Журнал обчислювальної та прикладної математики **2022, № 2** Journal of Numerical & Applied Mathematics

УДК 519.7
MSC 68U15

A SYSTEM OF INTELLECTUAL ANALYSIS AND PREDICTION OF REACTIONS TO NEWS BASED ON DATA FROM TELEGRAM CHANNELS

O. Yu. Kosukha, Iu. M. Shevchuk

Faculty of Computer Science and Cybernetics, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine, E-mail: kosukha.o@knu.ua, shevchuk.iuliia@knu.ua

СИСТЕМИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ТА ПРОГНОЗУВАННЯ РЕАКЦІЙ НА НОВИНИ НА ОСНОВІ ДАНИХ ТЕЛЕГРАМ-КАНАЛІВ

О.Ю. Косуха, Ю.М. Шевчук

Факультет комп'ютерних наук та кібернетики, Київський національний університет імені Тараса Шевченка, Київ, Україна, E-mail: kosukha.o@knu.ua, shevchuk.iuliia@knu.ua

Abstract. This research paper provides a description of the system of intellectual analysis and prediction of reactions to news based on data from Telegram channels. In particular, the features of collecting and pre-processing datasets for the system, the methodology of thematic analysis of the received data, and the model used to obtain predictions of reactions to Telegram messages depending on their text are described.
Keywords: natural language processing, sentiment analysis, naive Bayes classifiers, social media, Telegram messenger.

Анотація. У даній статті представлено опис системи інтелектуального аналізу та прогнозування реакцій на новини на основі даних Телеграм-каналів. Зокрема, описано особливості збирання та попередньої обробки наборів даних для системи, методику тематичного аналізу отриманих даних та модель, що була використовується системою для отримання прогнозів реакцій на Телеграм-повідомлення в контексті від його тексту.
Ключові слова: обробка природної мови, аналіз тональності тексту, наївний баєсів класифікатор, соціальні медіа, Телеграм.

Вступ

Телеграм, починаючи від своєї появи в 2013 році, з кожним роком збільшує свою частку користувачів в медіа-сфері. Особливо цьому сприяє його політика конфіденційності та цензура інших соціальних медіа. При цьому виникає потреба для аналізу повідомлень в публічних каналах цього типу медіа, які мають свої характерні особливості, яких немає, наприклад, в Твіттері або Фейсбуці. Наприклад, це можливість використання чат-ботів

для автоматизації тих чи інших функцій ([1–3]). Також з 30 січня 2021 року користувачі Телеграму можуть висловлювати своє відношення до повідомлень за допомогою реакції — смайлу, який можна поставити на повідомлення, — і який бачить автор повідомлення та інші користувачі. Це дає можливості для покращення алгоритмів аналізу тональності текстів повідомлень та аналізу поведінки користувачів, тобто ширші можливості для тих напрямів аналітики, які активно розвивались і до цього (наприклад, [4, 5]).

Потрібно відмітити, що для інших масмедіа, зокрема, Facebook і Twitter, дослідження контенту користувачів є задачами, які активно розв'язуються ([6–10]).

Мета роботи полягає у формалізації принципів, технологій та алгоритмів, які можна рекомендувати для розробки ефективної системи інтелектуального аналізу та прогнозування реакцій на новини на основі даних Телеграм-каналів, а також створення прототипу такої системи.

1. Збір та попередня обробка даних

Однією із задач дослідження було формування навчальної вибірки для інтелектуального аналізу повідомлень в Телеграм-каналах.

Нами було створено датасет, сформований з 21455 новинних повідомлень з реакціями Телеграм-каналу ТСН [12] з 09 березня по 20 вересня 2022 р. Основними критеріями при його виборі були: україномовний контент; відсутність орієнтації на специфічну цільову аудиторію, як, наприклад, у Телеграм-каналів районів міст і т.п.; активність підписників (середнє охоплення повідомлення близько 200 000 переглядів).

Тобто на основі цього можна обгрунтовано припускати репрезентативність сформованого датасету повідомлень з цього джерела.

Для отримання доступу до даних Телеграм-каналу була використана Telegram API ([11]). Для цього, а також для синтаксичного аналізу даних було розроблено програмну реалізацію на мові Python, код якої розміщено у відкритому доступі [13].

Також в якості попередньої обробки даних були видалені повідомлення, які не мають тексту (наприклад, картинки чи фотографії) або які мають лише посилання на сторінки в мережі Інтернет.

Фінальний набір даних має такі поля: дата і час повідомлення, кількість переглядів, його текст, типи реакцій, кількість реакцій кожного типу відповідно.

Множину типів реакцій звужено до таких типів:

Наведені типи реакцій, окрім смайла лиця клоуна — стандартний набір реакції для користувача Телеграм.

Смайл лиця клоуна до 18 вересня 2022 р. був доступний тільки для преміум-користувачів, але все одно є дуже популярним, тому його включено в множину типів реакцій.

2. Особливості реалізації тематичного аналізу повідомлень

Наступним етапом дослідження було перетворення отриманого набору даних в стандартизований формат, з якими могли б працювати алгоритми штучного інтелекту.

Оскільки для нашого дослідження не була потрібна деталізована інформація про реакції користувачів, їх було поділено на дві категорії:

– позитивні

– негативні

У підсумку замість фіксування абсолютної кількості позначок для кожного типу реакцій зберігалась інформація про відносну частку реакцій реакцій конкретної категорії відносно всієї кількості реакцій. Це значним чином зменшило кількість пам'яті, потрібної для збереження тестової вибірки, та спростило подальшу роботу на етапі навчання моделі для прогнозування реакцій на повідомлення.

Також кожне повідомлення було промарковані як позитивне або негативне: якщо відносна сумарна частка позитивних реакцій більша ніж 0.5, то повідомлення маркувалось як позитивне, і як негативне в іншому випадку.

У сформованому наборі даних налічується 13333 позитивних та 8122 негативних повідомлення (Рис. 1).

Також було проведено попередню обробку тексту:

– вилучено всі смайли
– видалено знаки пунктуації;
– видалено гіперпосилання на зовнішні ресурси;
– весь текст переведено в нижній регістр.

Також було проведено аналіз слів, які найчастіше використовуються в повідомлення, які ми розмістили як позитивні (Рис. 2) або негативні (Рис. 3).

При цьому були виключені з розгляду шумові слова та проведена лематизація слів (прикладом лематизації є групування слова «добре» за такими словами, як «краще» та «найкраще»). Потрібно відмітити, що ці дії не дали покращення для точності результатів моделі для прогнозу реакцій на повідомлення (детальніше про це в Розділі 3).

Рис. 1. Результати тематичного аналізу навчального набору даних

Рис. 2. Word cloud часто використовуваних слів в позитивних повідомленнях

Для візуалізації результатів використали представлення word cloud, для яких характерна прямопропорційний залежність розміру шрифту слова від частоти його вживання. Як можна відмітити, перетин множини слів, характерних для позитивних та негативних новин, є непустою множиною. Звідси можна припустити, що орієнтуючись під час використання алгоритмів природної обробки мови виключно на текстову частину новинного повідомлення, може відбутись спрощення сентиментального змістовного контексту

повідомлення, в той час як використання реакцій на повідомлення дозволяє цього уникнути.

Рис. 3. Word cloud часто використовуваних слів в негативних повідомленнях

3. Методика прогнозування реакцій на Телеграм-повідомлення

Для переведення тексту в числовий вид було застосовано два методи:

- модель “торба слів”, яка кожному слову співвідносить його частоту появи в тексті;
- TF-IDF, де вага слова пропорційна кількості вживань цього слова у повідомленні, і обернено пропорційна частоті вживання слова у інших повідомленях датасету.

Було використано два алгоритми:

- наївний баєсів класифікатор;
- метод опорних векторів.

Наївний баєсів класифікатор [14] для кожного слова формує умовні ймовірності того, що, якщо це слово є в повідомленні, то повідомлення віднесеться в відповідний клас. При застосуванні методу, речення розбивається на слова і всі раніше вирахувані умовні ймовірності слів множаться між собою. Такі добутки обчислюються для двох випадків: клас позитивний або негативний. Отримані дві ймовірності порівнюються і надається перевага тому класу, ймовірність якого більша. Цей метод активно використовується для розв’язку прикладних проблем у різних сферах (наприклад, [15–18]).

При роботі методу опорних векторів [19] для тренувального набору даних формується модель, яка відображає повідомлення у точку в просторі. Дане відображення працює так, щоб розділити повідомлення різних категорій якнайширшою прогалиною. При застосуванні методу, повідомлення

Табл. 1. Таблиця F-мір для різних моделей

Назва	F-міра без видалення шумових слів	F-міра з видаленням шумових слів
nb_cv	0.851	0.849
nb_tf_idf	0.861	0.859
svc_cv	0.853	0.851
svc_tf_idf	0.848	0.848

відображається на отриманий простір та належить тій категорії, на бік якої відносно прогалини воно потрапило [20–23].

Навчання проводилось з використання кожної комбінації методів, також ми спробували провести таке ж навчання з видаленням шумових слів.

Отримані F-міри на випробувальному датасеті для кожної моделі:

У Таб. 1 такі використовуються наступні скорочення:

- префікси в назвах: **nb** — наївний баєсівський класифікатор; **svc** — метод опорних векторів;
- суфікси в назвах: **cv** — переведення тексту в числовий вид за допомогою "торби слів"; **tf_idf** — переведення тексту в числовий вид за допомогою методики TF-IDF.

Можна зробити висновок, що видалення шумових слів не покращує точність моделі. Варто зазначити, що якщо слово зустрічається часто, то методика TF-IDF надає йому малу вагу. Таким чином, вплив деякої частини шумових слів нівелюється при використанні TF-IDF. Також маємо найкращу модель — наївний баєсівський класифікатор з використанням TF-IDF і F-мірою 0.86.

Висновки

З наведених вище результатів можна зробити висновок, що інтелектуальний аналіз повідомлень в Телеграм-каналах можна звести до використання вже розроблених раніше алгоритмів обробки природної мови. Але, якщо брати до уваги, що для української мови поки відносно мала кількість промаркованих наборів даних, в тому числі й змістовних словників тональності української мови (вони зараз на початковому рівні формування), то цей напрямок досліджень є доволі перспективним. Також потрібно відмітити потребу в дослідженнях, пов'язаних з розробкою інструментів для аналізу реакцій, адже сам Телеграм на даний момент не пропонує цей функціонал як базову частину аналітики для Телеграм-каналів.

Такі дослідження є особливо актуальними на фоні збільшення популярності Телеграму як нецензурованого майданчику для поширення новин та фейків.

На основі результатів, отриманих в цій роботі, можна формувати більші набори даних, використовуючи повідомлення одразу з багатьох каналів, та на їх базі формувати ефективніші моделі. На основі отриманих моделей можна розробляти системи для оцінки звучання тексту, що може допомогти

в написанні текстів в повсякденному житті, а також за допомогою моделі проводити аналітику повідомлень звичайних користувачів для оцінки громадського настрою.

Література

1. Mondal A., Dey M., Das D., Nagpal S., Garda K. Chatbot: An automated conversation system for the educational domain. *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*. 2018. P. 1–5.
2. Gunawan T. S., Babiker A. B. F., Ismail N., Effendi M. R. *Development of Intelligent Telegram Chatbot Using Natural Language Processing. In 2021 7th International Conference on Wireless and Telematics (ICWT)*. 2021. P. 1–5
3. Karimpour D., Chahooki M. A. Z., Hashemi A. User recommendation based on Hybrid filtering in Telegram messenger. *26th International Computer Conference, Computer Society of Iran (CSICC)*. 2021. P. 1–7.
4. Hashemi A., Zare Chahooki M. A. Telegram group quality measurement by user behavior analysis. *Social Network Analysis and Mining*. 2019. 9(1). P. 1–12.
5. Karimpour D., Zare Chahooki M. A., Hashemi, A. User recommendation in Telegram messenger by graph analysis and mathematical modeling of users' behavior. *Journal of Information and Communication Technology*. 2021. 49(49). 151.
6. Eichstaedt J. C., Smith R. J., Merchant R. M., Ungar L. H., Crutchley P., Preofiuc-Pietro D., Schwartz H. A. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*. 2018. 115(44). P. 11203–11208.
7. Kachamas P., Akkaradamrongrat S., Sinthupinyo S., Chandrachai A. Application of artificial intelligent in the prediction of consumer behavior from Facebook posts analysis. *International Journal of Machine Learning and Computing*. 2019. 9(1). P. 91–97.
8. Han B., Cook P., Baldwin T. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*. 2014. 49. P. 451–500.
9. Jordan S. E., Hovet S. E., Fung I. C. H., Liang H., Fu K. W., Tse Z. T. H. Using Twitter for public health surveillance from monitoring and prediction to public response. *Data*. 2018. 4(1). P. 6.
10. Essien A., Petrounias I., Sampaio P., Sampaio S. A deep-learning model for urban traffic flow prediction with traffic events mined from twitter. *World Wide Web*. 2021. 24(4). P. 1345–1368.
11. Telegram APIs. Режим доступу: https://core.telegram.org/
12. Телеграм-канал ТСН. Режим доступу: https://t.me/ТСН_channel
13. Код програмної реалізації. Режим доступу: https://github.com/KosukhaOlexandr/ reactions_prediction/blob/main/clear_dataset.py
14. Mosteller F., Wallace D. L. Inference and disputed authorship: The Federalist. *Stanford Univ Center for the Study*. 2007.
15. Козак Є. Б. Принципи впровадження моделей машинного навчання у сфері інтелектуального обслуговування промислового обладнання. *Таврійський науковий вісник. Серія: Технічні науки*. 2021. (3). С. 19–28.
16. Білецький Т. П., Федасюк Д. В. Прогнозування дефектів у програмному забезпеченні алгоритмами глибинного навчання CNN та RNN. *Науковий вісник НЛТУ України*. 2021. 31(2). С. 114–120.

17. Ahmad F., Tang X. W., Qiu J. N., Wrzblewski P., Ahmad M., Jamil I. Prediction of slope stability using Tree Augmented Naive-Bayes classifier: Modeling and performance evaluation. *Math. Biosci. Eng.* 2022. 19. P. 4526–4546.
18. Kewsuwun N., Kajornkasirat S. A sentiment analysis model of agritech startup on Facebook comments using naive Bayes classifier. *International Journal of Electrical & Computer Engineering* (2088-8708). 2022. 12(3).
19. Cortes C., Vapnik V. Support-Vector Networks. *Machine Leaming.* 1995. 20. P.273—297.
20. Jose C., Goyal P., Aggrwal P., Varma M. Local deep kernel learning for efficient non-linear svm prediction. *In International conference on machine learning.* 2013. P. 486–494.
21. Покідін Д. Економетрична модель Національного банку України для оцінки кредитного ризику банку та альтернативний метод опорних векторів. *Вісник Національного банку України.* 2015. 234. С. 53.
22. Верлань А. І., Олексій А. О. Огляд та порівняння методів машинного навчання для розпізнавання гідроакустичних сигналів. *Інфокомунікаційні та комп'ютерні технології.* 2022. 1 (03). С. 296–306.
23. Ramasamy L. K., Kadry S., Nam Y., Meqdad M. N. Performance analysis of sentiments in Twitter dataset using SVM models. *Int. J. Electr. Comput. Eng.* 2021. 11(3). P. 2275–2284.

Надійшла: 04.09.2022 / Прийнята: 10.10.2022