УДК 517
https://doi.org/10.17721/1812-5409.2022/3.7

О.Г. Наконечний[1],*д.ф.-м.н., проф.*
О.А. Капустян[1,2], *д.ф.-м.н., с.н.с.*
Ю.М. Шевчук[1,2], *к.ф.-м.н.*
М.В. Лосєва[1],
О.Ю. Косуха[1]

Інтелектуальна система аналізу реакцій на новини на основі даних Телеграм-каналів

[1] Київський національний університет імені Тараса Шевченка, 83000, м. Київ, пр-т. Глушкова 4д,
e-mail: iuliia.shevchk@knu.ua
[2] Університет м.Лаквіла, Італія, 67100 Лаквіла, будівля "Ренато Рікамо"(Коппіто 1), вул. Ветойо, Коппіто,
e-mail: iuliia.shevchuk@univaq.it

O.G. Nakonechnyi[1], *Dr.Sci.*
O.A. Kapustian[1,2], *Dr.Sci.*
Iu. M. Shevchuk[1,2], *PhD*
M.V. Loseva[1],
O. Yu. Kosukha[1]

A intellectual system of analysis of reactions to news based on data from Telegram channels

[1]Taras Shevchenko National University of Kyiv, 83000, Kyiv, 4d Glushkova str.,
e-mail: iuliia.shevchk@knu.ua
[2]The University of L'Aquila, Italy, 67100 L'Aquila, Edificio "Renato Ricamo" (Coppito 1), Via Vetoio, Coppito,
e-mail: iuliia.shevchuk@univaq.it

У даній статті представлено опис системи інтелектуального аналізу та прогнозування реакцій на новини на основі даних Телеграм-каналів. Запропоновано механізм збирання та попередньої обробки наборів даних для запропонованої системи, описано методику тематичного аналізу отриманих даних та модель, що 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.

This paper describes the system of intellectual analysis and prediction of reactions to the news based on data from Telegram channels. In particular, the features of collecting and pre-processing datasets for the intelligence systems, 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. We show the work of this system in the example of the Ukrainian news Telegram channel. The results are estimations of probability of emojis for the news from the testing dataset. Also, we give F-measures for our approaches to precise input data and models.

Key Words: natural language processing, sentiment analysis, naive Bayes classifiers, social media, Telegram messenger.

Статтю представив д.ф.-м.н. Пашко А.О.

Телеграм, починаючи від своєї появи в 2013 році, з кожним роком збільшує свою частку користувачів у медіа-сфері. Особливо цьому сприяє його політика конфіденційності та цензура інших соціальних медіа. При цьому виникає потреба для аналізу повідомлень у 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] - [2]). Також з 30 січня 2021 року користувачі Телеграму можуть висловлювати своє ставлення 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, [3] -[4]).

Потрібно відмітити, що для інших масмедіа, зокрема, Facebook і Twitter, дослідження контенту користувачів є задачами, які активно розв'язуються ([5] - [7]).

Мета роботи полягає у формалізації прин-

ципів, 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 у Телеграм-каналах.

Нами було створено датасет, сформований з 21455 новинних повідомлень з реакціями Телеграм-каналу ТСН [9] 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 обґрунтовано припускати репрезентативність сформованого датасету повідомлень із цього джерела.

Для отримання доступу до даних Телеграм-каналу була використана Telegram API ([8]). Для цього, а також для синтаксичного аналізу даних було розроблено програмну реалізацію на мові Python, код якої розміщено у відкритому доступі [10].

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 обличчя клоуна - стандартний набір реакції для користувача Телеграм. Смайл обличчя 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 у стандартизований формат, з яким 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 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).

Рис. 1: Результати тематичного аналізу навчального набору даних

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 повідомленнях, 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. 2: Word cloud часто використовуваних слів у позитивних повідомленнях

Для візуалізації результатів використали представлення word cloud, для яких характерна прямопропорційна 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.

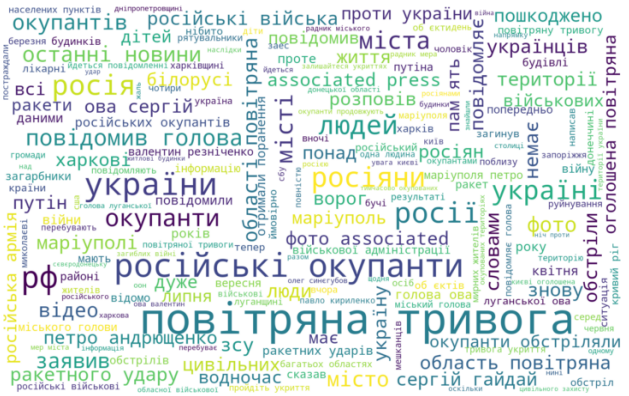

Рис. 3: Word cloud часто використовуваних слів у негативних повідомленнях

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 [11] 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, [12] - [15]).
При роботі методу опорних векторів [16] 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 відображається на отриманий простір та належить тій категорії, на бік якої відносно прогалини воно потрапило ([17] - [20]).

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:

Табл. 1: F-міри 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

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.

4 Подальші напрямки розвитку дослідження

Як було відмічено, в даному дослідженні порівнюються результати роботи для двох різних алгоритмів з видаленням та невидаленням зашумлень. У підсумку отримуємо, що F-міри для цих випадків різняться в сотих. Очевидно, що подальші дослідження мають стосуватись принципового покращення F-мір для запропонованих алгоритмів.

Однми із перспективними напрямками є:

- збільшувати навчальні вибірки для тренування алгоритмів аналізу новин, причому як для запропонованого каналу, так і спроби навчання на новинах з інших Телеграм-каналів;
- використовувати для попередньої обробки алгоритм word2vec, який використовує нейромережну модель для навчання пов'язаностей слів із великого корпусу тексту й виявлення слів-синонімів;
- дослідження можливості покращення результатів за рахунок використання нейронних мереж різних типів;
- використання трансормерів: моделей глибинного навчання, які базуються на використанні механізму уваги: роздільно зважуючи важливість кожної частини даних входу;
- зміна цільової функції: оскільки на даний момент вхідні датасет (навчальний та тестовий) ділиться за типами реакції на два умовні класи, а потім відбувається класифікація, то можна модифікувати цей підхід, щоб спочатку обчислювався прогноз відсотку кожної реакції та вже потім порівнювати отримані результати.

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

Дана робота виконана за підтримки Міністерства освіти та науки України у рамках виконання спільного українсько-литовського науково-дослідного проєкту "Моделювання ролі людського потенціалу для забезпечення оборони країни під час новітніх загроз"(Договір № М/27-2022 від 23.05.2022).

Список використаних джерел

1. *Mondal A.* Chatbot: An automated conversation system for the educational domain / A. Mondal , M. Dey, D. Das, S. Nagpal, K. Garda // International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP). — 2018. — P. 1-5.

2. *Karimpour D.* User recommendation based on Hybrid filtering in Telegram messenger / D. Karimpour, M.A.Z. Chahooki , A. Hashemi // 26th International Computer Conference, Computer Society of Iran (CSICC). — 2021. — P. 1-7.

3. *Hashemi A.* Telegram group quality measurement by user behavior analysis / A. Hashemi, M.A. Zare Chahooki // Social Network Analysis and Mining. — 2019. — № 9(1). — P. 1-12.

4. *Karimpour D.* User recommendation in Telegram messenger by graph analysis and mathematical modeling of users' behavior / D. Karimpour , M.A. Zare Chahooki , A. Hashemi // Journal of Information and Communication Technology. — 2021. — № 49. — P. 151-172.

5. *Eichstaedt J. C.* Facebook language predicts depression in medical records / J.C. Eichstaedt, R.J. Smith, R.M. Merchant, L.H. Ungar, P. Crutchley, D. Preoiuc-Pietro, H.A. Schwartz // Proceedings of the National Academy of Sciences. — 2018. — №115(44). — P. 11203-11208.

6. *Kachamas P.* Application of artificial intelligent in the prediction of consumer behavior from Facebook posts analysis / P. Kachamas , S. Akkaradamrongrat, S. Sinthupinyo, A. Chandrachai// International Journal of Machine Learning and Computing. — 2019. — №9(1). — P. 91-97.

7. *Essien A.* A deep-learning model for urban traffic flow prediction with traffic events mined from twitter / A. Essien , I. Petrounias , P. Sampaio , S. Sampaio // World Wide Web. — 2021. — №24(4). — P. 1345-1368.

8. Telegram APIs. Режим доступу: https://core.telegram.org/

9. Телеграм-канал ТСН. Режим доступу: https://t.me/TCH_channel

10. Код програмної реалізації. Режим доступу: https://github.com/KosukhaOlexandr/reactions_prediction/blob/main/clear_dataset.py

11. *Mosteller F.* Inference and disputed authorship: The Federalist / F. Mosteller , D.L. Wallace // Stanford Univ Center for the Study. — 2007.

12. *Козак Є. Б.* Принципи впровадження моделей машинного навчання у сфері інтелектуального обслуговування промислового обладнання /Є.Б. Козак // Таврійський

науковий вісник. Серія: Технічні науки. — 2021. — №3. — С. 19-28.

13. *Білецький Т. П.* Прогнозування дефектів у програмному забезпеченні алгоритмами глибинного навчання CNN та RNN / Т. П. Білецький , Д. В. Федасюк // Науковий вісник НЛТУ України. — 2021. — №31(2). — С. 114-120.

14. *Ahmad F.* Prediction of slope stability using Tree Augmented Naive-Bayes classifier: Modeling and performance evaluation. /F. Ahmad, X.W. Tang, J.N. Qiu , P. Wrblewski , M. Ahmad , I. Jamil // Math. Biosci. Eng. — 2022. — №19. — P. 4526-4546.

15. *Kewsuwun N.* A sentiment analysis model of agritech startup on Facebook comments using naive Bayes classifier / N. Kewsuwun, S. Kajornkasirat //International Journal of Electrical & Computer Engineering (2088-8708. — 2022. — №12(3).

16. *Cortes C.* Support-Vector Networks / C. Cortes, V Vapnik //Machine Leaming. — 1995. — №20. — P.273-297.

17. *Jose C.* Local deep kernel learning for efficient non-linear svm prediction / C. Jose, P. Goyal, P. Aggrwal, M. Varma //In International conference on machine learning. — 2013. — P. 486-494.

18. *Покідін Д.* Економетрична модель Національного банку України для оцінки кредитного ризику банку та альтернативний метод опорних векторів / Д. Покідін // Вісник Національного банку України. — 2015. — №234. — С. 53-67.

19. *Верлань А. І.* Огляд та порівняння методів машинного навчання для розпізнавання гідроакустичних сигналів / А.І. Верлань , А.О. Олексій //Інфокомунікаційні та комп'ютерні технології. — 2022. — №1(03). — С. 296-306.

20. *Ramasamy L. K.* Performance analysis of sentiments in Twitter dataset using SVM models/ L. K. Ramasamy , S. Kadry , Y. Nam , M.N. Meqdad // Int. J. Electr. Comput. Eng. — 2021. — №11(3). — P. 2275-2284.

References

1. MONDAL A., DEY M., DAS D., NAGPAL S. and GARDA K. (2018) Chatbot: An automated conversation system for the educational domain. *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*. pp. 1-5.

2. KARIMPOUR D., CHAHOOKI M. A. Z. and HASHEMI A.(2021) User recommendation based on Hybrid filtering in Telegram messenger. *26th International Computer Conference, Computer Society of Iran (CSICC)*. pp. 1-7.

3. HASHEMI A. and ZARE CHAHOOKI M. A. (2019) Telegram group quality measurement by user behavior analysis. *Social Network Analysis and Mining*. 9(1). pp. 1-12.

4. KARIMPOUR D., ZARE CHAHOOKI M. A. and HASHEMI A. (2021) User recommendation in Telegram messenger by graph analysis and mathematical modeling of users' behavior. *Journal of Information and Communication Technology*. 49(49). pp. 151-172.

5. EICHSTAEDT J. C., SMITH R. J., MERCHANT R. M., UNGAR L. H., CRUTCHLEY P., PREOIUC-PIETRO D. and SCWARTZ H. A. (2018) Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*. 115(44). pp. 11203-11208.

6. KACHAMAS P., AKKARADAMRONGRAT S., SINTHUPINYO S. and CHANDRACHAI A. (2019) Application of artificial intelligent in the prediction of consumer behavior from Facebook posts analysis. *International Journal of Machine Learning and Computing*. 9(1). pp. 91-97.

7. ESSIEN A., PETRUONIAS I., Sampaio P. and SAMPAIO S. (2021) A deep-learning model for urban traffic flow prediction with traffic events mined from twitter. *World Wide Web*. 24(4). pp. 1345-1368.

8. Telegram APIs. Available from: https://core.telegram.org/

9. Телеграм-канал TCH. Available from: https://t.me/TCH_channel

10. Code of the program realization. Available from: https://github.com/KosukhaOlexandr/reactions_prediction/blob/main/clear_dataset.py

11. MOSTELLER F. and WALLACE D. L. (2007) Inference and disputed authorship: The Federalist. *Stanford Univ Center for the Study.*

12. KOZAK YE.B. (2021) Pryncypy vprovadzhennya mashynnogo navchannya v sheri intelektualnogo obslugovuvannya promuslovogo obladnannya. *Tavriiskyi naukovyi visnyk. Seria: Tehnichni nauky.* (3). pp. 19-28.

13. BILETSKIY T.P. and FEDASYK D.V. (2021) Prognozuvannya defektiv v programnomu zabezpechenni algorytmamu glubunnogo navchannya CNN ta RNN. *Naukovyi visnyk NLTU.* 31(2). pp. 114-120.

14. AHMAD F., TANG X. W., QIU J. N., WRBLEWSKI P., AHMAD M. and JAMIL I. (2022) Prediction of slope stability using Tree Augmented Naive-Bayes classifier: Modeling and performance evaluation. *Math. Biosci. Eng.* 19. pp. 4526–4546.

15. KEWSUWUN N. and KAJORNKASIRAT S. (2022) A sentiment analysis model of agritech startup on Facebook comments using naive Bayes classifier. *International Journal of Electrical & Computer Engineering* (2088-8708). 12(3).

16. CORTES C. and VAPNIK V. (15) Support-Vector Networks. *Machine Leaming.* 20. pp.273—297.

17. JOSE C., GOYAL P., AGGRWAL P. and VARMA M. (2013) Local deep kernel learning for efficient non-linear svm prediction. *In International conference on machine learning.* pp. 486–494.

18. POKIDIN D. (2015). Ekonometrychna model Nacionalnogo banku Ukrainy dlya ocinky kredytnogo ryzyku banku ta alternatyvnyi metod opornyh vectoriv. *Visnyk Nacionalnogo banku Ukrainy.* 234. pp. 53.

19. VERLAN A. I. and OLEKSII A. O. (2022) Oglyad ta porivnyannya methodiv mashynnogo navchannya dlya rozpiznavannya gidrroakustychnyh sygnaliv. *Informaciini ta kompyyterni tehnologii.* 1(03). pp. 296-306.

20. RAMASAY L. K., KADRY S., NAM Y. and MEQDAD M. N. (2021) Performance analysis of sentiments in Twitter dataset using SVM models. *Int. J. Electr. Comput. Eng.* 11(3). pp. 2275-2284.

Надійшла до редколегії 29.08.2022