ISSN 2224-087X. Електроніка та інформаційні технології. 2021. Випуск 15. С. 16–23
Electronics and information technologies. 2021. Issue 15. P. 16–23

УДК 004.932 DOI: https://doi.org/10.30970/eli.15.2

СИСТЕМА АВТОМАТИЧНОГО ВИЗНАЧЕННЯ ТОНАЛЬНОСТІ ТЕКСТУ

І. Оленич, М. Притула, О. Сінькевич, О. Хамар

Львівський національний університет імені Івана Франка,
вул. Драгоманова, 50, 79005 Львів, Україна
iolenych@gmail.com

У роботі запропоновано систему емоційного аналізу україномовних текстів, яка ґрунтується на словниках і правилах. Проведено комп'ютерний аналіз текстової інформації за п'ятьма емоційними категоріями: дуже позитивним, позитивним, нейтральним, негативним і дуже негативним. Врахування емоційного навантаження різних частин мови та розширення словникової бази даних за допомогою словника синонімів дає змогу підвищити точність і обґрунтованість сентимент-аналізу. У результаті агрегування даних щодо різних емоційних критеріїв засобами нечіткого моделювання одержано кількісну оцінку тональності текстової інформації.

Ключові слова: комп'ютерний аналіз тексту, сентимент-аналіз, токенізація, словник тональності, нечітке моделювання.

Розвиток інформаційних технологій призвів до значного збільшення обсягів інформації, яка підлягає зберіганню, обробці та передачі за допомогою засобів комп'ютерних систем та мереж. Текст є однією з основних форм обміну інформацією у суспільстві. Текстова інформація у різних форматах складає значну частку інформаційних ресурсів комп'ютерних систем. У сучасному світі, де обмін текстовими повідомленнями відбувається по усіх можливих каналах спілкування, важливим є оперативний аналіз переданої інформації. Завдяки аналізу текстової інформації комерційні компанії мають змогу відслідковувати потреби користувачів та вчасно реагувати на їх відгуки про свої продукти та послуги [1].

Особливу увагу привертає проблема аналізу думок користувачів Інтернету, який ґрунтується на виділенні та подальшому розпізнаванні емоційно забарвленої лексики у тексті. Здебільшого такий аналіз застосовується до клієнтських текстів, таких як відгуки та відповіді на опитування, соціальні медіа та мережі, а також різноманітні програми, які варіюються від маркетингу до обслуговування клієнтів [2–4]. На основі проведеного сентимент-аналізу оцінюють реакцію на рекламу, прогнозують фондові ринки, здійснюють моніторинг якості життя у реальному часі, що дає змогу попереджувати небезпечні ситуації у суспільстві [5–7]. Застосування низки різноманітних IT-інструментів допомагає вченим не тільки зрозуміти колективну поведінку та оцінити настрої громадської реакції на маркетингові заходи, але й здійснювати вплив на соціальні системи та розробляти засади державної політики у різних галузях [8,9]. З огляду на значний потенціал практичного застосування створення і розвиток технологій обробки тексту є актуальною задачею на усіх етапах розвитку інформаційних систем.

Аналіз емоційного забарвлення тексту, як і багато інших задач обробки природної мови, можна розглядати як задачу класифікації, у якій повинно бути вирішено два завдання: класифікація суб'єктивності (оскільки важливою рисою думок та їх тональності є суб'єктивність) і класифікація тексту як вираження позитивної, негативної чи нейтральної думки, відома як класифікація тональності. Крім того, часто розглядають емоційне забарвлення тексту. Такі інформаційні системи фокусуються на виявленні почуттів та емоцій (злість, щастя, сум тощо) чи виявляють наміри оповідача (наприклад: зацікавлений чи незацікавлений). Для врахування суб'єктивних факторів і нечіткої інформації, що притаманно вираженню людських емоцій, у процесі аналізу тональності тексту та визначення емоційного забарвлення текстової інформації доречно використовувати методи та підходи нечіткої логіки, яка формалізує здатність людини до наближених міркувань і допускає відмінні від бінарного значення істинності нечітких висловлювань [10–12].

Незважаючи на численні способи використання аналізу тональності, проектів, які б зосередили свою увагу на емоційному відклику для україномовних текстів, є надзвичайно мало. Тому мета роботи полягала у розробці інформаційної системи для визначення емоційного забарвлення текстів українською мовою.

Існує низка методів та алгоритмів для реалізації систем емоційного аналізу тексту, які виражають два основних підходи до автоматичної класифікації текстової інформації: методи, що покладаються на технології машинного навчання [13–15], і методи, які ґрунтуються на словниках і правилах [16–18]. Найбільш популярними методами є методи машинного навчання з учителем та без учителя, зокрема такі ймовірнісні алгоритми як наївний класифікатор Байєса та лінійна регресія, не ймовірнісні векторні машини підтримки та алгоритми глибокого навчання, які намагаються імітувати роботу людського мозку, застосовуючи для опрацювання даних штучні нейронні мережі. Всі ці методи відрізняються за точністю та швидкодією.

Згідно з іншим підходом, лінгвістичний аналіз здійснюється шляхом знаходження емоційно забарвленої лексики в тексті за наперед укладеними тональними словниками і правилами. За сукупністю знайденої емотивної лексики текст можна оцінити за шкалою, яка містить значення для негативних та позитивних слів. Зазначені методи використовують наступні основні процедури: спочатку кожному слову з тексту присвоюють значення тональності зі словника (за умови його присутності в словнику), а потім обчислюють загальну тональність всього тексту шляхом підсумовування значення тональностей за кожною з емоційних категорій [19]. Головна проблема методів, які засновані на словниках і правилах – це трудомісткість процесу створення словника. Для класифікації текстової інформації з високою точністю лексика словника повинна мати вагу, адекватну для предметної області документа. Перевага цього підходу полягає у можливості аналізувати емоційно забарвлену лексику на рівні окремих речень.

Досить часто тональні словники можуть містити декілька емоційних критеріїв (наприклад, радість, задоволення, сум, жах, тривога та ін.). У цих словниках слова віднесені до цього чи іншого критерію цілком (приналежність позначається нулем або одиницею) або вказаний ступінь приналежності дробовим числом. У випадку сентимент-аналізу україномовних текстів є обмежена кількість словників тональності, які охоплюють значно меншу кількість слів і виражають лише позитивну чи негативну оцінку. Зокрема, використані у роботі тональні словники української мови [20,21] містять 3442 і 6859 слів, відповідно. У першому випадку слова мають не нейтральну дискретну тональність і розподілені за чотирма категоріями: двома позитивними і двома

негативними (рис. 1*а*). Інша словникова база доповнена словами з нейтральною тональністю і містить п’ять категорій: дуже негативна (-2); негативна (-1); нейтральна (0); позитивна (1); дуже позитивна (2), як це показано на рис. 1*б*.

удача	1
удачливий	1
удостоювати	1
удостоюватися	1
уїдливий	-1
українофобство	-2
украсти	-2
улесливий	-1
ультрасучасний	1
улюблений	1
умертвити	-2
умисел	-1
умисний	-1
унікальність	1
упійманий	-2
упир	-2
упускати	-1
ура	1
уразити	-2
уразливий	-1
урегулювання	1
урожайний	1
урочистість	1
урочистий	2
усвідомлювати	1
ускладнення	-1
ускладнювати	-1
ускладнюватися	-1
уславлений	1
усміх	1
усміхатися	1
усміхнений	1
усмішка	1
успіх	2

а)

крихітний	0
крихкий	-1
кричущий	-1
кришка	0
кришталевий	1
кровожерливий	-2
кровопролитний	-1
кровопролиття	-2
кровотеча	-1
круїзний	0
крупнокаліберний	0
крутий	1
крутитися	0
круто	0
ксенофобія	-1
кубометр	0
кулемет	0
кулеметник	0
кульгати	-1
культ	0
культовий	0
культура	0
культурний	0
куля	-1
купка	-1
куратор	0
курганський	0
курдський	0
курирувати	0
курок	0
куртуазний	-1
курячий	0

б)

Рис. 1. Фрагменти тональних словників української мови [20] (*а*) і [21] (*б*).

Проведення більш якісного аналізу емоційного забарвлення україномовного тексту вимагає розширення словникової бази або застосування додаткових засобів та інструментів сентимент-аналізу. Зазначені особливості враховані у запропонованому алгоритмі визначення емоційного забарвлення текстової інформації українською мовою, який схематично зображений на рис. 2.

Для реалізації алгоритму використовувалась мова програмування Python 3.7, яка дає змогу легко і швидко опрацьовувати текстову інформацію. Створена програма забезпечує завантаження україномовного тексту у різних форматах. На наступному етапі до завантаженого тексту застосовано процедуру токенізації – розділення тексту на прості одиниці (tokens), тобто осмислені групи символів, які відповідають певним шаблонам (наприклад, слова) [17]. Зазвичай, токенізацію використовують як початковий і фундаментальний крок у багатьох задачах аналізу даних. У розробленій системі процедуру токенізації реалізовано з допомогою бібліотеки Python Tokenize UK. У результаті отримуємо масив слів, який після виключення розділових знаків та стоп-слів використовувався для морфологічного аналізу.

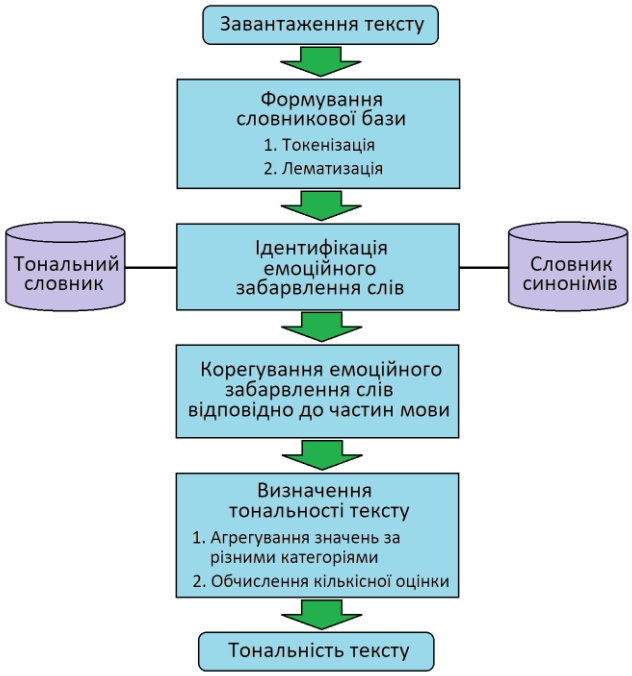

Рис. 2. Алгоритм визначення емоційного забарвлення текстової інформації.

Морфологічний аналіз слів, а саме процес лематизації здійснювався за допомогою бібліотеки pymorphy2. Під час лематизації утворені флективним способом слова приводять до базової граматичної форми (наприклад, для іменників і прикметників основною є форма називного відмінка однини, для дієслів – неозначена форма). Повернення базової, тобто словникової, форми слова сприяє формуванню ширших можливостей для емоційного аналізу текстової інформації.

Наступний етап алгоритму полягає у послідовному пошуку кожного слова з одержаного масиву у словнику тональності. У випадку виявлення, обраному слову присвоюють відповідне значення тональності. У протилежному випадку до пошуку залучають словник синонімів української мови, що дає змогу розширити словникову базу емоційно забарвлених слів. У разі, коли слово не виявлено ні у тональному словнику, ні у словнику синонімів, то зазначене слово не беруть до уваги в процесі сентимент-аналізу.

Крім того, для підвищення точності та обґрунтованості сентимент-аналізу текстової інформації у запропонованій системі автоматичного визначення тональності передбачена можливість врахування емоційного навантаження різних частин мови. З цією метою використано бібліотеку pymorphy2, яка дає змогу працювати з атрибутами тегів для визначення частин мови. Оскільки різні частини мови по-різному впливають на загальну оцінку тональності тексту, тому до отриманих значень тональності слів були застосовані корегувальні коефіцієнти. Зокрема, для прикметників і дієприкметників, які несуть головне емоційне навантаження, коефіцієнт дорівнював одиниці. Дієсловам та

іменникам присвоювались коефіцієнти у межах 0,6–0,8 та 0,4–0,6, відповідно. Застосовані коефіцієнти зменшують значення тональності слів, які несуть менше емоційне забарвлення і, відповідно, менше впливають на результати сентимент-аналізу.

Завершальний етап алгоритму охоплює процедури підсумовування значень тональності слів у кожній з категорій, їх нормалізацію та визначення не тільки якісної, але й кількісної оцінки емоційного забарвлення тексту. Результати тестування запропонованої системи сентимент-аналізу на україномовних текстах з різних джерел і різного емоційного забарвлення ілюструє рис. 3.

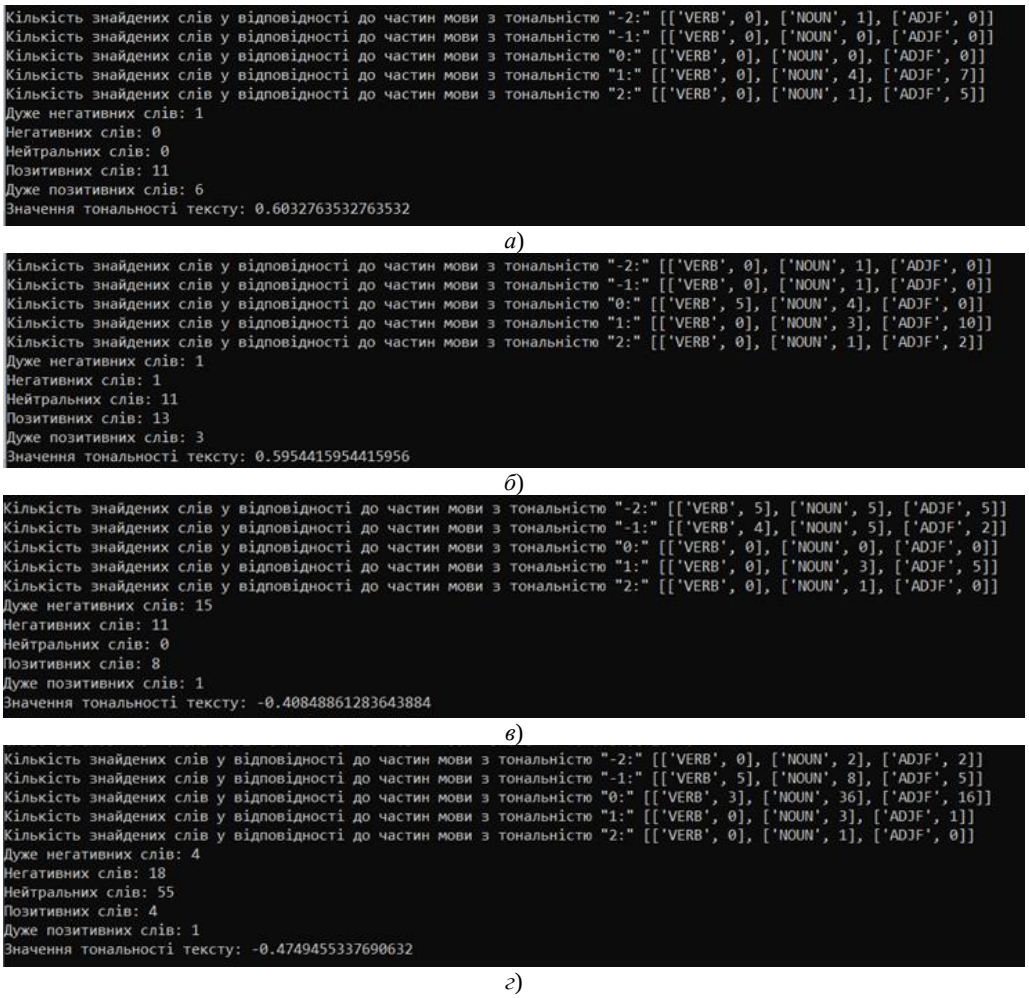

Рис. 3. Результати сентимент-аналізу текстів позитивної (*а*, *б*) та негативної тональностей (*в*, *г*) за допомогою тональних словників [20] (*а*, *в*) і [21] (*б*, *г*).

У найпростішому випадку якісного аналізу емоційну оцінку тексту визначають за тією категорією тональності, яка містить найбільшу кількість слів або характеризується найбільшим сумарним значенням тональності слів, віднесених до цієї категорії. Зокрема,

текст, результати аналізу якого зображені на рис. 3*а* і рис. 3*б*, можна віднести до позитивної тональності, а текст, якому відповідає рис. 3*в*, до дуже негативної. Цей ж текст у разі сентимент-аналізу за допомогою тонального словника [21], що містить нейтральні слова, можна віднести до нейтральної категорії (рис. 3*г*). Отже, у зазначеному випадку тональність текстової інформації залежить від використаного словника. Крім того, такий підхід не враховує внесок інших категорій тональності, який інколи може бути досить суттєвим, хоч і не визначальним.

Для забезпечення повнішого оцінювання емоційного забарвлення текстової інформації застосовано методи нечіткого моделювання, які не тільки вказують на приналежність елемента до тієї чи іншої множини за деякою характеристикою, але й зазначають ступінь приналежності у діапазоні значень від 0 до 1. Щоб застосувати засоби нечіткої логіки, сумарні значення тональності за кожною емоційною категорією були нормалізовані. Агрегування одержаних даних щодо різних категорій тональності та визначення кількісної оцінки тональності тексту здійснювалося за допомогою процедури дефазифікації методом центру ваги для одноелементних множин [22]:

$$T = \frac{\sum_{i=1}^{n} c_i t_i}{\sum_{i=1}^{n} t_i}, \tag{1}$$

де T – кількісна оцінка тональності тексту, c_i – категорія тональності, t_i – нормалізоване значення тональності слів i категорії. Такий підхід дає змогу враховувати слова з нейтральним емоційним забарвленням не змінюючи ранжування інших категорій тональності.

За допомогою розробленого програмного забезпечення проведено аналіз текстової інформації українською мовою за різними критеріями тональності. Крім того, здійснено порівняння результатів, отриманих без застосування додаткових засобів сентимент-аналізу і з врахуванням емоційного навантаження різних частин мови та використанням словника синонімів. Використання додаткових інструментів дає змогу проаналізувати більшу кількість слів у тексті та ступінь їх емоційного забарвлення і в результаті отримати коректніше значення тональності.

Отже, запропонована у роботі систем сентимент-аналізу дає змогу опрацьовувати україномовні тексти і забезпечує визначення тональності за п’ятьма емоційними категоріями: дуже позитивним, позитивним, нейтральним, негативним і дуже негативним. Внаслідок агрегування отриманих значень засобами нечіткого моделювання одержано кількісну оцінку тональності текстової інформації.

Список використаних джерел

[1] *Gallagher C.* The Application of Sentiment Analysis and Text Analytics to Customer Experience Reviews to Understand What Customers Are Really Saying / C. Gallagher, E. Furey, K. Curran // International Journal of Data Warehousing and Mining. – 2019. – Vol. 15(4). – P. 21–47.

[2] *Drus Z.* Sentiment Analysis in Social Media and Its Application: Systematic Literature Review / Z. Drus, H. Khalid // Procedia Computer Science. – 2019. – Vol. 161. – P. 707–714.

[3] *Poecze F.* Social Media Metrics and Sentiment Analysis to Evaluate the Effectiveness of Social Media Posts / F. Poecze, C. Ebster, C. Strauss // Procedia Computer Science. – 2018. – Vol. 130. – P. 660–666.

[4] *Gursoy U.T.* Social Media Mining and Sentiment Analysis for Brand Management / U.T. Gursoy, D. Bulut, C. Yigit // Global Journal of Emerging Trends in e-Business, Marketing and Consumer Psychology. – 2017. – Vol. 3. – P. 497–551.

[5] *Schumaker R.P.* Evaluating sentiment in financial news articles / R.P. Schumaker, Y. Zhang, C.N. Huang, H. Chen // Decision Support Systems. – 2012. – Vol. 53, No 3. – P. 458–464.

[6] *Hao J.* Social Media Content and Sentiment Analysis on Consumer Security Breaches / J. Hao, H. Dai // Journal of Financial Crime. – 2016. – Vol. 23, No 4. – P. 855–869.

[7] *Mansour S.* Social Media Analysis of User's Responses to terrorism using sentiment analysis and text mining / S. Mansour // Procedia Computer Science. – 2018. – Vol. 140. – P. 95–103.

[8] *Feldman R.* Techniques and Applications for Sentiment Analysis / Ronen Feldman // Communications of the ACM. – 2013. – Vol. 56, No 4. – P. 82–89.

[9] *D'Andrea A.* Approaches, Tools and Applications for Sentiment Analysis Implementation / A. D'Andrea, F. Ferri, P. Grifoni, T. Guzzo // International Journal of Computer Applications. – 2015. – Vol. 125, No 3. – P. 26–33.

[10] *Howellsa K.* Applying fuzzy logic for sentiment analysis of social media network data in marketing / K. Howellsa, A. Ertugan // Proc. Comp. Sci. – 2017. – Vol. 120. – P. 664–670.

[11] *Jain P.* Aspect Based Sentiment Analysis by Fuzzy Logic / P. Jain, A. Srivastava, V. Singh, B. Hazela // International Journal of Current Engineering and Technology. – 2019. – Vol. 9, No 2. – P. 243–248.

[12] *Liu H.* Fuzzy Rule Based Systems for Interpretable Sentiment Analysis / H. Liu, M. Cocea // International Conference on Advanced Computational Intelligence. – 2017. – P. 129–136.

[13] *Pang B.* Opinion Mining and Sentiment Analysis / B. Pang, L. Lee // Foundations and Trends in Information Retrieval. – 2008. – Vol. 2. – P. 1–135.

[14] *Khurshid A.* Affective Computing and Sentiment Analysis: Metaphor, Ontology, Affect and Terminology / A. Khurshid – Berlin: Springer Science & Business Media, 2011 – 164 p.

[15] *Jain U.* A Review on the Emotion Detection from Text using Machine Learning Techniques / U. Jain, A. Sandhu // International Journal of Current Engineering and Technology. – 2015. – Vol.5, No.4. – P. 2645–2650.

[16] *Chopra F.K.* Sentiment Analyzing by Dictionary based Approach / F.K. Chopra, R. Bhatia // International Journal of Computer Applications. – 2016. – Vol. 152, No.5. – P. 32–34.

[17] *Thelwall M.* Sentiment strength detection in short informal text / M. Thelwall, K. Buckley, G. Paltoglou, A. Kappas, D. Cai // Journal of the American Society for Information Science and Technology. – 2010. – No. 61. – P. 2544–2558.

[18] *Denecke K.* Using SentiWordNet for Multilingual Sentiment Analysis / K. Denecke // International Conference on Data Engineering Workshops. – 2008. – P. 507–512.

[19] *Dang Y.* A lexicon enhanced method for sentiment classification: An experiment on online product reviews / Y. Dang, Y. Zhang, H. Chen // IEEE Intelligent Systems. – 2010. – Vol. 25, No.4. – P. 46–53.

[20] Український тональний словник [Електронний ресурс]. – Режим доступу: https://github.com/lang-uk/tone-dict-uk/blob/master/tone-dict-uk.tsv

[21] Український тональний словник [Електронний ресурс]. – Режим доступу: https://github.com/lang-uk/tone-dict-uk/blob/master/tone-dict-uk-manual.tsv

[22] *Takagi T.* Fuzzy Identification of Systems and Its Applications to Modeling and Control / T. Takagi, M. Sugeno // IEEE Transactions on Systems, Man and Cybernetics. – 1985. – Vol. 15. – P. 116–132.

SYSTEM OF AUTOMATIC DETERMINATION OF TEXT TONE

I. Olenych, M. Prytula, O. Sinkevych, O. Khamar

Ivan Franko National University of Lviv,
50 Drahomanov St., UA–79005 Lviv, Ukraine
iolenych@gmail.com

In the work, the methods of sentiment analysis of the text are considered. The system of emotion detection of Ukrainian-language texts based on dictionaries and rules is proposed. The developed software downloads text information in various formats and carries out tokenization and lemmatization procedures using the Python Tokenize UK and pymorphy2 libraries. As a result, an array of words in the basic grammatical form using for determining the tone of the text is formed. The obtained word base was analyzed using a tonal dictionary of the Ukrainian language. A dictionary of synonyms was used to expand the vocabulary. If there is no word in the tonal dictionary, the tonality value of its nearest synonym is used for further calculations.

Computer analysis of textual information was performed in five emotional categories namely: very positive, positive, neutral, negative, and very negative tone of words. To increase the accuracy and validity of sentiment analysis, coefficients were used that take into account the various emotional load of words of different speech parts and their dissimilar impact on the overall assessment of the text tone. The proposed system of sentiment analysis assumes a greater emotional influence of adjectives compared to verbs and nouns. Since the tone of textual information and the expression of human emotions are subjective factors, the means of fuzzy modeling are used for sentiment analysis of texts. This approach makes it possible to take into account the contribution of all emotional categories in the final evaluation of the text. As a result of an aggregation of normalized data on different emotion categories and defuzzification by the method of the center of gravity for one-element sets, a quantitative estimate of the emotional tone of texts was obtained.

The developed system of sentiment analysis was tested on Ukrainian-language texts from different sources and different emotional tones. The use of additional tools provides the analysis of more words in the text and the degree of their emotional tone, which leads to more correct detection of the tone of the whole text.

Key words: computer text analysis, sentiment analysis, tokenization, tonality dictionary, fuzzy modeling.

Стаття: надійшла до редакції 21.05.2021,
доопрацьована 24.06.2021,
прийнята до друку 25.06.2021