https://doi.org/10.15407/csc.2024.04.034
УДК 681.3.062

О.О. МАРЧЕНКО, доктор фіз.-мат. наук, професор, зав. відділом,
Міжнародний науково-навчальний центр інформаційних технологій і систем НАН та МОН України,
просп. Академіка Глушкова, 40, Київ, Україна, 03187,
ORCID: https://orcid.org/0000-0002-5408-5279,
omarchenko@univ.kiev.ua

Е.М. НАСІРОВ, кандидат фіз.-мат. наук, ст. наук. співробітник,
Міжнародний науково-навчальний центр інформаційних технологій і систем НАН та МОН України,
просп. Академіка Глушкова, 40, Київ, Україна, 03187,
ORCID: https://orcid.org/0009-0006-9016-2602,
enasirov@gmail.com

Д.О. ВОЛОШЕНЮК, кандидат техн. наук, ст. досл., зав. лаб.,
Міжнародний науково-навчальний центр інформаційних технологій і систем НАН та МОН України,
просп. Академіка Глушкова, 40, Київ 03187, Україна,
ORCID: https://orcid.org/0000-0003-3793-7801,
p-h-o-e-n-i-x@ukr.net

ФОРМУВАННЯ УКРАЇНОМОВНОЇ НАВЧАЛЬНОЇ ВИБІРКИ ДЛЯ ВИЗНАЧЕННЯ ЕМОЦІЙНОГО ЗАБАРВЛЕННЯ ТЕКСТІВ

Проведено аналіз наявних тональних словників української мови. Створено розширений тональний словник української мови. Для поєднання різних стилістичних забарвленостей та особливостей мови зібрано набір текстів українською мовою з різних типів джерел. Побудовано вибірку текстів українською мовою для навчання нейромереж для визначення емоційного забарвлення текстів, визначивши тональність текстів зібраного набору застосувавши статистичні моделі.

Ключові слова: *штучний інтелект, комп'ютерна лінгвістика.*

Вступ

Щодня зростає кількість новин, сторінок у соціальних мережах та чатів в мережі Інтернет, відповідно відбувається збільшення інформації, яка несе емоційне навантаження. При цьому зростає і кількість інформаційних загроз.

За цих умов побудова систем визначення емоціонального забарвлення текстів стає надзвичайно актуальною.

Саме потоки новин створюють основу для надлишкової соціальної напруженості, не зумовленої безпосереднім досвідом та спостереженнями людей. Завдяки значному поширен-

Cite: Марченко О.О., Насіров Е.М., Волошенюк Д.О. Формування україномовної навчальної вибірки для визначення емоційного забарвлення текстів. *Control Systems and Computers*, 2024, 4, 34—38. https://doi.org/10.15407/csc.2024.04.034

ню засобів електронної комунікації люди самі, навіть без новинних агенцій, формують потоки власних новин, що можуть швидко поширюватися, і мати властивості "вірусності" та "меметичності".

Окремі особи (для позначення яких часто вживають напів-формальний термін "інфлюенсер" та більш формальний термін "лідер суспільної думки") можуть мати вплив, співмірний із впливом засобів масової інформації, оскільки легше передати емоційний сигнал, у тому разі, якщо людина, яка його передає, є цікавою для читача (слухача чи глядача).

Більшість інформації є доступною у текстовому вигляді. Хоча візуальність сприйняття значно переважає над текстовою складовою, проте все одно під зображенням майже завжди є коментарі, як мінімум від автора повідомлення, а часто й від аудиторії (у соціальних мережах). Для відеозаписів зазвичай доступні субтитри.

Для успішної боротьби з привнесеною соціальною напруженістю необхідно прослідковувати емоційну складову новин різних видів та з різних джерел.

Постановка задачі

Емоційний посил в повідомленнях можна віднайти та класифікувати за допомогою засобів штучного інтелекту. Існують методи визначення емоційного забарвлення текстів, що базуються на статистичних методах та глибокому навчанні з використанням нейромереж. Хоча нейромережі показують значно кращі результати в задачах кластеризації текстів, для їх навчання необхідно мати навчальну вибірку текстів з попередньою оцінкою їх емоційного забарвлення. Такі розмічені навчальні вибірки існують для новин та текстів англійською мовою, проте, на даний момент, не створено доступної навчальної вибірки україномовних новин та текстів.

Тому було поставлено задачу збору великого набору текстових даних українською мовою, що міститиме тексти різного типу(новини, користувацькі повідомлення) та різних тематик. Для текстів з зібраного набору текстів необхідно визначити їхнє емоційне забарвлення.

Методи визначення емоційного забарвлення текстів

Емоційне забарвлення може мати такі значення:

- *positive* — слова чи словосполучення, що самостійно виражають позитивне значення;
- *negative* — слова чи словосполучення, що самостійно виражають негативне значення;
- *neutral* — слова чи словосполучення, що самостійно не виражають ні позитивного, ні негативного значення.

Аналіз тональності може бути розділений на дві окремі категорії:

- ручний (або аналіз тональності експертами);
- автоматизований.

В автоматизованих системах автоматизованого аналізу тональності застосовують алгоритми машинного навчання, інструменти статистики і обробки природної мови, що дозволяє обробляти великі масиви тексту, включаючи веб-сторінки, онлайн-новини, тексти користувацьких груп в мережі Інтернет, соціальні медіа та інше.

Існують такі автоматизовані підходи до аналізу тональності:

- підходи, засновані на правилах;
- підходи, засновані на словниках;
- підходи, засновані на машинному навчанні з учителем;
- підходи, засновані на машинному навчанні без учителя.

Методи, засновані на правилах і словниках базуються на пошуку емотивної лексики (лексичної тональності) в тексті по заздалегідь складеним тональним словників і правилам із застосуванням лінгвістичного аналізу. За сукупністю знайденої емотивної лексики текст може бути оцінений за шкалою, що містить кількість негативної та позитивної лексики.

Таблиця 1. **Тональність слів країнської мови**

Слово	Тональність
агресія	–1
бездоганно	2
багно	–3
волонтер	1
банкрут	–1

Даний метод може використовувати як списки правил, що вставляються у регулярні вирази, так і спеціальні правила для аналізу зв'язків тональної лексики всередині речення. Аналіз тексту складається з наступних кроків:

• кожному слову тексту присвоюється його тональне значення зі словника (якщо воно є у словнику);

• обчислюється загальна тональність тексту шляхом підсумовування тональних значень кожного окремого речення.

Основна проблема методів, що базуються на словниках і правилах, полягає у високій трудомісткості створення словника. Для досягнення високої точності класифікації документа терміни словника мають відповідати предметній області документа і мати відповідну вагу.

Методи, що засновані на машинному навчанні. В основі цього підходу лежить ідея, що терміни, які найчастіше зустрічаються в цьому тексті і в той же час присутні в невеликій кількості текстів у всій колекції, мають найбільшу вагу.

Виділивши ці терміни, а потім визначивши їх тональність, можна зробити висновок про тональності всього тексту.

В методах машинного навчання з вчителем використовується навчальна вибірка розмічених заздалегідь текстів для тренування. Проте, за відсутності такої навчальної вибірки, данні методи не можуть бути застосовані.

У випадку методів машинного навчання без вчителя для тренування алгоритму використовується навчальна вибірка нерозмічених заздалегідь текстів. При такому підході найбільшу вагу отримують терміни, що найбільш часто зустрічаються в тексті, але, які при цьому присутні тільки в обмеженій кількості текстів всієї множини. Недоліком таких методів — є низька точність.

Набір текстів

Щоб покрити більшу кількість стилістичних забарвленостей та особливостей мови, було поєднано різні джерела як базовий набір текстів для розмітки, а саме:

1. новини з український ресурсів;

2. повідомлення користувачів соціальних мереж та месенджерів;

3. відгуки на товари з інтернет-магазинів.

У такий спосіб було отримано набір, що містить 6605448 текстів українською мовою.

Тональний словник

Для автоматичної розмітки текстів було вирішено використовувати словник тонального забарвлення(валентності) слів та частотний аналіз входження таких слів у текстах.

У найпростішому вигляді тональний словник містить список слів і словосполучень зі значенням тональності для кожного слова.

Як базовий словник було використано Український тональний словник [1], що містить 3442 слів української мови, які мають не нейтральну тональність (–2, –1, 1, 2) (Табл. 1). Словник наповнено з двох джерел: слова оцінені експертами та згенеровані автоматично за допомогою алгоритмів машинного навчання, а також із використанням векторів слів *word2vec* та *lex2vec*. Усі слова було приведено до нижнього регістру та зведено до нормальної форми.

Тональний словник було доповнено словами з доступного в мережі Інтернет словника [2], що дало змогу додати ще 3319 слів.

Використано англомовний тональний словник *NRC Emotion Lexicon* [3], також доступний у вигляді перекладу на декілька мов, включно з українською. Попри деякі культурні відмінності, авторами було показано, що більшість афективних норм є сталими в різних мовах. Таким чином, перекладаючи англійські терміни за допомогою *Google Translate*, було сформовано словники понад 100 мовами. Словник містить 14155 слів із вісьмома основними емоціями та почуттями, з яких 8903 є — емоційно забарвленими як негативні чи позитивні.

У такий спосіб, після об'єднання трьох словників було отримано тональний словник української мови, що містить 8730 унікальних слів, які мають не нейтральну тональність.

Розмітка текстів

Для якісної розмітки текстів було виконано декілька кроків попередньої обробки, а саме:

1. видалення *HTML* тегів;

2. видалення пунктуації та спецсимволів;
3. зведення до нижнього регістру;
4. видалення стоп-слів;
5. зведення слів до нормальної форми.

Після виконання попередньої обробки було визначено емоційне забарвлення кожного тексту.

Оскільки попередньої розмітки, яка б дозволила виконати розмітку методами машинного навчання, немає, було використано методи, що засновані на правилах і словниках.

Ці методи базуються на пошуку емотивної лексики(лексичної тональності) в тексті за заздалегідь складеними тональними словниками та правиламии із застосуванням лінгвістичного аналізу. За сукупністю знайденої емотивної лексики текст може бути оцінений за шкалою, що визначає кількість негативної та позитивної лексики.

Нами було обрано модель *VADER* (*Valence Aware Dictionary for sEntiment Rea-soning*) [4] на основі правил для загального аналізу настроїв, що використовує комбінацію якісних і кількісних методів. Лексичні особливості тонального словника поєднуються з урахуванням п'яти загальних правил, які втілюють граматичні та синтаксичні підходи для вираження та акцентування інтенсивності почуттів.

Дана модель доступна в пакеті *NLTK* [5] для *Python* і демонструє високу точність [4]. Для розмітки корпусу текстів модель було наповнено словами та їхньою тональністю з побудованого словника.

Оцінка отриманої розмітки

При використанні лише Українського тонального словника [1] було отримано 4561563 текстів із визначеною емоційністю. Об'єднавши три словники, вдалося збільшити кількість розмічених текстів на 16.6 %, у такий спосіб було отримано 5318783 текстів, з яких 2933008 позитивних та 2385775 негативних. Отже, вдалось визначити тональність для 80 % текстів.

Для перевірки отриманої розмітки текстів було виконано експертну оцінку. Для цього випадковим чином було обрано 100 текстів, емоційне забарвлення яких було визначено експертом. Порівняння отриманих результатів показало точність розмітки у 98 %, що є достатнім для якісного навчання нейронних мереж.

Висновки

Здійснено аналіз методів статистичного визначення тональності текстів, а також аналіз наявних тональних словників української мови та побудовано об'єднаний тональний словник, що містить 8730 слів.

У такий спосіб, було побудовано великий корпус текстів та їхнього емоційного забарвлення з експертно оціненою точністю розмітки у 98 %, що містить 5318783 різних типів текстів українською мовою. Побудований текстовий корпус може бути застосовано для навчання та тестування нейронних мереж з метою визначення емоційного забарвлення текстів.

Дослідження підготовлено за грантової підтримки Національного фонду досліджень України в рамках реалізації проєкту "Інформаційна технологія визначення тональності та класифікації текстової інформації для виявлення інформаційних загроз" (реєстраційний номер 2023.04/0053). в рамках конкурсу "Наука для зміцнення обороноздатності України".

ЛІТЕРАТУРА

1. Український тональний словник. URL: https://github.com/lang-uk/tone-dict-uk (дата звернення: 05.02.2024).
2. Ukrainian-Sentiment-Analysis. URL: https://github.com/skupriienko/Ukrainian-Sentiment-Analysis (дата звернення: 08.05.2024).
3. Mohammad S., Turney P. Crowdsourcing a Word-Emotion Association Lexicon, Computational Intelligence. 2013. 29 (3). pp. 436–465.
4. Hutto, C.J. & Gilbert, E.E. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

5. Bird S, Klein E, Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. «O'Reilly Media, Inc.» 2009.

Надійшла 01.11.2024

REFERENCES

1. Ukrainian sentiment vocabulary. [online] Available at: <https://github.com/lang-uk/tone-dict-uk> [Accessed: 05 Feb. 2024].
2. Ukrainian-Sentiment-Analysis. [online] Available at: <https://github.com/skupriienko/Ukrainian-Sentiment-Analysis> [Accessed: 08 May 2024].
3. Mohammad, S. and Peter Turney, P. (2013). "Crowdsourcing a Word-Emotion Association Lexicon", *Computational Intelligence*, 29 (3), pp. 436–465.
4. Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Eighth International Conference on Weblogs and Social Media* (ICWSM-14). Ann Arbor, MI, June 2014.
5. Bird S, Klein E, Loper E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. «*O'Reilly Media, Inc*».

Received 01.11.2024

*O.O. Marchenk*o, Doctor (Physical and Math.), Professor, Head of the Department, International Research and Training Center for Information Technologies and Systems NAS and MES of Ukraine,
Glushkov ave., 40, Kyiv, Ukraine, 03187,
ORCID: https://orcid.org/0000-0002-5408-5279,
omarchenko@univ.kiev.ua

E.M. Nasirov, PhD (Physical and Math.), senior Researcher, International Research and Training Center for Information Technologies and Systems NAS and MES of Ukraine,
Glushkov ave., 40, Kyiv, Ukraine, 03187,
ORCID: https://orcid.org/0009-0006-9016-2602,
enasirov@gmail.com

D.O. Volosheniuk, PhD (Eng.), Head of the Laboratory, International Research and Training Center for Information Technologies and Systems NAS and MES of Ukraine,
Glushkov ave., 40, Kyiv, Ukraine, 03187,
ORCID: https://orcid.org/0000-0003-3793-7801,
p-h-o-e-n-i-x@ukr.net

BUILDING THE UKRAINIAN-LANGUAGE TRAINING DATASET FOR DETERMINING THE SENTIMENT ANALYSIS OF TEXTS

Introduction. Every day, the number of news, pages on social networks and chats on the Internet is increasing, accordingly, there is an increase in information that carries an emotional load. At the same time, the number of information threats is also growing. Under such conditions, the construction of systems for determining the emotional color of texts becomes extremely relevant.

Purpose. Emotional messages can be found and classified using artificial intelligence, namely based on neural network methods. For the process of learning neural networks, it is necessary to have a training sample of texts with a preliminary assessment of their emotional coloring. Such marked learning samples exist for news and texts in English, however, at the moment, no accessible learning sample of Ukrainian news and texts has been created.

Methods. Using statistical methods of sentiment analysis for detecting text tonality with an extended vocabulary.

Results. Extended tonality vocabulary of the Ukrainian language was built. A large corpus of texts and their emotional coloring was constructed with an expertly assessed markup accuracy of 98%, containing 5,318,783 texts of various types in the Ukrainian language.

Conclusion. The built text corpus can be used to train and test neural networks for sentiment analysis of Ukrainian-language texts.

***Keywords**: artificial intelligence, computational linguistics.*