УДК 004.94:519.72 **doi: 10.26906/SUNZ.2022.3.071**

Є. В. Івохін, М. Ф. Махно, В. О. Рець

Київський національний університет імені Тараса Шевченка, Київ, Україна

ПРО ОДИН СПОСІБ АНАЛІЗУ ТОНАЛЬНОСТІ ТЕКСТІВ ЗА ДОПОМОГОЮ ШТУЧНИХ НЕЙРОННИХ МЕРЕЖ

Анотація. У статті детально розглянуто теоретичні відомості основного підходу до тонального аналізу тексту, проведено дослідження на тему автоматизації цього процесу за допомогою машинного навчання та використання штучних нейронних мереж. Проаналізовано основний тип архітектури нейронних мереж для роботи з класифікацією тексту та визначено оптимальну конфігурацію для програмної реалізації аналізу тональності даних. Зроблено висновок, що серед можливих конфігурацій моделей штучних нейронних мереж найкраще виконують свою функцію мережі із двонаправленими LSTM шарами у вигляді GRU конфігурації. При цьому, точність результатів напряму залежить від наявних наборів даних. Наведено результати порівняння точності результатів тренувань на обраних датасетах. Описано структуру програмної реалізації роботи сконфігурованої моделі ШНМ. Наведено приклад її застосування.

Ключові слова: аналіз тональності текстів, штучні нейронні мережі, класифікатори, двонаправлена модель, алгоритм роботи

Вступ

Аналіз тональності (іншими словами, емоцій, сентиментів, настроїв) є галуззю обчислювальних досліджень, метою якої є класифікація текстів як позитивних, негативних, нейтральних або ж належних до певної тональної групи [1]. Існуючі методи ефективних обчислень та аналізу настроїв можна розділити на три категорії: методи, засновані на знаннях, статистичні методи та гібридні методи. Технологія, заснована на знаннях, поділяє тексти на емоційні категорії на основі наявності досить відвертих емоційних слів. Поширені джерела слів-сентиментів або багатослівних виразів включають ефективний лексикон [2], схему лінгвістичних анотацій, WordNet-Affect, SentiWordNet, SenticNet [3] та ймовірнісні бази знань [4]. Основна слабкість методів, заснованих на знаннях, полягає в тому, що вони погано розуміють емоції, коли йдеться про мовні правила [5]. Більше того, дійсність методів, заснованих на знаннях, значною мірою залежить від глибини та широти задіяних ресурсів. Статистичні методи, такі як машини опорних векторів і глибоке навчання (англ. deep learning), широко використовувалися для класифікації настроїв тексту, тому дослідники використовували їх, наприклад, для таких проектів, як класифікатори відгуків на фільми [6]. Забезпечуючи велику кількість навчальних наборів з текстом емоційних анотацій для алгоритмів машинного навчання, така система може не тільки вивчати емоційну валентність ключових слів афекту, а й враховувати валентність інших ключових слів і частоти спільного зустрічання слів. Слід зауважити, що ці методи є ефективними для категоризації тексту користувача лише на рівні сторінки або абзацу та погано працюють з меншими текстовими одиницями, такими як речення або слово. Гібридний підхід використовує як методи, засновані на знаннях, так і статистичні методи для виконання таких завдань, як розпізнавання емоцій та виявлення полярності з текстових або мультимодальних даних. Традиційні методи класифікації настроїв в основному засновані на правилах і машинному навчанні. Підходи, засновані на правилах, в основному використовують лексикони тональностей, шаблони та статистичні характеристики, отримані з досвіду чи експертних думок, для класифікації настроїв тексту, ця техніка зазвичай вимагає значного ручного втручання [7]. А метод машинного навчання розглядає аналіз настроїв як проблему класифікації, він спочатку створює навчальний набір, позначаючи вручну частину даних, потім витягує та вивчає навчальні дані для побудови моделі класифікації, яка використовується для класифікації та прогнозування тестових даних з невідомими тегами.

Одним з найбільш ефективних підходів до розв'язання задачі класифікації тексту є використання методу глибокого (машинного) навчання, що проводиться на основі створення нейронної мережі для класифікації тексту.

Найбільш часто використовуваними моделями глибокого навчання в задачах аналізу настроїв є згорткові нейронні мережі (CNN) і повторювані нейронні мережі (RNN). CNN дозволяє виділяти високорозмірні особливості між локально сусідніми словами, використовуючи різні розміри "розсувних вікон" для векторів слів усіх слів речення. Однак, фільтр CNN має обмежену ємність слів і не може вловити довгострокові залежності, тому він, як правило, не дозволяє отримати семантичний зв'язок між несуміжними словами в реченні. На відміну від CNN, мережі RNN призначені для моделювання послідовності з контекстним семантичним захопленням, яке може застосувати вміст пам'яті до поточного сценарію. Завдяки таким функціям мережі RNN частіше використовуються для класифікації тексту. Однак для довгої послідовності даних традиційні нейронні мережі RNN можуть викликати "вибух" градієнта або його зникнення. Мережі довготривалої пам'яті LSTM (Long Short-Term Memory) [8], є особливим типом RNN, які дозволяють вивчати довгострокові залежності та використовувати ефект довготривалої короткочасної пам'яті в якості прихованих одиниць. LSTM мережі враховують залежність порядку між послідовностями слів, тому вони здатні залучати до розрахунків залежності як на близькі відстані між словами, так і на великі. Враховуючи потужність мережі в обробці довгих текстових повідомлень, LSTM відіграють велику роль в обробці

природної мови (NLP). Базові LSTM мережі сканують послідовності лише в одному напрямку, двонаправлена довготривала пам'ять (BiLSTM) [9] є її подальшим вдосконаленням, за яким сканування послідовності здійснюється в обох напрямках, забезпечуючи одночасний доступ як до прямого, так і до зворотного контекстів. Тому BiLSTM може вирішувати задачі моделі послідовності краще, ніж LSTM. Ці моделі нейронних мереж досягли великого успіху в завданні аналізу емоційної класифікації.

Розробка підходу до розв'язання задачі аналізу тональності

Для проведення неперервного аналізу та класифікації текстових повідомлень необхідно створити методику ефективної обробки та оцінки емоційного забарвлення (тональності) текстів. В якості конструктивного підходу пропонується використовувати BiLSTM архітектуру нейронної мережі у варіанті конфігурації з двонаправленим скануванням тексту. Реалізація у цьому випадку передбачає:

1) пошук датасетів, які містять ознаки емоційної класифікації з чотирма [10] або 6 класами [11];
2) утворення та тренування BiLSTM моделі на знайдених наборах даних;
3) порівняння ефективності використання BiLSTM з двома типами шарів – звичайним LSTM та з Gated Recurrent Unit (GRU) [12];
4) розробку програмних засобів для класифікації тональності вхідного тексту за допомогою отриманої нейронної мережі;
5) розробку сервісу по періодичному парсингу новинних ресурсів;
6) наповнення бази даних для збереження отриманих результатів;
7) створення мережі з використанням технологій брокера повідомлень для неперервної обробки текстів новин класифікатором емоцій;
8) написання демонстраційної веб-сторінки для візуалізації отриманих даних.

У якості тренувальних наборів даних було прийнято рішення використати англомовні датасети з додаванням автоматичного перекладу до етапів попередньої обробки тексту.

Серед наявних англомовних наборів даних найпопулярніші з них використовують бінарну класифікацію тональності тексту, тобто визначення лише "негативного" або "позитивного" емоційного забарвлення. Найповнішим датасетом такої класифікації є набір даних від IMDb [13], що включає в себе 50 тисяч рецензій на фільми з відповідною оцінкою. Очевидно, така класифікація не є достатньо повною для досягнення поставленої мети розробки, але саме цей набір даних використовується у більшості сучасних досліджень.

Для проведення повноцінного дослідження тональності необхідно мати набір даних з описом не менш, ніж 4 класів емоцій (у більшості випадків такі датасети не мали більше 3-4 тисяч класифікованих записів, що не дозволило б мати достатню точність під час тренування нейронної мережі). Кінцевими знахідками виявились два набори даних – на 20 та на 40 тисяч записів, перший з шістьма класами, другий з тринадцятьма. Обидва набори містять тексти повідомлень соціальної мережі Twitter, але перший має лише прості тексти з видаленими знаками пунктуації та літерами нижнього регістру, тоді як другий містить повноцінні повідомлення різної довжини. При цьому перший датасет є відносно збалансованим (рис. 1, 2)

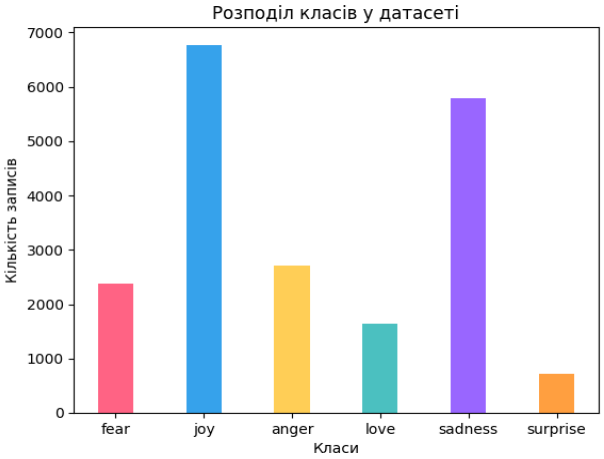

Рис 1: Структура навчального набора даних з шістьма класами

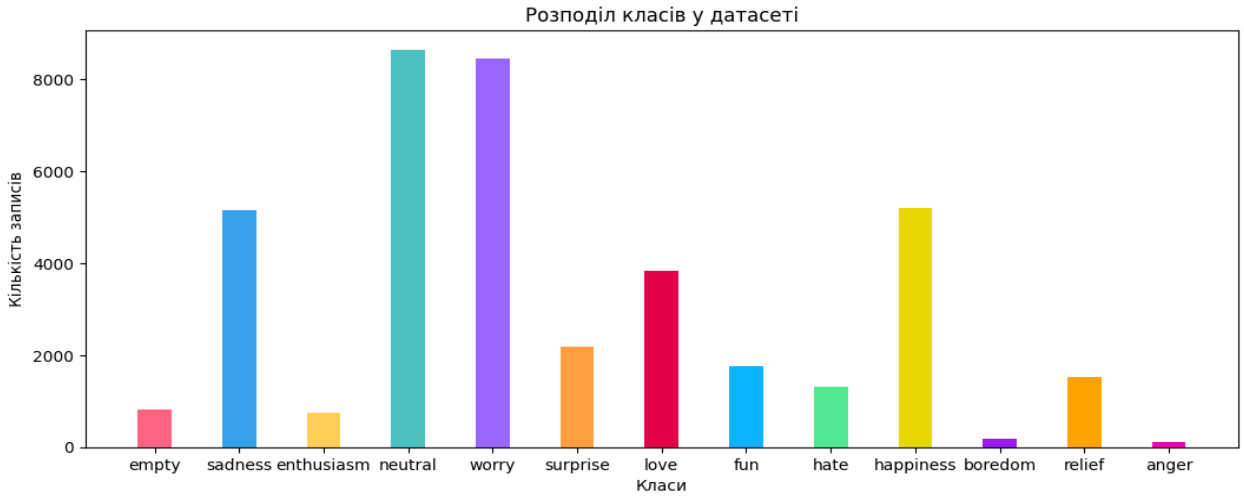

Рис 2: Структура навчального набора даних з тринадцятьма класами

Відповідно, точність тренування моделі на першому наборі матиме вищу точність, але менший спектр результатів, другий же датасет буде мати меншу точність але більшу кількість класів у результаті. Перед тренуванням мережі запропоновано проводити етапи попередньої обробки: видалення всіх спеціальних символів hex-формату; розбиття тексту на слова; видалення службових слів та артиклів; зведення слів до їх початкових форм; об'єднання слів назад в текст. Спростивши текст, маємо набір ключових слів в їх початкових формах, що полегшує роботу нейронної мережі, оскільки таким чином загальна архітектура мережі зводиться до однієї цілі – класифікації емоцій, замість додавання шарів для попередньої обробки і, як наслідок, ускладнення процесу тренування. В якості першого шару моделі, що обробляє текст, обрано вбудований шар, який виконує перетворення слів у векторну форму, тобто у вигляд, що найоптимальніше сприймається нейронною мережею [14]. Такий шар було отримано з попередньо розрахованого шару з 1 мільйоном оброблених слів ресурсу Wikipedia [15].

Архітектура моделі ШНМ для аналізу тональності тексту

Для повноцінного аналізу сформуємо архітектуру нейронної мережі, що складається з трьох основних шарів:

1. Вбудований шар для переведення слів у векторну форму
2. Двонаправлений шар з довгою короткочасною пам'яттю (BiLSTM)
3. Ущільнюючий шар для виведення результатів.

Для порівняння отриманих результатів проведено тренування на двох датасетах з використанням звичайного LSTM та GRU шарів, результати (точність тренування кожної конфігурації) представлені у табл. 1:

Таблиця 1 – **Порівняння точності результатів тренувань на обраних наборах даних**

Датасет \ Тип класифікатора	LSTM	GRU
20k записів	0.7290	0.7965
40k записів	0.3524	0.3696

З наведених результатів видно, що перший набір даних виконує свою функцію краще з точки зору тестування. При цьому, у випадку тестування на справжніх даних, натренована на другому наборі модель дає кращі результати. Це можна помітити, перевіривши роботу нейронних мереж (GRU конфігурації, оскільки вона має вищу точність тренування) на текстах новин і художніх текстів (табл. 2).

Таблиця 2 – **Приклад застосування ШНМ з GRU класифікатором**

Тестовий текст \ Модель	**Мережа 1 датасету**	**Мережа 2 датасету**
Латвія за два місяці віддала Україні третину оборонного бюджету, - президент Левітс	Joy (радість)	Neutral (нейтральність)
Російське МЗС ввело санкції у відповідь проти США. Їх запровадили проти всіх 398 членів Палати представників – нижньої палати Конгресу	Joy (радість)	Neutral (нейтральність)
Європейський союз схвалив подальшу військову допомогу Україні на загальну суму 1,63 мільярда доларів. Новий пакет фінансуватиме забезпечення Збройних Сил України обладнанням та предметами постачання, у тому числі засобами індивідуального захисту, аптечками та пальним, а також військовою технікою, що призначена для доставки летального озброєння в оборонних цілях.	Joy (радість)	Worry (занепокоєння)
Максимум квітів – від немофіл до сакури – в одному кадрі чекає туристів у парку Хіояма в японському місті Сімоносекі. Заодно можна помилуватися величезний міст, що з'єднує острови Хонсю і Кюсю.	Surprise (здивування)	Happiness (щастя)
Коли сакура опадає, це теж страшенно красиво. Земля або вода, рівномірно всипані рожевими пелюстками, чудові. А найкраще йти під дощем із пелюсток і відчувати, як краса буквально тече крізь тебе.	Sadness (смуток)	Love (любов)

Недоцільні дані при використанні, здавалося б, більш точної моделі першого датасету, пояснюються природою його даних: внаслідок вже проведеної попередньої обробки вони могли втратити попередній сенс та/або могли бути сконцентровані на іншій тематиці за рахунок меншої кількості записів у наборі даних у порівнянні з другим набором. Низька точність тренування другого датасету, в свою чергу, може бути спричинена незбалансованістю набору даних, але завдяки різноманіттю тем цей набір дає змогу натренувати та використати нейронну мережу для класифікації текстів різної тематики. Таким чином, для практичного використання у системі аналізу тональності було обрано нейронну мережу, натреновану на другому датасеті з шаром GRU.

Програмна реалізація роботи моделі ШНМ

Виходячи, з того, що програмна система, яка моделює роботу отриманої ШНМ складається з декількох незалежних сервісів, було прийнято рішення

розробити її засобами оточення Docker Compose [16], що дозволяє побудувати Docker контейнери для кожного описаного сервісу, після чого об'єднати їх в ізольовані мережі. Розробка на основі Docker контейнерів – в даному випадку найшвидший спосіб розробити сервіси, які в можуть бути встановленими у середовищі будь-якої серверної архітектури за мінімальний час і які можуть дублюватися для прискорення обробки великих наборів вхідних даних.

Серед запропонованих сервісів є три основні компоненти: парсер новин, що збирає тексти новин та їх метадані для подальшої обробки; класифікатор тональності отриманих текстів; візуалізатор оброблених даних. Для обміну даними між парсером та класифікатором було реалізовано сервіс-брокер повідомлень RabbitMQ в окремому контейнері, що забезпечує збереження вхідних даних (в цьому випадку текстів новин) до моменту обробки їх класифікатором. В свою чергу, оброблені дані зберігаються у базі даних PostgreSQL, яка теж реалізована в окремому контейнері. Візуалізація відбувається завдяки поєднанню двох сервісів: бекенд-частини системи на NodeJS + Fastify, що представляє собою сервер з доступом до бази даних і з відкритим інтерфейсом (API) для доступу даних у базі даних; фронтенд-частини на NodeJS + ReactJS, який завдяки функціональній архітектурі та реалізації отримання даних з бекенду на основі хуків глобального стану подає накопичені дані у вигляді графіків відповідно до обраного джерела новин.

Висновки

В роботі детально розглянуто теоретичні відомості основного підходу до тонального аналізу тексту, проведено дослідження на тему автоматизації цього процесу за допомогою машинного навчання та використання штучних нейронних мереж.

Проаналізовано основний тип архітектури нейронних мереж для роботи з класифікацією тексту та визначено оптимальну конфігурацію для програмної реалізації аналізу тональності даних.

Зроблено висновок, що серед можливих конфігурацій моделей штучних нейронних мереж найкраще виконують свою функцію мережі із двонаправленими LSTM шарами у вигляді GRU конфігурації. При цьому, точність результатів напряму залежить від наявних наборів даних. Наведено результати порівняння точності результатів тренувань на обраних датасетах. Описано структуру програмної реалізації роботи зконфігурованої моделі ШНМ. Наведено приклад її застосування.

СПИСОК ЛІТЕРАТУРИ

1. *Cambria, Erik, et al.* A practical guide to sentiment analysis. – 2017.
2. *Ortony, Andrew, Gerald L. Clore, and Allan Collins.* The cognitive structure of emotions. Cambridge university press, 1990.
3. *Cambria, Erik, Daniel Olsher, and Dheeraj Rajagopal*. SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis// Twenty-eighth AAAI conference on artificial intelligence, 2014.
4. *Swapna Somasundaran, Janyce Wiebe, and Josef Ruppenhofer*. Discourse level opinion interpretation. In Proc. of the 22nd Int. Conf. on Comp. Linguistics, 2008. Vol.1 (COLING '08). Association for Computational Linguistics, USA, 801–808.
5. *S. Poria, E. Cambria, A. Gelbukh, F. Bisio, A. Hussain*. Sentiment data flow analysis by means of dynamic linguistic patterns, IEEE Comput. Intell. Mag. 10 (4). – 2013.
6. *R.Y. Lan, Y. Xia, Y. Ye*. A probabilistic generative model for mining cybercriminal networks from online social media, IEEE Comput. lntell. Mag. 9 (4). – 2014. – Pp. 31—43.
7. *Chesley, Paula, et al*. Using verbs and adjectives to automatically classify blog sentiment// Training 580.263 (2006): 233.
8. *Zhu, Xiaodan, Parinaz Sobihani, and Hongyu Guo.* Long short-term memory over recursive structures// Proc. International Conference on Machine Learning. PMLR, 2015.
9. *Alex Graves, Jürgen Schmidhuber*. Framewise phoneme classification with bidirectional LSTM and other neural network architectures// Neural Networks, 2005. – Vol. 18. – Iss.5–6.
10. *Izard, Carroll E.* The substrates and functions of emotion feelings: William James and current emotion theory// Personality and Social Psychology Bulletin 16.4 (1990): 626-635.
11. *Ekman, Paul*. "Emotions revealed." Bmj 328.Suppl S5 (2004).
12. *Yao, Kaisheng, et al*. Depth-gated recurrent neural networks// arXiv:1508.03790 9 (2015).
13. IMDb Datasets [Online] – Available from: https://www.imdb.com/interfaces/
14. *Luong, Minh-Thang, Richard Socher, and Christopher D. Manning.* Better word representations with recursive neural networks for morphology// Proceedings of the seventeenth conference on computational natural language learning. 2013.
15. English word vectors[Online] – Available from: https://fasttext.cc/docs/en/english-vectors.html
16. *Ian Miell, Aidan Hobson Sayers.* Docker in Practice. - Manning Publications. – 2019.

Received (Надійшла) 10.07.2022
Accepted for publication (Прийнята до друку) 24.08.2022

About one way to analyze the sentiment of texts using artificial neural networks

E. Ivokhin, M. Makhno, V. Rets

Abstract. The article discusses in detail the theoretical information of the main approach to tone analysis of the text, conducted research on the topic of automating this process using machine learning and the use of artificial neural networks. The main type of architecture of neural networks for working with text classification is analyzed and the optimal configuration for the implementation of data sentiment analysis is determined. It is concluded that among the possible configurations of artificial neural network models, networks with bidirectional LSTM layers in the form of a GRU configuration perform their function better. At the same time, the accuracy of the results directly depends on the available data sets. The results of comparing the accuracy of training results on the selected datasets are presented. The structure of the software implementation of the configured ANN model is described. An example of its application is given.

Keywords: text sentiment analysis, artificial neural networks, classifiers, bidirectional model, operation algorithm.