УДК 004.056.53:004.912

Власова А.С., Кушніренко Н.І., Назарова І.В.

Національний університет «Одеська політехніка»

АЛГОРИТМ ТЕКСТОВОГО АНАЛІЗУ ДЛЯ ПРОФІЛЮВАННЯ КОРИСТУВАЧІВ В OSINT ДОСЛІДЖЕННЯХ

Вступ. В умовах стрімкого зростання обсягів текстової інформації в соціальних мережах виникає потреба в автоматизованих методах аналізу цифрового сліду користувачів. Текстові дані містять найбільший обсяг інформації про особистість, погляди та поведінкові характеристики користувачів, що робить їх аналіз ключовим компонентом систем OSINT (Open Source Intelligence) [1].

Традиційні методи текстового аналізу часто не враховують морфологічні особливості української мови та специфіку інтернет-комунікації, що знижує точність профілювання. Існуючі міжнародні рішення демонструють точність лише 48-55% для україномовного контенту, що недостатньо для практичного застосування в системах розвідки з відкритих джерел [2]. Додаткову складність створює необхідність обробки змішаного контенту, де поєднуються українська та англійська мови, інтернет-сленг, емоджі та неологізми.

Сучасні дослідження показують, що навіть короткі текстові фрагменти містять достатньо лінгвістичних маркерів для побудови психологічного профілю автора [3], однак для їх ефективного виявлення необхідні спеціалізовані алгоритми, адаптовані до особливостей цільової мови та культурного контексту.

Мета. Підвищити точність профілювання користувачів соціальних медіа в OSINT-дослідженнях шляхом розробки алгоритму комплексного аналізу текстового контенту, адаптованого до української мови та особливостей інтернет-комунікації.

Основна частина

Розроблений алгоритм комплексного аналізу текстового контенту складається з п'яти послідовних етапів обробки, кожен з яких вирішує специфічні задачі аналізу з урахуванням особливостей української мови [4].

Перший етап включає нормалізацію та очищення тексту від технічних артефактів. Процес нормалізації передбачає приведення тексту до стандартного формату UTF-8 з корекцією можливих помилок кодування, видалення HTML тегів та спеціальних символів, які не несуть семантичного навантаження, обробку емоджі зі збереженням їх емоційного значення, оскільки вони є важливим маркером стилю комунікації.

Другий етап реалізує токенізацію та лематизацію з урахуванням морфологічних особливостей української мови. Використовується спеціалізований токенізатор, який враховує українські конструкції з прийменниками з апострофом, складні числівники та назви власні, специфічні інтернет-скорочення та абревіатури. Лематизація виконується з використанням морфологічних словників для української мови, що дозволяє привести всі словоформи до їх основної форми.

Третій етап включає автоматичне визначення мови тексту з використанням бібліотеки langdetect, що важливо для обробки змішаного контенту. Алгоритм

використовує статистичний підхід на основі частотного аналізу символьних n-грам для української та англійської мов. Система здатна визначати мову з точністю 91% для українських текстів, 92% для англійських та 87% для змішаного контенту.

Четвертий етап реалізує сентимент-аналіз текстового контенту для визначення емоційного забарвлення повідомлень користувача [2]. Використовується гібридний підхід, який поєднує словникові методи з елементами частотного аналізу. Система класифікує тексти за трьома основними категоріями емоційного забарвлення: позитивне, негативне та нейтральне. Для української мови створено спеціалізований словник емоційно забарвлених слів та виразів з урахуванням контекстуальних особливостей їх використання. Алгоритм досягає точності 74% для українських текстів, що на 17-19% вище за універсальні міжнародні рішення.

П'ятий етап включає частотний аналіз ключових слів та виділення сутностей з тексту користувача. Алгоритм використовує метод TF-IDF для автоматичного виявлення найбільш важливих термінів у корпусі текстів користувача:

$$TF - IDF(t,d) = \left(\frac{f(t,d)}{max_{freq(d)}}\right) \times \log\left(\frac{N}{|\{d \in D: t \in d\}|}\right), \quad (1)$$

де $f(t,d)$ – частота терміну t у тексті d,

$max\_freq(d)$ – максимальна частота будь-якого терміну в тексті d,

N – загальна кількість текстових фрагментів користувача,

$|\{d \in D: t \in d\}|$ – кількість текстових фрагментів, що містять термін t.

Метод TF-IDF дозволяє виявляти терміни, які є специфічними для конкретного користувача та відрізняють його тексти від загального корпусу. Система аналізує частотні характеристики лексики для визначення тематичних преференцій користувача та побудови профілю його інтересів з точністю 77%.

Експериментальне тестування алгоритму проводилось на вибірці з 100 користувачів україномовних соціальних медіа різних вікових категорій та рівнів активності. Результати точності алгоритмів текстового аналізу для різних типів мовного контенту представлені в таблиці 1.

Таблиця 1 – Результати точності алгоритмів текстового аналізу

Метрика	Українська мова	Англійська мова	Змішаний контент
Визначення мови	91%	92%	87%
Сентимент-аналіз	74%	71%	65%
Виділення ключових слів	77%	74%	69%

Інтеграція всіх п'яти етапів створює комплексну систему текстового аналізу, яка забезпечує багатоаспектну характеристику мовної поведінки користувача. Результати обробки включають лексичний профіль з показником багатства словника, емоційний профіль з розподілом позитивних, негативних та нейтральних висловлювань, тематичний профіль з виявленими ключовими інтересами та сферами діяльності. Ці характеристики використовуються для побудови інтегрованого цифрового профілю користувача та оцінки його

особистісних якостей на основі цифрового сліду.

Порівняльний аналіз розробленого алгоритму з існуючими міжнародними рішеннями показує його переваги саме для української аудиторії. Система враховує понад 800 популярних українських інтернет-скорочень та неологізмів, які активно використовуються в соціальних медіа. Створено спеціалізований словник емоційно забарвлених слів, адаптований до особливостей українського менталітету та культурного контексту, що суттєво підвищує точність сентимент-аналізу порівняно з універсальними багатомовними рішеннями.

Особливу увагу приділено обробці змішаного україно-англійського контенту, який є типовим для сучасної інтернет-комунікації українських користувачів. Алгоритм здатен коректно обробляти тексти з частковим перемиканням мов, транслітерацією українських слів латиницею та використанням англомовних термінів у україномовному контексті.

Висновок. У процесі дослідження було підвищено точність профілювання користувачів соціальних медіа шляхом розробки алгоритму комплексного аналізу текстового контенту, адаптованого до української мови та особливостей інтернет-комунікації. Алгоритм включає п'ять етапів: нормалізація тексту, токенізація та лематизація з морфологічними словниками української мови, визначення мови через langdetect, сентимент-аналіз та частотний аналіз ключових слів методом TF-IDF.

Експериментальне тестування на вибірці з 100 користувачів підтвердило підвищення точності: 74% для сентимент-аналізу українських текстів та 77% для виділення ключових слів, що на 17-19% вище за універсальні міжнародні рішення. Система забезпечує ефективну обробку змішаного контенту з точністю визначення мови 87% та враховує понад 800 популярних українських інтернет-скорочень.

Розроблений алгоритм може бути успішно застосований в OSINT дослідженнях для автоматизованого профілювання користувачів україномовних соціальних медіа та створює основу для побудови інтегрованих цифрових профілів користувачів.

Перелік використаних джерел.

1. Deeva I. Computational Personality Prediction Based on Digital Footprint of A Social Media User. Procedia Computer Science. 2019. Vol. 156. P. 185–193.

2. Birjali M., Kasri M., Beni-Hssane A. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. Knowledge-Based Systems. 2021. Vol. 226. P. 107134. URL: https://doi.org/10.1016/j.knosys.2021.107134

3. Azucar D., Marengo D., Settanni M. Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. Personality and Individual Differences. 2018. Vol. 124. P. 150–159.

4. Nandwani P., Verma R. A review on sentiment analysis and emotion detection from text. Social Network Analysis and Mining. 2021. Vol. 11, no. 1. URL: https://doi.org/10.1007/s13278-021-00776-6