Шейко Максим Олексійович

Київський національний університет імені Тараса Шевченка, Київ

0009-0002-3809-3487

**ВИКОРИСТАННЯ ВЕЛИКОЇ МОВНОЇ МОДЕЛІ GPT 4.1 ДЛЯ ВИЯВЛЕННЯ ДЕЗІНФОРМАЦІЇ У СОЦІАЛЬНИХ МЕРЕЖАХ**

У сучасному цифровому суспільстві відбувається зміна того, як люди споживають інформацію. Соціальні мережі, зокрема Twitter, дедалі частіше виступають головним джерелом новин для мільйонів користувачів. Даний процес став ще швидше під час епідемії COVID-19, коли майже всі сфери життя перейшли на «онлайн» режим. Це також стосується і новин. Наприклад, у 2024 році майже 25% опитуваних жителів Сполучених Штатів сказали, що вони часто споживають новини з соціальних мереж. Ще 29% сказали, що роблять це іноді. Загалом у всьому світі налічується 5,22 мільярда користувачів соціальних мереж[1]. Очевидно, що соціальні мережі надають інформацію з будь-яких акаунтів, включаючи ті, що ширять неправдиву інформацію. Отже, задача виявлення дезінформації у текстових повідомленнях є актуальною.

З активним розвитком сфери штучного інтелекту, з’являється все більше методів та підходів, здатних до знаходження закономірностей у текстах та виявленні неправдивої інформації. Наявні дослідження вказують на те, що навіть традиційні методи машинного навчання у задачах класифікації (метод опорних векторів, метод випадкового лісу та лінійна регресія) показують достатні результати на тренувальних наборах даних[2-7]. Використання моделей-трансформерів, таких як BERT та його варіантів здатне значно підвищити точність прогнозів.

Поточне дослідження направлене на порівняння наявних методів з комбінованим підходом, що використовує декілька трансформерних моделей для виявлення дезінформації. Розроблена система включає декілька етапів: попередню класифікацію повідомлень для фільтрації неінформативного контенту (наприклад, мемів), обробку медіа за допомогою моделі BLIP, генерацію ключових слів з використанням GPT-4.1, пошук актуальних джерел через DuckDuckGo, класифікацію повідомлення через модель DistilBERT та остаточну оцінку від великої мовної моделі з урахуванням зібраного контексту. Такий багаторівневий процес дозволяє досягти повного урахування контексту повідомлення для оцінки.

Для об'єктивної оцінки ефективності реалізованої системи виявлення дезінформації було проведено експеримент, що полягає у порівнянні точності виявлення з традиційними підходами, що були розглянуті. Метою дослідження стало визначення значення покращення точності класифікації повідомлень після впровадження комбінованого методу, який поєднує кілька джерел контексту, трансформерні моделі та отримання кінцевого результату від великої мовної моделі GPT 4.1. У рамках експерименту було створено новий валідаційний набір даних із 2137 повідомлень, з яких 467 були марковані як дезінформація, а 1670 — як правдиві. Маркування даних здійснювалося з використанням ChatGPT для оптимізації часових витрах. Вдалося отримати результати класичних моделей (TF-IDF + SVM, Logistic Regression), трансформерів (BERT, RoBERTa, DistilBERT) та запропонованого комбінованого підходу. Результати порівняння наведено на рисунку 1.

A graph with orange bars

AI-generated content may be incorrect.

Рисунок 1. Результат порівняння підходів класифікації фейкових повідомлень.

Найкращий результат було отримано саме за допомогою комбінованої системи – точність склала **82.5%,** що на **6% вище** за найближчу модель-конкурент (DistilBERT, навчена на первинному датасеті). Це свідчить про ефективність підходу, що інтегрує кілька шарів аналізу та пояснення.

Таким чином, експеримент показав, що врахування зовнішнього контексту, фільтрація повідомлень та інтеграція пояснювального модуля підвищують здатність системи точно класифікувати складні або неоднозначні повідомлення. Проте, результати також демонструють, що модель залишається чутливою до навчального датасету, що підтверджує необхідність подальшого вдосконалення системи генерації контексту та розширення бази знань моделі.

# **Література**

1. Social Media and News Fact Sheet [Електронний ресурс] // Pew Research Center. – Режим доступу: <https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/> (дата звернення: 29.04.2025).
2. Shushkevich E. Fake News Detection in Online Platforms : Doctoral thesis, Technological University, Dublin. – Dublin, 2024. – 161 p. – Режим доступу: <https://arrow.tudublin.ie/ittthedoc/16/> (дата звернення: 29.04.2025).
3. Bozzi M. The Role of Machine Learning in Automatic Fake News Detection : Master thesis, Luiss University. – Rome, 2023. – 72 p.
4. Alkhodair S. A., Ding S. H. H., Fung B. C. M., Liu J. Detecting breaking news rumors of merging topics in social media // Information Processing & Management. – 2020. – Vol. 57. – Режим доступу: <https://doi.org/10.1016/j.ipm.2019.102150> (дата звернення: 29.04.2025).
5. Shivam B., Parikh P. K., Atrey P. K. Media-Rich Fake News Detection: A Survey // Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval. – Miami, FL : IEEE, 2018.
6. Jain A., Kasbe A. Fake News Detection // 2018 IEEE International Students Conference on Electrical, Electronics and Computer Science (SCEECS). – Bhopal, India, 2018.
7. Chan A. GPT-3 and InstructGPT: technological dystopianism, utopianism, and “Contextual” perspectives in AI ethics and industry // AI and Ethics. – 2024. – P. 53–64.