УДК 004.8:004.912

ОСОБЛИВОСТІ ОБРОБКИ ТЕКСТОВИХ ДАНИХ СОЦІАЛЬНИХ МЕРЕЖ ПРИ АНАЛІЗІ ЕМОЦІЙНОЇ ТОНАЛЬНОСТІ

Ткаченко І.Є.
e-mail: ihor.tkachenko1@nure.ua
Науковий керівник – д.т.н., проф. Перова І.Г.
Харківський національний університет радіоелектроніки, каф. КМІТ
м. Харків, Україна

Are being considered the specific features of processing textual data from social networks for sentiment analysis. Social media comments, particularly on YouTube, reflect users' opinions and emotional reactions to digital content. However, such texts often contain slang, abbreviations, emojis, and spelling errors, which complicate automatic analysis. Therefore, text preprocessing plays an important role and includes data cleaning, normalization, and tokenization. Special attention is given to emojis that may carry additional emotional meaning. Modern Natural Language Processing methods and transformer models such as BERT improve the accuracy of sentiment classification. The use of the YouTube API allows automatic collection of comments for analysis. The obtained results can be visualized to better understand audience reactions.

Сучасні соціальні мережі є одним із найбільших джерел текстових даних, що відображають думки, емоції та реакції користувачів на різні події, продукти або інформаційний контент. Однією з популярних платформ для обміну думками є YouTube, де користувачі активно залишають коментарі під відеоматеріалами. Аналіз таких коментарів дозволяє отримати цінну інформацію щодо громадської думки, рівня задоволеності аудиторії та загальної емоційної реакції користувачів. Саме тому задача аналізу емоційної тональності текстів (sentiment analysis) стала важливим напрямом досліджень у галузі обробки природної мови (Natural Language Processing, NLP) [1, 2].

Аналіз тональності передбачає автоматичну класифікацію текстових повідомлень за їх емоційною характеристикою. Найчастіше використовується поділ на три основні класи: позитивний, нейтральний та негативний. У контексті соціальних мереж така класифікація дозволяє швидко оцінювати реакцію користувачів на певний контент без необхідності ручного аналізу великої кількості повідомлень [2].

Разом із тим текстові дані соціальних мереж мають ряд специфічних особливостей, які значно ускладнюють процес автоматичного аналізу. На відміну від формальних текстів, коментарі користувачів часто містять скорочення, сленг, емодзі, орфографічні помилки, а також неструктуровані речення. Крім того, тексти можуть бути короткими, емоційно насиченими або містити сарказм, що ускладнює правильну інтерпретацію змісту [2].

Однією з ключових проблем є наявність великої кількості шуму в текстових даних. До такого шуму можна віднести зайві символи, посилання, HTML-теги, повторювані символи, хештеги або емодзі. Без попередньої обробки такі елементи можуть негативно впливати на якість роботи алгоритмів машинного навчання. Тому важливим етапом аналізу є попередня підготовка тексту (preprocessing), яка включає очищення даних, нормалізацію тексту та його перетворення у формат, придатний для подальшої обробки.

Попередня обробка тексту зазвичай включає декілька етапів. Першим етапом є очищення тексту від небажаних елементів, таких як спеціальні символи, HTML-теги або посилання. Другим етапом є нормалізація тексту, що передбачає приведення слів до нижнього регістру, видалення зайвих пробілів та стандартизацію написання слів. Також застосовується токенізація – процес поділу тексту на окремі слова або токени, що дозволяє аналізувати структуру речення та підготувати текст до подальшої обробки моделями машинного навчання.

Ще однією важливою особливістю текстів соціальних мереж є використання емодзі та емоційних символів. Хоча такі елементи можуть розглядатися як шум, вони часто несуть важливу інформацію про емоційний стан автора повідомлення. Наприклад, смайли можуть підсилювати позитивну або негативну тональність тексту. Тому сучасні підходи до аналізу тональності намагаються враховувати такі елементи під час обробки текстових даних.

З розвитком технологій штучного інтелекту значного поширення набули трансформерні моделі, які демонструють високу ефективність у задачах обробки природної мови. На відміну від класичних алгоритмів машинного навчання, трансформери здатні враховувати контекст слів у реченні та краще розуміти семантичні зв’язки між ними. Це дозволяє підвищити точність класифікації текстів [3], зокрема у задачах аналізу емоційної тональності.

Серед популярних трансформерних моделей можна виділити BERT, RoBERTa та інші моделі, які використовуються для задач класифікації тексту. Такі моделі попередньо навчаються на великих масивах текстових даних, після чого можуть бути адаптовані до конкретних задач, зокрема аналізу коментарів соціальних мереж.

У процесі аналізу коментарів YouTube важливим етапом є отримання текстових даних. Для цього використовується офіційний API платформи YouTube, який дозволяє автоматично завантажувати коментарі до відео [4]. Після отримання даних виконується їх збереження, попередня обробка та підготовка до аналізу. Далі текст передається у модель BERT, яка виконує класифікацію повідомлень за емоційною тональністю на три основні категорії: позитивну, нейтральну та негативну. Загальний процес збору, обробки та аналізу коментарів представлено на рисунку 1.

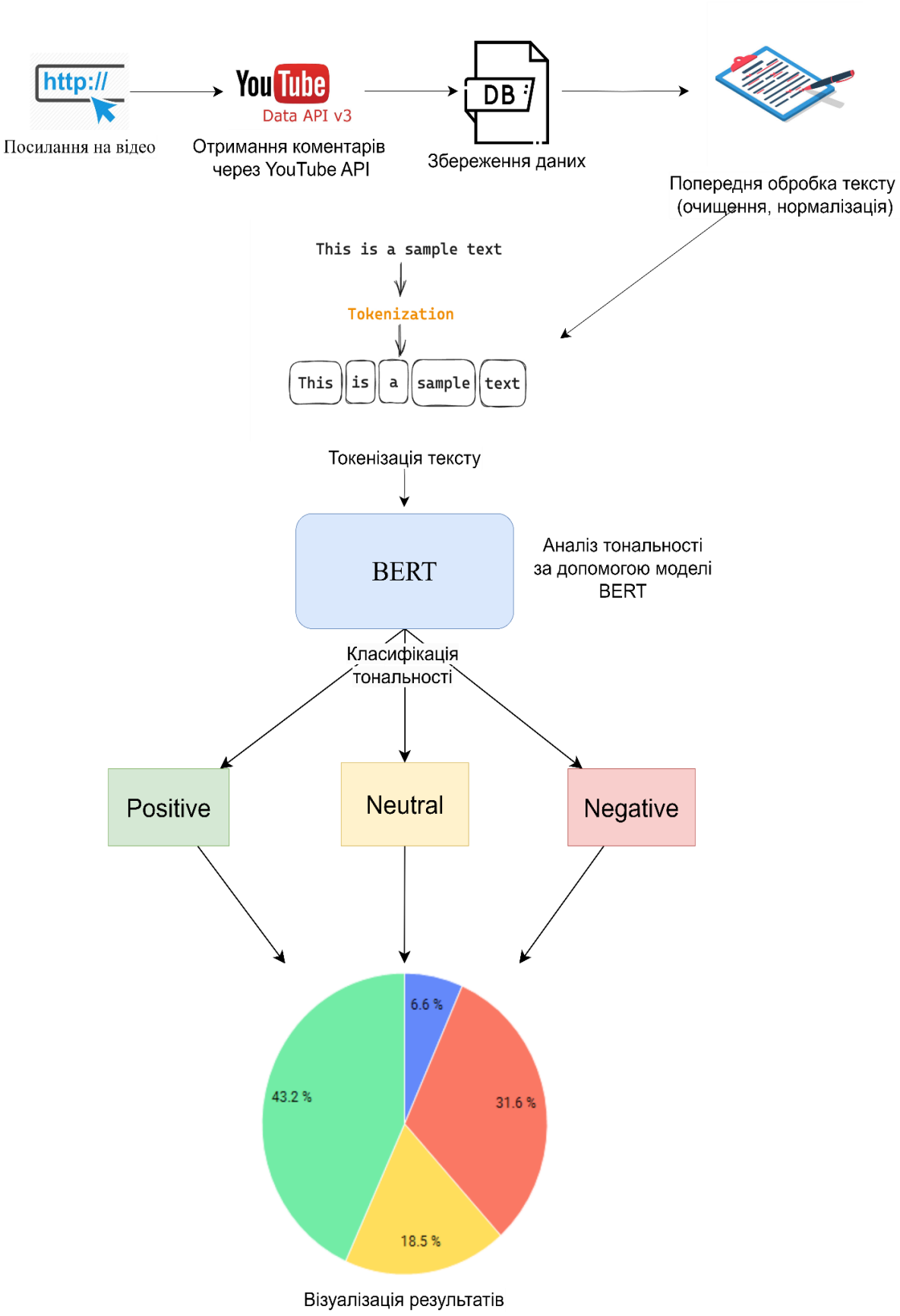

Рисунок 1 – Схема процесу аналізу емоційної тональності коментарів YouTube

Після класифікації результатів важливим етапом є їх інтерпретація та візуалізація. Наприклад, результати аналізу можуть бути представлені у вигляді діаграм або гістограм, що відображають співвідношення позитивних, негативних та нейтральних коментарів. Це дозволяє швидко оцінити загальну реакцію аудиторії на відеоконтент [3, 5].

Таким чином, аналіз емоційної тональності коментарів соціальних мереж є складною задачею, що потребує врахування специфічних особливостей текстових даних.

Використання сучасних методів обробки природної мови та трансформерних моделей дозволяє значно підвищити точність аналізу та отримувати більш достовірні результати.

Подальший розвиток таких підходів відкриває нові можливості для дослідження громадської думки, аналізу поведінки користувачів та підвищення ефективності цифрових сервісів.

Список використаних джерел:

1. Liu B. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, 2012. 167 p.

2. Jurafsky D., Martin J. H. Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 3rd ed. Pearson, 2020. URL: https://web.stanford.edu/~jurafsky/slp3/ (дата звернення: 08.03.2026).

3. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). 2019. P. 4171–4186.

4. YouTube Data API : вебсайт. URL: https://developers.google.com/youtube/v3 (дата звернення: 08.03.2026).

5. Vaswani A., Shazeer N., Parmar N. та ін. Attention Is All You Need // Advances in Neural Information Processing Systems (NeurIPS). 2017. Vol. 30.