УДК 004.93

РОЗРОБКА TELEGRAM-БОТА З ФУНКЦІЄЮ СЕНТИМЕНТ-АНАЛІЗУ ТЕКСТУ

Кулик Владислав Олександрович
асистент Петросюк Денис Валерійович
Національний університет «Одеська політехніка», УКРАЇНА

АНОТАЦІЯ. Досліджено існуючі підходи до визначення емоційної забарвленості текстів та визначено переваги Lexicon-based методу. Розроблена система являє собою бот з функцією сентимент-аналізу текстових повідомлень. Бот реалізовано для месенджера Telegram, використовуючи мову програмування Python та бібліотеку python-telegram-bot. Сентимент-аналіз здійснюється за допомогою Lexicon-based методу, доповненого словами-модифікаторами, який базується на аналізі тональних словників та їх впливу на загальний сентимент тексту. Розроблений бот може знайти застосування в різних сферах, включаючи моніторинг соціальних мереж, аналіз відгуків клієнтів, вимірювання публічного настрою.

Вступ. З розвитком інформаційних технологій та стрімкого накопичення великих масивів даних широкого розповсюдження набула така область комп'ютерної лінгвістики, як сентимент аналіз. З розвитком даних технологій стало можливим автоматично витягати з тексту виражену автором думку, а також оцінювати текст як позитивний, негативний, а при необхідності виокремлювати конкретні емоції (радість, гнів, сум тощо). Для виокремлення емоційної оцінки автора застосовуються підходи з використанням тональних словників і правил або застосовують методи машинного навчання [1,2] .

Розробка Telegram-ботів є актуальною темою в сучасному світі інформаційних технологій. У зв'язку зі зростанням кількості користувачів месенджерів та збільшенням обсягів обміну інформацією у мережі, з'явилась потреба в розробці Telegram-бота з функцією сентимент-аналізу тексту на основі Lexicon-based методу. Розробка такого бота забезпечує можливість швидкого та зручного аналізу текстів для користувачів Telegram. Окрім того, Telegram-бот з функцією сентимент-аналізу тексту може бути корисною для бізнесу, який бажає відстежувати реакції користувачів на продукт або послуги.

Мета роботи. Дослідження методики сентимент-аналізу текстових повідомлень та на її основі розробка Telegram-бота, які будуть забезпечувати оптимальну точність та високу швидкість аналізу, що дозволить користувачам визначати емоційний стан своїх співрозмовників в чатах Telegram та підвищить ефективність спілкування.

Основна частина роботи. Один з основних методів аналізу текстів чи повідомлень, пов'язаних із сентиментом, це аналіз сентименту на одному із рівнів, зазвичай виділяють три рівні: повідомлення, речення та об'єкту/аспекту. Сентимент аналіз на рівні повідомлення передбачає визначення загального настрою тексту – позитивного або негативного. Сентимент аналіз на рівні речення визначає, який настрій виражається в кожному окремому реченні – позитивний, нейтральний чи негативний.

Для визначення тональності тексту також використовують тональні словники [2], які містять як загальні емоційно-забарвлені слова, так і вузькоспеціалізовану лексику. Кожному слову або словосполученню дається оцінка, що характеризує позитивний чи негативний сентимент. Часто такі словники складаються вручну або напівавтоматизованими методами, але також є способи повністю автоматичного укладання тональних словників. Виконуючи аналіз із застосуванням словників, кожному слову в тексті присвоюється певне значення тональності, взяте із тонального словника (якщо слово присутнє в словнику).

У статті [3] автори пропонують лексиконний підхід до вилучення настрою з тексту. Вони розробили Semantic Orientation CALculator (SO-CAL), який використовує словники слів, анотованих їх семантичною орієнтацією (полярність та сила), та враховує інтенсифікацію та заперечення. SO-CAL застосовується до завдання класифікації полярностей, процесу

призначення позитивної або негативної марки тексту, яка передає думку тексту щодо його основного предмета.

Лематизація та стемінг [4] допомагають зменшити кількість слів, з якими потрібно працювати в аналізі тексту, що зменшує розмір даних та поліпшує точність аналізу. Стемінг – це процес видалення закінчень слів, залишаючи лише корінь слова. Цей метод може використовуватися для зведення подібних слів до одного базового слова, зменшуючи кількість унікальних слів в тексті. Наприклад, слова "біг", "бігти", "бігти" будуть зведені до базового слова "біг". Лематизація – це більш складний процес, який враховує граматичні та морфологічні характеристики слова для зведення до його базової форми. Лематизація може зменшити кількість унікальних слів більше ефективно, ніж стемінг, оскільки вона враховує смисловий контекст. Наприклад, слова "біжу", "біжить", "біг" будуть зведені до базового слова "бігти".

Використовуючи метод тональних словників було досягнуто оптимальної точності та швидкості сентимент-аналізу для розробленого Telegram-бота (табл. 1). Точність в 66-69% хоч і є невеликою в порівнянні з машинним навчанням. Головним критерієм комфортного користування Telegram-ботом є швидкість відгуку (табл. 2). І незважаючи на велику кількість тексту, бот повинен завжди проводити сентимент-аналіз швидко.

Таблиця 1 – Таблиця порівняння тональних словників та методів глибинного навчання

Показник	Тональні словники	Машинне навчання
Швидкість	Швидко, виконується без значного затримки	Потребує тривалого часу на навчання та настройку
Вимоги до даних	Мінімальні, можна використовувати з відкритими джерелами даних	Високі, потрібна достатня кількість якісних даних для навчання
Рівень складності	Низький, не потрібна спеціальна підготовка	Високий, потрібна спеціалізована підготовка та знання алгоритмів та методів глибинного навчання

Таблиця 2 – Таблиця порівняння точності тональних словників та глибинного навчання

Метод	Ітерація	Точність, %	Швидкість, с	Метод	Ітерація	Точність, %	Швидкість, с
Тональні словники	1	68,9	1,8	Машинне навчання	1	82,3	3,8
	2	66,5	1,7		2	86,2	3,4
	3	67,8	1,7		3	84,7	3,7
	4	66,1	1,3		4	88,2	3,9
	5	68,3	1,5		5	78,7	3,1
	6	65,7	2,1		6	86,7	3,7

Після додавання слів-модифікаторів до аналізу сентименту, точність сентимент-аналізу значно підвищилась (табл. 3).

Таблиця 3 – Таблиця порівняння точності після додавання слів-модифікаторів

Метод	Середня точність, %	Середня швидкість, с
Машинне навчання	85,2	3,5
Тональні словники	74,2	1,6

Слова-модифікатори є додатковими словами, які мають змінювати або підсилювати сентимент в тексті. Вони можуть бути прикріплені до існуючих тональних словників або використовуватись самостійно. Завдяки використанню слів-модифікаторів, система сентимент-

аналізу стає більш чутливою до контексту та може краще розрізняти субтильні відтінки сентименту. Наприклад, слово "дуже" може підсилити позитивний або негативний сентимент, а слово "не" може змінити сентимент з позитивного на негативний або навпаки. Це доповнення до аналізу сентименту значно поліпшує його точність, оскільки враховує більше факторів і деталей в тексті. Результатом є більш точна і надійна оцінка сентименту, що допомагає користувачам отримувати більш об'єктивну інформацію про відношення до певних тем, продуктів чи послуг.

Висновки. В результаті було розроблено Telegram-бота з функцією сентимент-аналізу тексту на основі лексикон-базового методу. Такий метод сентимент-аналізу показав ефективність в аналізі великої кількості повідомлень та є простим у реалізації.

Реалізована методика сентимент-аналізу текстових повідомлень для Telegram-бота забезпечує оптимальну точність на рівні 74% та достатню швидкість аналізу на рівні 1-2с, що дозволить користувачам визначати емоційний стан своїх співрозмовників в чатах Telegram та підвищити ефективність спілкування. Точність аналізу вдалося значно підвищити більш ніж на 5% за рахунок додавання слів-модифікаторів до аналізу сентименту.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Mishchenko M. V., Dorosh M. S. “Semantic analysis and classification of malware for UNIX-like operating systems with the use of machine learning methods”. Applied Aspects of Information Technology. Publ. Nauka i Tekhnika. Odessa: Ukraine. 2022; Vol.5 No.4: 371–386. DOI: https://doi.org/10.15276/aait.05.2022.25
2. Kaur C. Social Issues Sentiment Analysis using Python [Електронний ресурс] / C. Kaur, A. Sharma – https://www.researchgate.net/publication/346962692_Social_Issues_Sentiment_Analysis_using_Python.
3. Taboada, M., J. Brooke, M. Tofiloski, K. Voll and M. Stede (2011) Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics 37 (2): 267-307.
4. Stemming vs Lemmatization in NLP: Must-Know Differences [Електронний ресурс] – https://www.analyticsvidhya.com/blog/2022/06/stemming-vs-lemmatization-in-nlp-must-know-differences/.

DEVELOPMENT OF TELEGRAM BOT WITH SENTIMENT ANALYSIS FUNCTION OF TEXT

Vladyslav Kulyk
assistant Denys Petrosiuk
Odesa Polytechnic National University, UKRAINE

ANNOTATION. The existing approaches to determining the emotional coloration of texts are investigated and the advantages of the developed method in relation to the machine learning method are determined. The developed system is a Telegram bot with the function of sentiment analysis of text. The bot is implemented for the Telegram messenger using the Python programming language and the python-telegram-bot library. The sentiment analysis is performed using the Lexicon-based method with modifier words, which is based on the analysis of tone dictionaries and their impact on the overall sentiment of the text. The developed bot can be used in various fields, including social media monitoring, customer feedback analysis, and public sentiment measurement.