти маніпулятивні дописи, пропускаючи значну їх частину. Цікаво, що лематизація за допомогою Stanza навіть погіршила результат, знизивши повноту для негативного класу з 23 % до 17 %.

Модель LSTM демонструє кращу здатність до розпізнавання негативного класу (повнота 53 %), порівняно з базовими моделями, що підтверджує важливість врахування послідовності слів. Однак вона все ще пропускає майже половину маніпулятивних дописів.

Модель Transformers (Fine-Tuned) очікувано [9; 13] виявилась беззаперечним лідером. Вона досягла найвищих показників за всіма метриками, а головне – показала значно вищу повноту для негативного класу (76 %). Це свідчить про те, що архітектура трансформерів завдяки механізмам уваги та контекстуальному розумінню мови найкраще підходить для виявлення семантичних нюансів, характерних для «джинси» та маніпуляцій.

Основною проблемою дослідження є сильний дисбаланс класів у використаному датасеті (1 777 позитивних дописів проти 171 негативного). Наслідками дисбалансу є:

1. *Оманлива точність.* Моделі можуть досягати високої точності, просто прогнозуючи домінантний клас (pos). Тому метрики точності, повноти та міри F1 для міноритарного класу (neg) є значно більш показовими.

2. *Низька повнота для негативного класу.* Моделі недостатньо «вчаться» на прикладах маніпуляцій, через що погано їх розпізнають. Навіть найкраща модель (Transformers) ідентифікує лише приблизно три з чотирьох таких дописів.

Висновки. Донавчена модель на архітектурі Transformers (youscan/ukr-roberta-base) показала найкращу ефективність для задачі виявлення прихованої реклами та маніпуляцій, суттєво перевершивши інші розглянуті моделі за здатністю розпізнавати міноритарний «негативний» клас.

Головним обмеженням представлених досліджень виявився сильний дисбаланс класів у зібраному навчальному датасеті, що робить загальні показники точності оманливими та знижує практичну цінність моделей через низьку повноту.

Для підвищення якості розпізнавання потрібно збалансувати датасет, що можна зробити шляхом збору більшої кількості прикладів маніпулятивних дописів або застосувавши техніки аугментації даних (наприклад, синонімізації, зворотного перекладу) спеціально для міноритарного класу з подальшим повторним донавчанням моделі Transformers [13].

Подяка. Дослідження виконано в межах програми 2025.06 «Наука для зміцнення обороноздатності і національної безпеки України» Національного фонду досліджень України (проєкт № 2025.06/0090, номер державної реєстрації 0125U003181).

Список використаних джерел

1. Стеблина Н. Нам пишуть, що щось коїться – як регіональні телеграм-канали інформують українців під час Великої війни. *Інститут демократії імені Пилипа Орлика*. 20.12.2024.

URL: https://idpo.org.ua/reports/6136-nam-pishut-shho-shhos-koïtsya-yak-regionalni-telegram-kanali-informuyut-ukraïnciv-pid-chas-velikoï-vijni.html (дата звернення: 27.10.2025).

2. How Non-Institutionalized News Telegram-Channels Operate and Capture the Audience in Ukrainian Segment. Analytical Report – Kyiv, UMCI NGO, 2023. 69 p.

3. Регіональні медіа під час виборів 2020. *Committee of voters of Ukraine.* URL: http://cvu.od.ua/ua/library/monitoring-regionalnih-media-pid-chas-viboriv-2020-zvit_1269/ (дата звернення: 27.10.2025).

4. Контент українських медіа в 2021 році. Підсумки моніторингів. *IMI.* https://imi.org.ua/monitorings/kontent-ukrayinskyh-media-u-2021-rotsi-pidsumky-monitoryngiv-imi-i43114 (дата звернення: 27.10.2025).

5. Медіа без стандартів. Як популярні телеграм-канали показують достовірність інформації. *IMI*. 26.07.2024. URL: https://imi.org.ua/monitorings/media-bez-standartiv-yak-populyarni-telegram-kanaly-pokazuyut-dostovirnist-informatsiyi-i62753 (дата звернення: 27.10.2025).

6. Makogon I., Samokhin I. Targeted Sentiment Analysis for Ukrainian and Russian News Articles. *ICTERI 2021 Workshops*, Springer International Publishing. 2022. P. 538–549. DOI: 10.1007/978-3-031-14841-5_36.

7. Shynkarov Y., Solopova V. High quality sentiment analysis model for Ukrainian social media. *Proceedings of the 6th Masters Symposium MS-AMLV-2025.* Lviv, Ukraine, 2025. URL: https://apps.ucu.edu.ua/en/ms-amlv-2025-proceedings/ (дата звернення: 19.10.2025).

8. Zalutska O. Method for Analyzing the Ukrainian Language Texts Sentiment Using Natural Language Processing. *Системи контролю інформації та інтелектуальні технології. Досягнення та застосування*. Liha-Pres, 2025. P. 122–137. DOI: 10.36059/978-966-397-538-2-7.

9. Prytula M. Fine-tuning BERT, DistilBERT, XLM-RoBERTa and Ukr-RoBERTa models for sentiment analysis of ukrainian language reviews. *Artificial Intelligence.* 2024. Vol. 29(2). P. 85–97. DOI: 10.15407/jai2024.02.085.

10. Методологія оцінювання якості контенту в регіональних друкованих та он-лайн виданнях. *Інститут демократії імені Пилипа Орлика*. URL: https://idpo.org.ua/wp-content/uploads/methodology-2023.pdf (дата звернення: 27.10.2025).

11. Стеблина Н. Медіахолдингу немає, джинса триває. Тепер Ріната Ахметова вихваляють у телеграмі. *Детектор медіа.* 29.12.2022. URL: https://detector.media/monitorynh-internetu/article/206483/2022-12-29-mediakholdyngu-nemaie-dzhynsa-tryvaie-teper-rinata-akhmetova-vykhvalyayut-u-telegrami/ (дата звернення: 27.10.2025).

12. Ukrainian Roberta base model. *Hugging Face.* URL: https://huggingface.co/youscan/ukr-roberta-base (дата звернення: 19.10.2025).

13. Shynkarov Y. Developing a robust text-based sentiment analysis model for Ukrainian social media. Ukrainian Catholic University, Faculty of Applied Sciences, Department of Computer Sciences. Lviv 2025. x, 40 p. URL: https://hdl.handle.net/20.500.14570/5706 (дата звернення: 19.10.2025).