МЕТОД НЕЙРОМЕРЕЖЕВОГО ВИЯВЛЕННЯ ПРИЙОМІВ ПРОПАГАНДИ ЗА МАРКЕРАМИ У ТЕКСТОВОМУ КОНТЕНТІ З ВІЗУАЛЬНОЮ АНАЛІТИКОЮ ПРИЙНЯТИХ РІШЕНЬ

Молчанова М.О., m.o.molchanova@gmail.com

Хмельницький національний університет

Пропаганда, прихована під виглядом звичайних новин, поширюється вже багато десятиліть, однак сучасна цифрова епоха створює сприятливі умови для ще швидшого і більш масового її розповсюдження. Нові методи генерації текстів, які дедалі частіше мало відрізняються від створених людиною, призводять до стрімкого зростання кількості контенту [1]. Це підкреслює важливість створення автоматизованих методів для виявлення пропагандистських маніпуляцій, що сприятиме усвідомленому сприйняттю інформації користувачами і забезпеченню безпеки інформаційного середовища.

На сьогоднішній день ефективними засобами для автоматизованого виявлення та класифікації прийомів і об’єктів пропаганди у текстовому контенті є засоби штучного інтелекту. Однак, попри бурхливий розвиток галузі обробки природної мови, відсутній комплексний інструмент, який би не лише забезпечував виявлення та класифікацію прийомів пропаганди, а й показував на кого і на що вона спрямована. Також проблемою використання моделей глибокого навчання є їх низька інтерпретованість, що породжує недовіру до нейромережевого виявлення пропаганди.

Схема методу виявлення прийомів пропаганди за маркерами у текстовому контенті з візуальною аналітикою прийнятих рішень наведена на рис. 1. Працює метод шляхом перетворення вхідних даних у вигляді множини навчальних текстів для ідентифікації кожного маркера пропаганди, множини навчальних текстів для кожного прийому пропаганди та тестового тексту для виявлення використаних прийомів пропаганди у вихідні дані у вигляді множини навчених нейромережевих моделей для ідентифікації кожного з прийомів пропаганди, множини навчених нейромережевих моделей для ідентифікації кожного маркера з доповненої множини маркерів, та оціненого тексту щодо сил проявів кожного

прийому пропаганди, а також візуальна аналітика прийнятих класифікаторами рішень [2].

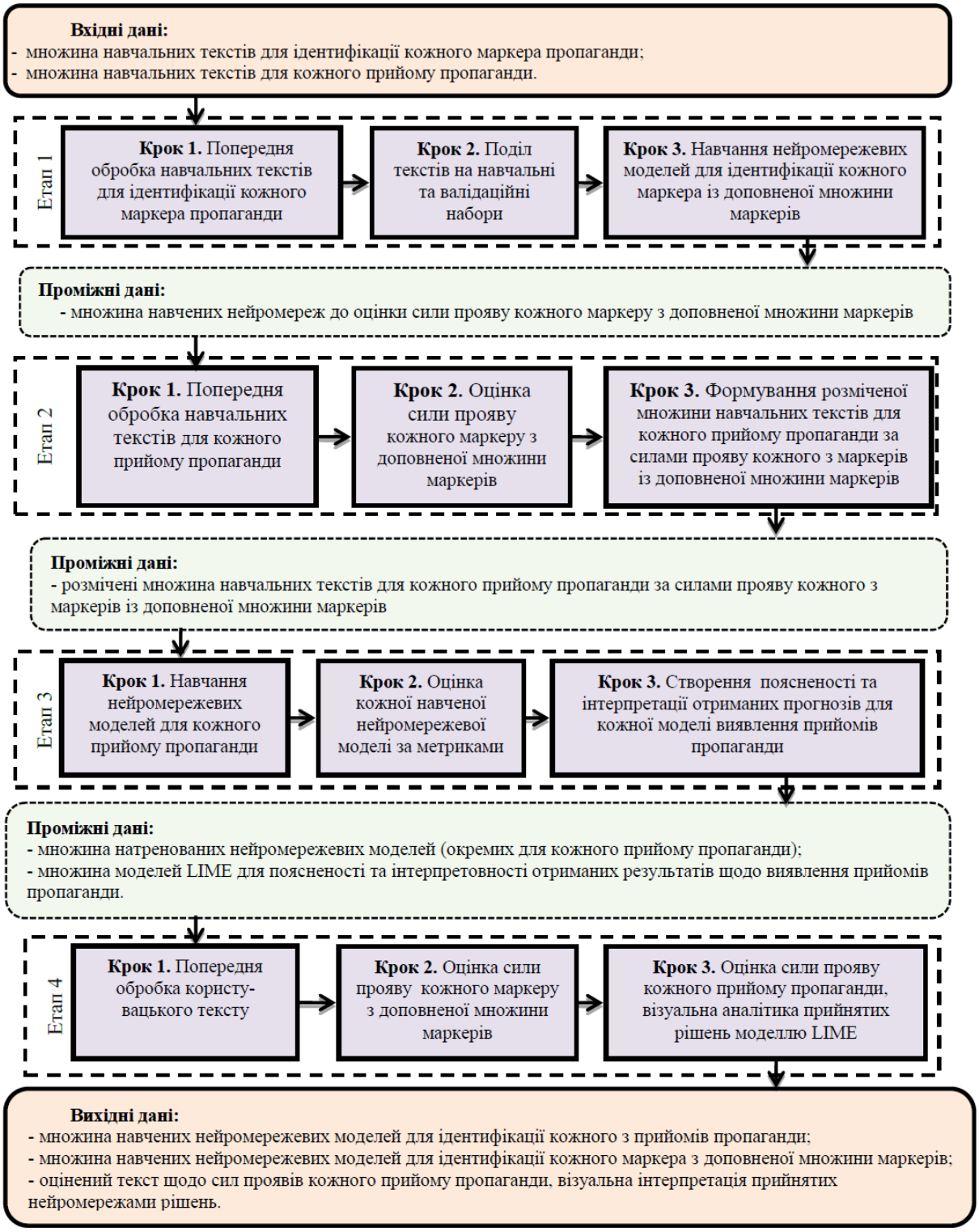

Рис. 1. Схема методу виявлення прийомів пропаганди за маркерами

Етап 1 складається з послідовного виконання кроків попередньої обробки навчальних текстів для ідентифікації кожного маркера пропаганди, поділу

текстів на навчальні та валідаційні набори та навчання нейромережевих моделей для ідентифікації кожного маркера із доповненої множини маркерів.

Етап 2 складається із виконання таких послідовних кроків, як: попередня обробка навчальних текстів для кожного прийому пропаганди, оцінка сили прояву кожного маркеру з доповненої множини маркерів та кроку формування розміченої множини навчальних текстів для кожного прийому пропаганди за силами прояву кожного з маркерів із доповненої множини маркерів.

Етап 3 складається із послідовного виконання таких кроків: навчання нейромережевих моделей для кожного прийому пропаганди (1 модель відповідає 1 окремому прийому пропаганди), оцінки кожної навченої нейромережевої моделі за метриками та створення поясненості у вигляді візуальної аналітики для отриманих прогнозів по кожної моделі виявлення прийомів пропаганди.

Етап 4 відповідає безпосередньо за класифікацію текстового контенту, та полягає у послідовному виконанні таких кроків: попередня обробка користувацького тексту, що відбувається у відповідності до використовуваних нейромережевих засобів, оцінці сили прояву кожного маркеру з доповненої множини маркерів та оцінка сили прояву кожного прийому пропаганди з візуальною аналітикою прийнятих рішень методом LIME.

Запропонований метод дозволяє як отримувати значення сили прояву застосованих прийомів пропаганди, так і дає додаткову поясненість у вигляді оцінок маркерів, а також візуальної аналітики. В ході навчання нейромережевих моделей, що відповідають за класифікацію використаних прийомів пропаганди, було досягнуто значень за метрикою Accuracy від 0.82 до 0.97. Більш детально дані дослідження наведені на рис. 2.

Розроблений метод протестовано на ідентифікації 17 прийомів пропаганди [3]. Однак, їх перелік можна розширити, на що і будуть спрямовані подальші дослідження.

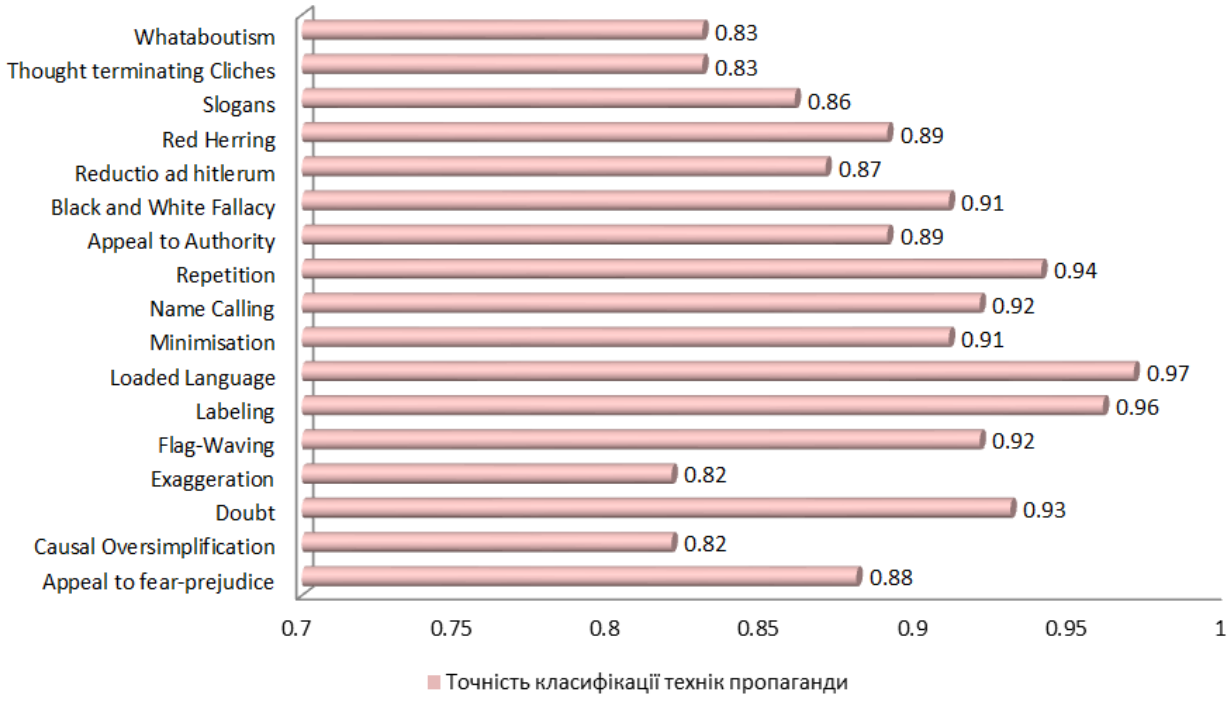

Рис. 2. Досягнута точність класифікації використаних прийомів пропаганди

Отже, розроблено метод виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень, який відрізняється від існуючих використанням доповненої множини маркерів для виявлення прийомів пропаганди, що дозволяє пояснити отримані результати і підвищити точність та якість виявлення пропаганди.

Список використаних джерел

1. Faye G., Icard B., Casanova M., Chanson J., Maine F., Bancilhon F., Gadek G., Gravier G., Egre P. Exposing propaganda: an analysis of stylistic cues comparing human annotations and machine classification. *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language.* 2024. P. 62–72.
2. Молчанова М.О. Метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання. *Вісник Хмельницького національного університету*. 2024. 5(341). С. 344-350.
3. Молчанова М.О. Нейромережеве виявлення і класифікація прийомів та об’єктів пропаганди у текстовому контенті. *Міжнародний науково-технічний журнал «Вимірювальна та обчислювальна техніка в технологічних процесах».* 2024. № 4. С. 153-161.