Кібернетика та комп'ютерні технології

УДК 004[023+62] + 519.7 DOI:10.34229/2707-451X.23.2.6

М.О. ЩЬОГОЛЄВ, О.В. АНДРІЙЧУК

МАТЕМАТИЧНІ МЕТОДИ ОБРОБКИ ПРИРОДНОЇ МОВИ У СИСТЕМІ ОПЕРАТИВНОГО ВИЗНАЧЕННЯ РІВНЯ НАПРУЖЕНОСТІ В СУСПІЛЬСТВІ

Розроблено архітектуру та математичне забезпечення системи оперативного визначення рівня напруженості в суспільстві на основі даних з соціальних мереж. У статті подано опис математичних методів обробки природної мови, що використовувались у цій системі, продемонстровано приклад її застосування.

Ключові слова: *рівень напруженості в суспільстві, соціальні мережі, сентимент-аналіз, TF-IDF, Word2vec, нейронні мережі.*

Вступ. Найголовнішими чинниками, що спричиняють підвищення рівня напруженості в суспільстві, є кризові явища та інформаційні операції. До кризових явищ можна віднести економічні, політичні, соціальні кризи, а також епідемії та війни.

Протягом кількох останніх років українське суспільство потерпає від кризових явищ у різних сферах та на різних стадіях, у тому числі, на стадії найбільшого загострення. Зростання рівня напруженості, спричинене цими явищами, посилює внутрішні конфлікти, знижує можливості державних органів щодо ефективного реагування на зовнішні загрози. Інформаційні атаки в основному працюють у полі існуючих конфліктів, поглиблюючи їх.

Сьогодні основним способом визначення рівня напруженості, що виникає у деякої групи людей у зв'язку з певною подією, вважаються соціологічні опитування. Одним з недоліків такого підходу з погляду державних посадових осіб, що мають приймати оперативні рішення у кризових ситуаціях, є його ретроспективний характер: на момент отримання результатів ситуація може частково або повністю змінитись. Витрата часу на пошук респондентів, саме опитування та обробку анкет призводить до суттєвого звуження можливостей щодо оперативного реагування на виклики. До того ж, інформація про динаміку зміни напруженості, визначеної за допомогою соціологічних опитувань, дуже обмежена, оскільки опитування на ту ж саму тему в основному проводяться не частіше, ніж раз на кілька місяців.

Альтернативним джерелом даних є соціальні мережі. У порівнянні з результатами опитувань респондентів ці дані мають низку переваг. По-перше, вони завжди відображають актуальну ситуацію щодо рівня напруженості в суспільстві, оскільки люди починають активно коментувати новинні події відразу ж після появи перших публікацій, що ці події висвітлюють. Як наслідок, дані з соціальних мереж дозволяють дослідити рівень напруженості у динаміці, зокрема, дослідити динаміку реакції суспільства на інформаційні

атаки. По-друге, процес збору і обробки даних з соціальних мереж легше автоматизувати. По-третє, використання таких даних не потребує пошуку каналу комунікації між соціологом та респондентом, більше того, не потребує участі респондента в опитуванні як такому. Суть дослідження полягає у зборі та обробці даних, які з'являються у вільному доступі невимушено – з власної волі їхніх авторів, керованих бажанням поділитись власними думками та емоціями з іншими читачами та дописувачами соціальних мереж. Однак такий підхід має свої недоліки. Серед найбільших – необхідність визначення та відсіювання повідомлень, згенерованих ботами, необхідність класифікувати публікації за темами, складність визначення напруженості в окремих соціальних групах.

В рамках цього дослідження соціальною напруженістю (СН) будемо називати емоційно-психологічний стан людей, викликаний впливом зовнішніх подразнюючих факторів, до яких входять як реальні кризові явища (війна, пандемія, безробіття, підвищення цін, низька соціальна захищеність тощо), так і інформаційні впливи (інформаційні атаки, політична агітація, новини, що генерують почуття тривожності). Припускається, що напруженість може бути визначена на основі аналізу емоційності коментарів користувачів соціальних мереж на певні новинні публікації; напруженість, спровокована певною новиною, може бути визначена як узагальнена емоційність коментарів до новинних публікацій, що висвітлюють цю новину.

З 2017 по 2021 рік в рамках проєкту HATO CyRADARS на базі кафедри прикладної математики НТУУ «КПІ ім. І. Сікорського» та за участі інших українських та закордонних наукових інституцій розроблялась комплексна система визначення інформаційних загроз в Україні для Міністерства культури та інформаційної політики України й Міністерства оборони України. Вона, зокрема, включала у себе систему оперативного визначення рівня напруженості в суспільстві за даними із соціальних мереж. Оперативність визначення напруженості надавала перевагу у швидкості прийняття рішень працівникам відповідних установ. Опис прототипу системи подано у статті [1] та тезах [2].

Для побудови системи оперативного визначення рівня напруженості в суспільстві розроблено унікальну архітектуру, що включає у себе методи сентимент-аналізу, нейронну мережу, базу знань системи підтримки прийняття рішень, базу даних, модуль автоматичного збору даних та інші структурні елементи. У даній статті подано математичний опис методів обробки природної мови, що застосовуються у цій системі.

Запропонована система оцінює рівень напруженості за текстовими даними з соціальної мережі Facebook – за коментарями до новинних публікацій. З практичного погляду під оцінкою рівня напруженості в рамках цього дослідження розуміються усереднена оцінка емоційності за певний проміжок часу або щодо певної новинної події у рамках одного новинного ресурсу та оцінка впливу новинних подій на підвищення загального рівня напруженості. У рамках цієї статті розглядається спосіб визначення першої з цих двох оцінок. Система оновлюватиме оцінку рівня напруженості щогодини, що дозволить досліджувати динаміку її зміни у розрізі кожної новинної події.

Аналіз літератури та постановка проблеми. Існують різні методологічні підходи до визначення рівня напруженості (в деяких джерелах – рівня депривації, стресу) в суспільстві. Соціологи поділяють всі методи на дві умовні групи: анкетні опитування та аналіз статистичного матеріалу [3]. Анкетні опитування надають багато можливостей для різних видів аналізу, однак дослідження будуть обмежуватись лише певною соціальною групою. Аналіз статистичного матеріалу надає можливість досліджувати динаміку зміни напруженості, а також дозволяє визначати напруженість у межах великих регіонів та на загальнонаціональному рівні, однак такі дослідження носитимуть ретроспективний характер.

Метод анкетних опитувань був використаний у статтях [4–6], метод аналізу статистичного матеріалу в [7–9].

Існує багато досліджень, у яких автори використовують методи машинного навчання та сентимент-аналізу текстових даних, однак ці дані не пов'язані з СН. Таким дослідженням присвячені

статті [10–13]. У них досліджується якість продуктів інтернет-магазинів на основі емоційності коментарів споживачів; ставлення студентів до коледжів, у яких вони навчаються, на основі їхніх відгуків; рейтинги фільмів на основі рецензій на них. У всіх дослідженнях використовуються згорткові або рекурентні нейронні мережі.

Опис систем, у яких використовуються методи машинного навчання та сентимент-аналізу і які призначені для аналізу даних, пов'язаних з емоційністю та СН, подано у [14–23]. У більшості з них коротко або більш розлого подається опис попередньої обробки текстових даних, що включає у себе очистку тексту (видалення стоп-слів, знаків пунктуації, спеціальних символів, емодзі, гештеґів тощо), токенізацію тексту, побудову вкладань слів та інші методики. Для класифікації даних використовуються штучні нейронні мережі з різними архітектурами, метод опорних векторів, логістична регресія, метод *k*-найближчих сусідів, наївний баєсівський класифікатор та інші.

Жодна з перелічених робіт не досліджує напруженість у розрізі кожної новинної події, не визначає її вплив на загальну напруженість і не ставить на меті визначення динаміки її зміни. Тож проблема дослідження СН під таким кутом залишається актуальною.

Мета та задачі дослідження – підвищення ситуаційної обізнаності представників державних установ щодо поточного рівня СН, спровокованої кризовими явищами, новинними подіями чи інформаційними операціями. Отримана інформація допомагатиме державним посадовим особам приймати швидкі рішення для подолання цих кризових явищ та протидії дезінформації.

Основна задача дослідження – розробка архітектури та математичного забезпечення системи оперативного визначення рівня напруженості в суспільстві за даними з соціальних мереж.

Загальний опис архітектури системи оперативного визначення рівня напруженості в суспільстві. Для зображення архітектури системи та способів переходу даних з одного блоку в інший створено діаграму потоків даних, яка показана на рис. 1.

Робота системи починається зі збору даних. Основне джерело даних – це соціальна мережа Facebook. У рамках цього дослідження збираються дані з офіційної сторінки ТСН у цій соціальній мережі. Дані включають у себе текст публікацій, час публікацій, посилання на повні статті на сайті tsn.ua та теґи статей (теґи збираються безпосередньо з сайту tsn.ua). Окрім того, збираються усі коментарі до кожної публікації, час публікації коментарів та інформація про рівень їхньої вкладеності. Усі дані зберігаються в базі даних у вигляді двох таблиць, пов'язаних за ключем, у ролі якого виступає номер публікації.

Перед класифікацією коментарі проходять попередню обробку. З них видаляються стоп-слова, знаки пунктуації, спеціальні символи, емодзі, гіперпосилання тощо. Окрім того, з них видаляються посилання на інших дописувачів, що дозволяє повністю анонімізувати оброблювані дані. Далі з коментарів виділяються ключові слова, що відображають їхню основну суть. Кожному ключовому слову ставиться у відповідність його контекстний вектор. Набір ключових слів разом з контекстними векторами, що їм відповідають, зберігається у словнику, що містить вичерпний перелік таких пар.

Для визначення емоційності коментарів використовується штучна нейронна мережа, на вхід якої подається набір ключових слів для одного коментаря, а на виході отримується його оцінка емоційності. Усі оцінки емоційності зберігаються в базі даних. На основі цих оцінок підраховується статистичний розподіл емоційності коментарів для кожної публікації. Набори цих розподілів для кожної ітерації роботи системи зберігаються в окремі файли.

Дані статистики використовуються для визначення ступеня впливу новин на підвищення загального рівня напруженості. Для вирішення цієї задачі використано ієрархічну базу знань системи підтримки прийняття рішень. За результатами її роботи новинні події та новинні публікації ранжуються за рівнем напруженості, яку вони спричиняють у дописувачів досліджуваного новинного ресурсу. Опис принципу її роботи подано у [1].

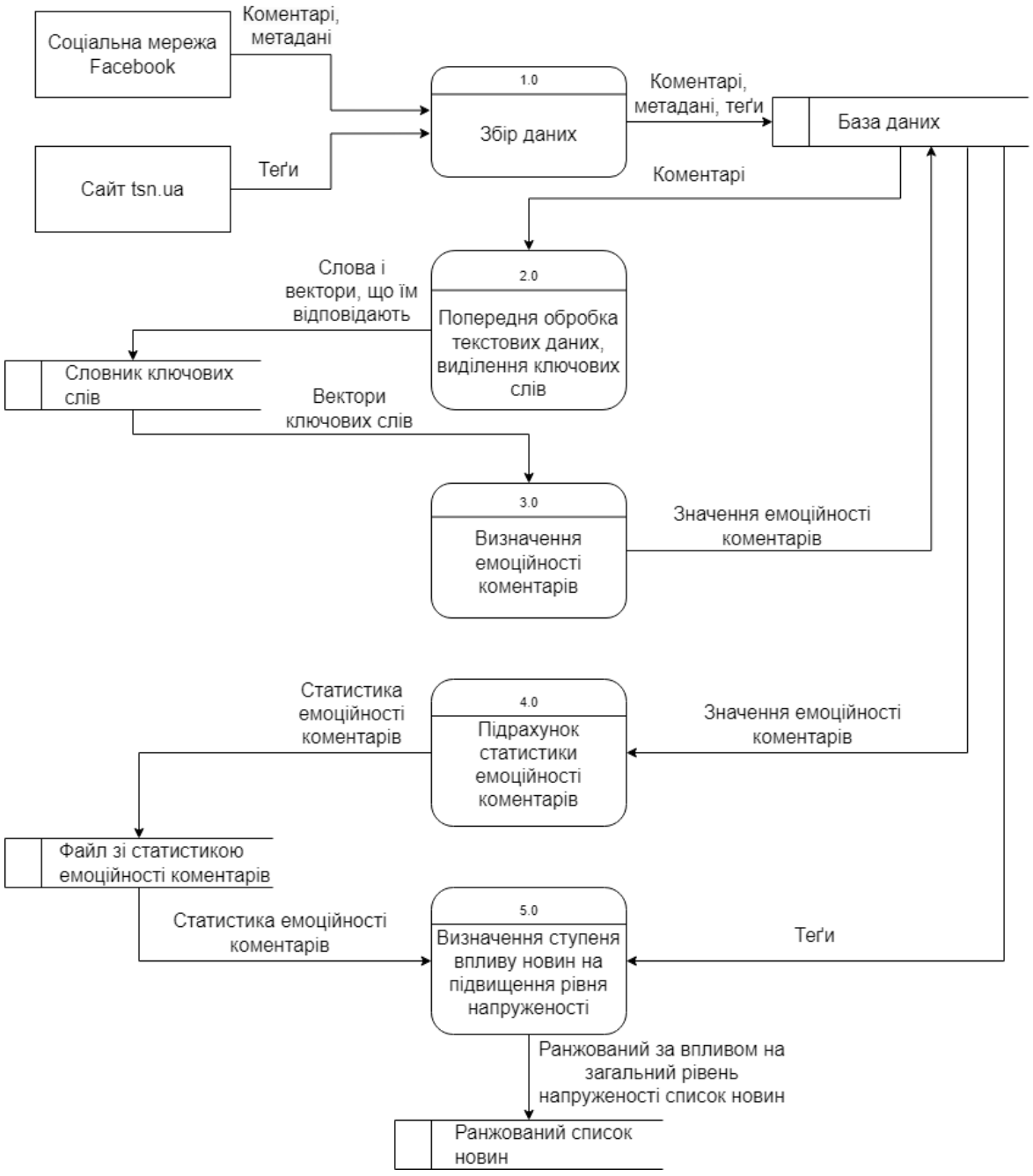

РИС. 1. Діаграма потоків даних системи оперативного визначення рівня напруженості в суспільстві

Попередня обробка текстових даних. Перед класифікацією коментарі до новинних публікацій мають пройти етап попередньої обробки. Далі подається опис математичних методів, які можуть для цього використовуватись.

Після очистки тексту для оцінки важливості слів у контексті кожного коментаря пропонується застосувати метод TF-IDF або його модифікації. Тут буде подано математичний опис класичного методу TF-IDF [24].

Показник TF-IDF складається з двох величин: TF (term frequency – частота терміна у документі) та IDF (inverse document frequency – обернена частота документа). Міра TF підраховується за формулою

$$tf_{t,d} = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}, \quad (1)$$

де $f_{t,d}$ – число входжень терміна t у документ d, у знаменнику – загальна кількість термінів у документі. Міра IDF підраховується за формулою

$$idf_t = \log \frac{N}{df_t}, \quad (2)$$

де N – загальна кількість документів у колекції, df – кількість документів, у яких зустрічається слово t (якщо $f_{t,d} \neq 0$). Загалом показник TF-IDF підраховується так:

$$f - idf_{t,d} = tf_{t,d} \times idf_t . \quad (3)$$

З кожного коментаря виділятиметься певна фіксована кількість слів (3–7), що матимуть найбільші показники TF-IDF. Кожному з цих слів буде ставитись у відповідність його контекстний вектор. Контекстний вектор визначатиметься за допомогою різновиду методу Word2vec – методу CBOW. Загальний опис методу Word2vec подано у [25, 26], детальний математичний опис наведено у [27].

Метод CBOW використовує одношарову нейронну мережу, приблизну архітектуру якої показано на рис. 2. Слід зауважити, що ключове слово подається на вхід нейронної мережі разом з кількома словами до та після нього, що утворюють його контекст.

Для визначення результуючого вектора прихованого шару нейронної мережі модель CBOW підраховує середнє значення векторів вхідних контекстних слів та повертає добуток матриці ваг із вхідного шару на прихований і підрахованого вектора середніх значень:

$$h = \frac{1}{C} W^T (x_1 + x_2 + \ldots + x_C) = \frac{1}{C} (v_{w1} + v_{w2} + \ldots + v_{wC})^T, \quad (4)$$

де C – кількість слів у контексті, $x_1, \ldots, x_C$ – вектори контекстних слів на вхідному шарі, $w_1, \ldots, w_C$ – контекстні слова, v_w – вхідний вектор слова w на прихованому шарі.

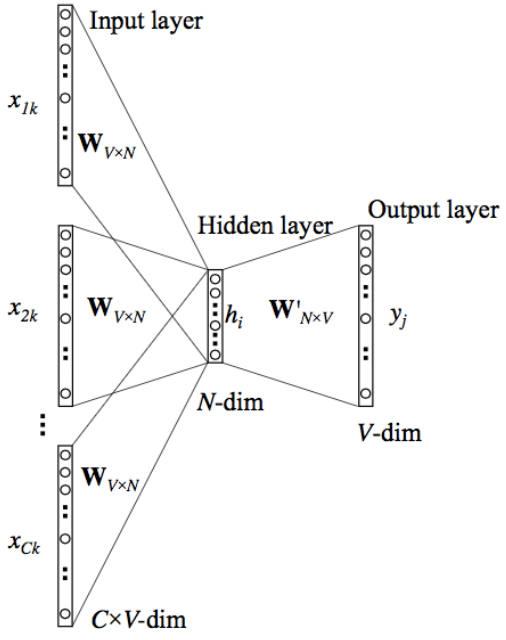

РИС. 2. Архітектура нейронної мережі методу CBOW [27]

Функція втрат має вигляд:

$$E = -\log p\left(w_O \middle| w_{I,1}, \dots, w_{I,C}\right) = -u_{j^*} + \log \sum_{j'=1}^{V} \exp\left(u_{j'}\right) = -{v'_{w_O}}^T \cdot h + \log \sum_{j'=1}^{V} \exp({v'_{w_j}}^T \cdot h) \quad (5)$$

з урахуванням того, що в ролі функції активації нейронів використовується функція Softmax, тобто:

$$p\left(w_O \middle| w_{I,1}, \dots, w_{I,C}\right) = \frac{\exp(u_{j^*})}{\sum_{j'=1}^{V} \exp(u_{j'})}, \quad (6)$$

а також того, що

$$u_j = {v'_{w_j}}^T \cdot h, \quad (7)$$

$$u_{j^*} = {v'_{w_O}}^T \cdot h, \quad (8)$$

де w_O – результуюче слово, $w_{I,1}, \dots, w_{I,C}$ – вхідні контекстні слова.

Рівняння для ваг з прихованого шару на результуючий має вигляд:

$${v'_{w_j}}^{(new)} = {v'_{w_j}}^{(old)} - \eta \cdot e_j \cdot h, \quad (9)$$

де $j = 1, 2, \dots, V$. Це рівняння застосовується для кожного елемента матриці ваг з прихованого шару на результуючий для кожного тренувального екземпляра.

Рівняння ваг із вхідного шару на прихований має вигляд:

$${v_{w_{I,c}}}^{(new)} = {v_{w_{I,c}}}^{(old)} - \frac{1}{C} \cdot \eta \cdot EH^T, \quad (10)$$

де $c = 1, 2, \dots, C$, $v_{w_{I,c}}$ – вхідний вектор c-того слова у вхідному контексті, $\eta > 0$ – частота навчання, $EH = \frac{\partial E}{\partial h_i}$, h_i – результуюче значення i-того нейрона прихованого шару.

Отримані на виході методу CBOW контекстні вектори слів потрапляють до словника ключових слів, а потім – на вхід нейронної мережі.

Словник можна представити як послідовність впорядкованих пар $(w_s, l_s) \in (\mathbb{N}^n \times \mathbb{R}^m)$; $s = \overline{1, \dots, q}$; q – поточна кількість слів у словнику, $w_s \in \mathbb{N}^n$ – слова у форматі тексту, $l_s \in \mathbb{R}^m$ – слова у векторному представленні.

Нейронна мережа. Загалом нейронну мережу можна представити у вигляді функції $F: \mathbb{R}^{n \times m} \to \mathbb{R}$ такої, що

$$F(w_1, \dots, w_n) = r_k, \quad (11)$$

де $w_1, \dots, w_n$ – ключові слова у векторній формі, m – довжина вектора ключового слова, $r_k \in \mathbb{R}$ – оцінка емоційності коментаря k, $r_k \in [0, 2]$, де коментарі з позитивною емоційністю набувають значення 0, коментарі з негативною емоційністю – значення 2.

Для визначення емоційності коментарів за ключовими словами пропонується використовувати нейронну мережу зворотного поширення помилки. У цій роботі буде подано математичний опис класичної нейронної мережі з алгоритмом зворотного поширення помилки, описаної у [28]. Така нейронна мережа складається із вхідного шару, певної кількості прихованих шарів та результуючого шару. Всі зв'язки йдуть лише від вхідного шару у напрямку результуючого. Зв'язків у зворотному напрямку та зв'язків у межах прихованого шару немає, однак деякі зв'язки можуть проходити повз приховані шари.

Вхідне значення x_j довільного j-го нейрона на прихованому шарі – це лінійна комбінація результуючих значень на попередньому шарі:

$$x_j = \sum_i y_i w_{ji}, \quad (12)$$

де y_j – результуючі значення нейронів на попередньому шарі, що мають зв'язок з j-тим нейроном, w_{ji} – ваги синоптичних зв'язків.

Результуючі значення y_j можуть підраховуватись за різними формулами, зокрема, може бути використана нелінійна сигмоїдна функція:

$$y_j = \frac{1}{1+e^{-x_j}}. \quad (13)$$

Мета тренування нейронної мережі зворотного поширення помилки це пошук такого набору ваг синоптичних зв'язків, щоб для кожного вхідного набору значень результуюче значення дорівнювало або було близьким до очікуваного. Порівнюючи дійсні та бажані результати роботи мережі, можна підрахувати помилку. Загальна помилка може бути визначена за формулою:

$$E = \frac{1}{2}\sum_c \sum_j (y_{j,c} - d_{j,c})^2, \quad (14)$$

де c – індекс пар вхідних та вихідних значень, j – індекс нейронів результуючого шару, y – дійсне результуюче значення, d – бажане результуюче значення.

Для мінімізації помилки E методом градієнтного спуску потрібно підрахувати часткові похідні цього виразу по кожній з ваг мережі, причому ці часткові похідні мають підраховуватись у два проходи: прямому та зворотному. Прямий прохід було описано за допомогою формул (12) та (13). Розглянемо зворотний прохід.

Зворотний прохід починається з підрахунку часткової похідної $\partial E/\partial y$ для кожного нейрона результуючого шару. Диференціюючи рівняння (14), отримуємо:

$$\frac{\partial E}{\partial y_j} = y_j - d_j. \quad (15)$$

Застосовуючи правило диференціювання складної функції, можна підрахувати $\partial E/\partial x_j$:

$$\frac{\partial E}{\partial x_j} = \frac{\partial E}{\partial y_j} \cdot \frac{\partial y_j}{\partial x_j}. \quad (16)$$

Диференціюючи рівняння (13) та підставляючи отримане значення у рівняння (16), отримаємо:

$$\frac{\partial E}{\partial x_j} = \frac{\partial E}{\partial y_j} \cdot y_j(1 - y_j). \quad (17)$$

Підрахуємо часткову похідну для ваги w_{ji}:

$$\frac{\partial E}{\partial w_{ji}} = \frac{\partial E}{\partial x_j} \cdot \frac{\partial x_j}{\partial w_{ji}} = \frac{\partial E}{\partial x_j} \cdot y_i. \quad (18)$$

Застосовуючи правило диференціювання складної функції до $\partial E/\partial y_i$ для виходу i-того нейрона, отримуємо:

$$\frac{\partial E}{\partial x_j} \cdot \frac{\partial x_j}{\partial y_i} = \frac{\partial E}{\partial x_j} \cdot w_{ji}. \quad (19)$$

Беручи до уваги усі зв'язки, що виходять з i-того нейрона, отримуємо:

$$\frac{\partial E}{\partial y_i} = \sum_j \frac{\partial E}{\partial x_j} \cdot w_{ji}. \quad (20)$$

За формулою (20) можна підраховувати значення $\partial E/\partial y$ для кожного нейрона передостаннього шару на основі значень $\partial E/\partial y$ нейронів останнього шару. Ця процедура має повторюватись для всіх попередніх шарів. Водночас підраховуються значення $\partial E/\partial w$.

Корегувати ваги можна двома способами: або за допомогою простішої формули

$$\Delta w = -\varepsilon \frac{\partial E}{\partial w}, \quad (21)$$

де кожна вага змінюється на величину, пропорційну до акумульованого $\partial E/\partial w$, або за допомогою більш вдосконаленої формули

$$\Delta w(t) = -\varepsilon \frac{\partial E}{\partial w(t)} + \alpha \Delta w(t-1), \quad (22)$$

де t збільшується на 1 після кожного проходу за всіма екземплярами, α – експоненційний фактор розпаду на інтервалі від 0 до 1, що визначає відносний внесок поточного та попередніх градієнтів у зміну ваги.

Підрахунок середніх значень напруженості. Отримані оцінки емоційності коментарів використовуються для побудови графіку розподілу емоційності для кожної публікації. Схематично цей графік можна подати у вигляді однієї з восьми гістограм (рис. 3).

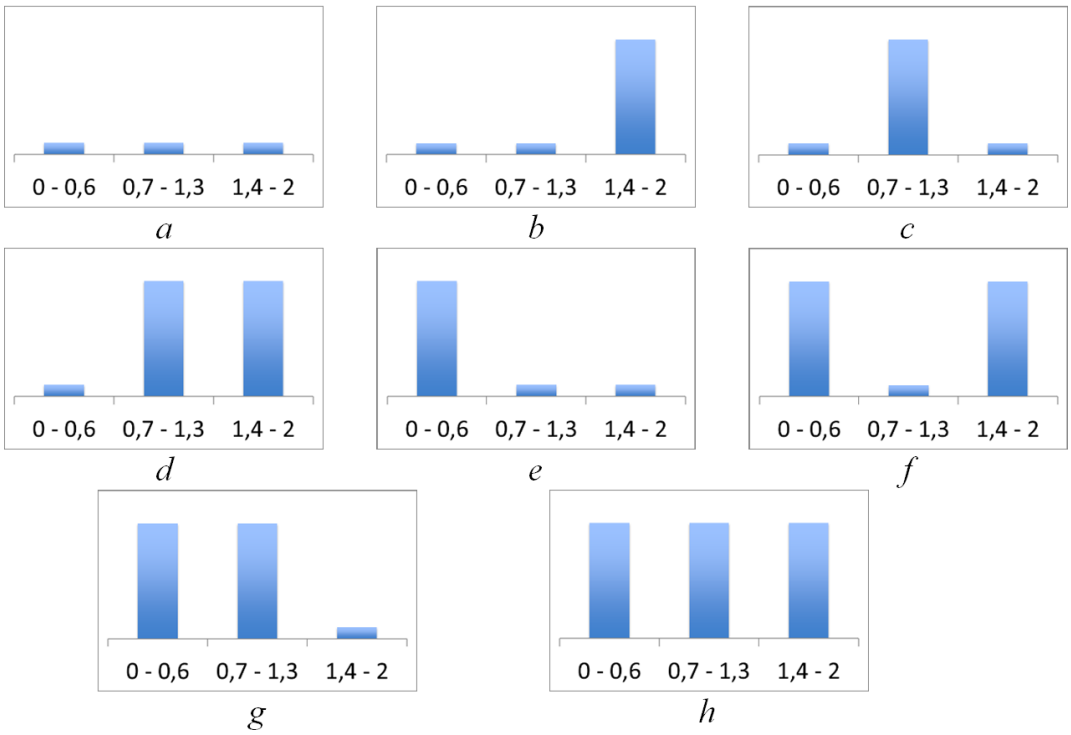

РИС. 3. Схематичні графіки розподілу емоційності для кожної публікації

Якщо кількість коментарів з оцінками емоційності 1.4–2 досить велика, це означає, що досліджувана публікація провокує підвищення рівня напруженості. Тому надалі увага зосереджуватиметься насамперед на тих публікаціях, емоційний рисунок яких має вигляд *b*, *d*, *f* або *h*.

Середня кількість коментарів з оцінками 1.4–2 підраховується як зважене середнє арифметичне:

$$x_1 = \frac{\sum_{i=1}^{n_1} N_i w_i}{n_1}, \tag{23}$$

де $N_i \in \mathbb{N}$ – кількість коментарів з певною оцінкою емоційності, $n_1 = 7$ – кількість можливих оцінок емоційності на інтервалі 1.4-2, $w_i = i/10 + 0{,}6$ – вагові коефіцієнти, $i \in [1, n_1]$.

Загальна середня кількість коментарів підраховується як звичайне середнє арифметичне:

$$x_2 = \frac{\sum_{i=1}^{n_2} N_i}{n_2}, \tag{24}$$

де $N_i \in \mathbb{N}$ – кількість коментарів з певною оцінкою емоційності, $n_2 = 21$ – загальна кількість можливих оцінок емоційності, $i \in [1, n_2]$.

Щоб підрахувати відношення між загальною середньою кількістю коментарів та середньою кількістю коментарів з негативним емоційним забарвленням, вводиться коефіцієнт k:

$$k = \frac{x_2}{x_1}. \tag{25}$$

Основну увагу буде зосереджено на тих публікаціях, для яких $k \leq 1$, тобто де середня кількість коментарів з високою негативною емоційністю перевищує загальну середню кількість коментарів. При цьому підраховуються значення n_{neg}, що показує кількість досліджуваних публікацій, для яких $k \leq 1$, та n_{pos}, що показує кількість досліджуваних публікацій, для яких $k > 1$.

Щоб врахувати популярність новинних ресурсів, підраховується значення k^* для кожного k:

$$k^* = 1 - a(1 - k)\,, \tag{26}$$

де a – кількість підписників, нормована за всіма новинними ресурсами, $a \in (0; 1]$. Графік значень k^* для різних a та k на інтервалі $[0; 1]$ показано на рис. 4.

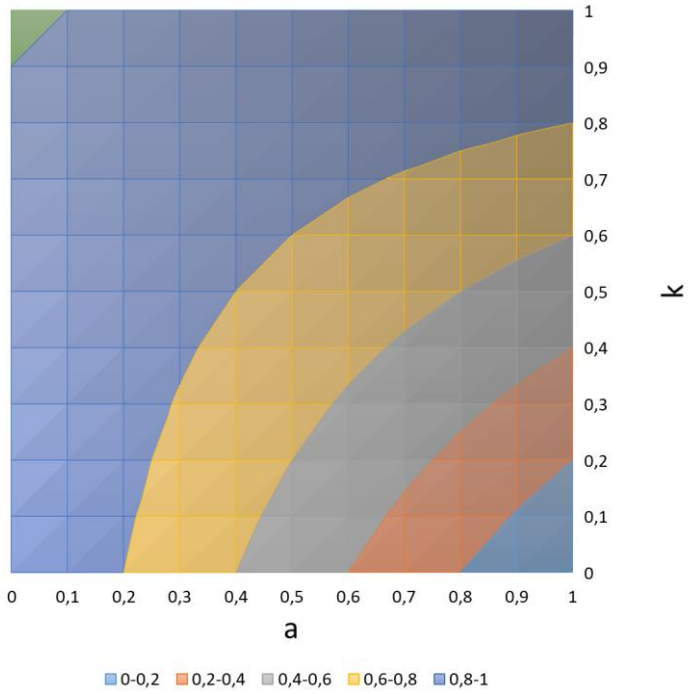

РИС. 4. Графік значень k^* для різних a та k на інтервалі $[0; 1]$

Для підрахунку середнього рівня напруженості застосовується формула середнього геометричного:

$$\bar{X} = \sqrt[n]{\prod_{j=1}^{n} {k_j}^*}\,, \tag{27}$$

$\bar{X} \in [0, 1]$. В залежності від поставленої задачі за цією формулою може бути підрахована середня негативна напруженість або для певного новинного ресурсу, або для певної новини.

Приклад підрахунку коефіцієнта k . Розрахуємо значення рівня напруженості, спровокованої новинною публікацією, яка на сайті ТСН вийшла під назвою «Путін наляканий і у відчаї покладає останню надію на зиму в Європі — CNN» 26 жовтня 2022 року о 12:06. На офіційній сторінці Facebook посилання на цю статтю було додане о 12:08. Станом на 14:27 того ж дня до публікації з цим посиланням було додано 39 коментарів, з них 1 коментар містив лише зображення, тож його зміст до уваги не брався. Перші 15 коментарів з рівнями вкладеності та значеннями емоційності, визначеними експертами, подано у таблиці.

ТАБЛИЦЯ. Коментарі з офіційної сторінки ТСН у соціальній мережі Facebook на новинну публікацію «Путін наляканий і у відчаї покладає останню надію на зиму в Європі — CNN»

№	Значення емоційності	Рівень вкладеності	Коментар
1	1,4	0	Європа ВЖЕ знайшла заміну 2/3 поставок газу з рассеі, а ми і на дровах проживе аби воно здохло, тому вибачте пуйло, але надіятись нема на шо
2	0,5	0	Не дочекається українці силний народ незамерзли у сибірах і тепер з Божою допомогою перезимуєм
3	0,3	0	Бог дасть нам теплу зиму і перезимуємо.....українців не зламати....
4	0,8	0	Відвернути увагу всього світу від військових злочинів, скоєних путіним в Україні, вже неможливо і саме це є запорукою того, що путіну настане кінець, як би він не старався задурити голови росіянам
5	0,9	0	А ми сподіваємося на скорий суд Божий!
6	1,3	1	Уже кінець світу був у 2012 році, стадо хаває казки біблейські
7	1,1	2	Я про кінець світу не писав нічого. Про 2012 рік Біблія не говорить нічого. Я писав про інше
8	1,6	2	Ну правильно, в загальному ви не берете події, а лиш те що написано, що скоро, скоро буде коли віруючі повбивають себе , бо в очікування суду, але суд щось чекають в Гаазі. Чи може хтось згвалтував дитину, і чекає гвалтівник суд божий? Не тупіть, і не пишіть що попало, але такі як ви, це хаває, і моляться в коментарях, що ж ви робили,, коли не було соц мереж, стояли біля ікон в будинках, і молилися в кутках
9	1,8	0	Та грохніть вже ту наволоч, як зігріється всвоєму газі і нафті
10	1,3	0	Цією зимою цього бункерного довбня переламають
11	0,7	0	В Європі опалення включили ще в кінці вересня,так що дарма надіється
12	1,4	1	Не кажіть дурниць.
13	1,3	2	Кажу з власного досвіду,24 вересня
14	1,7	2	А дурниці без знання пишете ви
15	1	2	В Європі тепла, навіть дуже осінь. Люди навіть не включають отоплення

Для цієї новинної публікації гістограма розподілу коментарів за емоційністю показана на рис. 5

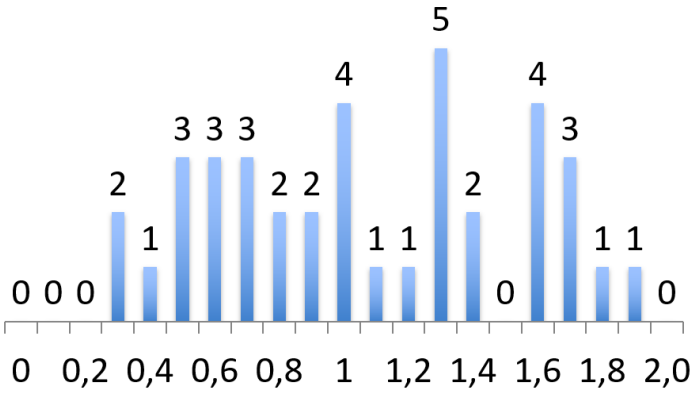

РИС. 5. Діаграма розподілу коментарів за емоційністю для новинної публікації «Путін наляканий і у відчаї покладає останню надію на зиму в Європі — CNN»

Розрахуємо за формулою (23) значення x_1:

$$x_1 = \frac{2 \cdot 0.7 + 4 \cdot 0.9 + 3 \cdot 1.0 + 1 \cdot 1.1 + 1 \cdot 1.2}{7} = 1.47. \quad (28)$$

Розрахуємо за формулою (24) значення x_2:

$$x_2 = \frac{38}{21} = 1.81. \quad (29)$$

Розрахуємо за формулою (25) коефіцієнт k:

$$k = \frac{1.81}{1.47} = 1.23. \quad (30)$$

Оскільки $k > 1$, досліджувана новинна публікація не провокує значного підвищення напруженості.

Далі розраховується значення k^* для отриманого k, але оскільки досліджувана новинна публікація взята з офіційної сторінки новинної агенції ТСН у соціальній мережі Facebook, а станом на 10 липня 2023 року сторінка ТСН є найпопулярнішою новинною сторінкою в українському сегменті Facebook (має 2,4 мільйона підписників), то $a = 1$. Отже, відповідно до формули (26), $k^* = k$.

Результати. Розглянемо переваги та недоліки розробленої системи.

Важлива перевага цієї системи це оперативність її роботи, оскільки розрахунок оновленої оцінки рівня напруженості відбувається щогодини. Окрім того, за отриманими оцінками можна спостерігати динаміку зміни СН, що якісно відрізняє отримувані результати від тих, які можна отримати за допомогою традиційних методів соціологічних досліджень.

Серед недоліків окремо слід виділити проблему спотворення результатів коментарями з фейкових облікових записів. Для розробки програми, здатної відслідковувати підозрілу активність таких облікових записів, потрібно проводити додаткові дослідження. Недоліком також є необхідність дотреновувати нейронну мережу на нових даних та оновлювати словник ключових слів, оскільки нові новинні події часто супроводжуються новою лексикою.

Серед потенційних напрямів подальших досліджень можна виділити вдосконалення архітектури нейронної мережі та пошук оптимальних способів її дотреновування.

Висновки. Напруженість у суспільстві має всеосяжний вплив на економічні, політичні, соціальні та інші сфери життя країни в цілому. Її дослідження відіграє значну роль у процесі ідентифікації загроз та ризиків, інформація про які допомагає державним посадовим особам у прийнятті рішень. Розроблена система оперативного визначення рівня напруженості в суспільстві допомагає швидко визначати новинні події та новинні публікації, які найбільше впливають на підвищення рівня соціальної напруженості у масштабах країни у певні конкретні момент часу. Використання системи з опорою на соціальні мережі дозволяє будувати на основі актуальних даних оцінки, за якими можна досліджувати динаміку зміни соціальної напруженості, пов'язаної з певною новинною подією чи новинною публікацією.

У роботі подано опис архітектури та математичного забезпечення цієї системи, розглянуто математичні методи сентимент-аналізу, зокрема, методи TF-IDF, Word2vec (CBOW) та нейронну мережу зворотного поширення помилки. Також розглянуто спосіб підрахунку середніх значень напруженості на основі оцінок емоційності коментарів. Окрім того, продемонстровано приклад підрахунку коефіцієнта k для новинної події на основі даних з соціальної мережі Facebook. Наприкінці подано основні переваги та недоліки обраного підходу визначення соціальної напруженості та визначено основні напрямки його подальшого вдосконалення.

Список літератури

1. Shchoholiev M., Andriichuk O., Tsyganok V., Tretynyk V. Decision-making and computational linguistic tools application for overall estimation of the level of social tension. *Journal of Physics: Conference Series*. 1780. 9 p.

2. Щьоголєв М.О. Методи підтримки прийняття рішень та обчислювальної лінгвістики для визначення загальної оцінки рівня напруженості в суспільстві. *Прикладна математика та комп'ютинг*. ПМК, 2022 : XV Наук. конф. магістрантів та аспірантів, Київ, 16–18 лист. 2022 р. : зб. тез. доп. С. 192–200.
3. Рудаченко О., Клебанова Т. Сучасні підходи до аналізу соціальної напруженості. *Науковий вісник Ужгородського національного університету. Серія: Міжнародні економічні відносини та світове господарство*. 2020. Вип. 30. С. 140–144.
4. Slyusarevskyy M.M., Chunikhina S., Flaherty M. Social Tension as a macro indicator of the psychological well-being of society. *Wiadomości Lekarskie*. 2021. Vol. **74**, No. 11, part 1. P. 2813–2817. https://wiadlek.pl/wp-content/uploads/archive/2021/WLek202111123.pdf
5. Artemov G., Aleinikov A., Daur A., Pinkevich A. Social Tension: the Possibility of Conflict Diagnosis (on the Example of St. Petersburg). *Economics and Sociology*. 2017. Vol. **10**, No. 1. P. 192–208. https://www.researchgate.net/publication/316256578_Social_Tension_the_Possibility_of_Conflict_Diagnosis_on_the_Example_of_St_Petersburg
6. Баранова Г., Фролов В., Кондрашин А. Особенности социальной напряженности в регионах России. *Социологический исследования.* 2011. № 6. С. 48–55.
7. Андрієнко О., Мордовцев С. Інтегральна оцінка напруженості регіонів. *Соціально-економічні проблеми і держава.* 2011. Вип. 13(2). С. 161–168. http://nbuv.gov.ua/UJRN/Sepid_2015_2_21
8. Яценко Л., Коломієць О. Регіональні аспекти соціальної напруженості в Україні: стан, фактори ормування та шляхи подолання. Аналітична доповідь. 2015. https://niss.gov.ua/sites/default/files/2015-09/socialna_napruzhenist-fc95d.pdf
9. Быковский В. Социальная напряженность на муниципальном уровне: методика оценки работы администрации. *Социологические исследования.* 2005. № 10. С. 22–26.
10. Iqbal A., Amin R., Iqbal J., Alroobaea R., Binmahfoudh A., Hussain M. Sentiment Analysis of Consumer Reviews Using Deep Learning. *Sustainability*. 2022. Vol. **14**, No. 17. https://www.mdpi.com/2071-1050/14/17/10844
11. Sri Devi T., Dhanalakshmi R., Sankar S. An Improved Framework for Sentiment Analysis for Collage Rewiews. *International Journal of Advanced Trends in Computer Science and Engineering*. 2020. Vol. **9**, No. 2. P. 1959–1963. https://doi.org/10.30534/ijatcse/2020/162922020
12. Bhimanadham H., Areti S. A., Kolluri B. S., Yanam P., Peda Gopi A. Sentimental Analysis on Movie Review System using deep learning approach. *Juni Khyat*. 2021. Vol. 11, issue 1. P. 371–380. http://junikhyatjournal.in/no_1_Online_21/47.pdf
13. Sri Lalitha Y., Vijendar Reddy G., Swapnika K., Akunuri R., Jahagirdar H. K. Analysis of Customer Review using Deep Neural Network. *2022 First International Conference on Artificial Intelligence Trends and Pattern Recognition (ICAITPR).* 2022. 5 p. https://doi.org/10.1109/ICAITPR51569.2022.9844183
14. Bekesiene S., Smaliukiene R., Vaicaitiene R. Using Artificial Neural Networks in Predicting the Level of Stress among Military Conscripts. *Mathematics*. 2021. Vol. **9**, No. 6. 626. 23 p. https://doi.org/10.3390/math9060626
15. Donchenko D., Ovchar N., Sadovnikova N., Parygin D., Shabalina O., Ather D. Analysis of Comments of Users of Social Networks to Assess the Level of Social Tension. *Procedia Computer Science*. 2017. No. 119. P. 359–367. https://doi.org/10.1016/j.procs.2017.11.195
16. Naga Mounika S., Kanumuri P. K., Narasimha rao K., Manne S. Detection of Stress Levels in Students using Social Media Feed. *Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2019).* 2020. P. 1178–1183. https://doi.org/10.1109/ICCS45141.2019.9065720
17. Chakraborty S., Talukdar M. B. U., Adib M. Y. M., Mitra S., Alam Md. G. R. LSTM-ANN Based Price Hike Sentiment Analysis from Bangla Social Media Comments. *2022 25th International Conference on Computer and Information Technology (ICCIT).* 2023. P. 733–738. https://doi.org/10.1109/ICCIT57492.2022.10055290
18. Selvadass S., Malin Bruntha P., Priyadharsini K. Stress Analysis in Social Media using ML Algorithms. *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT).* 2022. P. 1502–1506. https://doi.org/10.1109/ICSSIT53264.2022.9716396
19. Ahmed U., Lin J. C.-W. Deep Explainable Hate Speech Active Learning on Social-Media. *IEEE Transactions on Computational Social Systems*. 2022. P. 1–11. https://doi.org/10.1109/TCSS.2022.3165136
20. Danuri M. S. N., Rahman R. A., Mohamed I., Amin A. The Improvement of Stress Level Detection in Twitter: Imbalance Classification Using SMOTE. *2022 IEEE International Conference on Computing (ICOCO).* 2023. P. 294–298. https://doi.org/10.1109/ICOCO56118.2022.10031684
21. Ghosh S., Anwar T. Depression Intensity Estimation via Social Media: A Deep Learning Approach. *IEEE Transactions on Computational Social Systems*. 2021. Vol. **8**, No. 6. P. 1465–1474. https://doi.org/10.1109/TCSS.2021.3084154
22. Sharma C., Saxena P. Stress Analysis for Students in Online Classes. *2021 Grace Hopper Celebration India (GHCI).* 2021. 5 p. https://doi.org/10.1109/GHCI50508.2021.9514059

23. Cheng L.-C., Tsai S.-L. Deep Learning for Automated Sentiment Analysis of Social Media. *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).* 2020. P. 1001–1004. https://doi.org/10.1145/3341161.3344821
24. Manning C.D., Raghavan P., Schütze H. Introduction to Information Retrieval. *Cambridge University Press.* 2008. 506 p. https://doi.org/10.1017/CBO9780511809071
25. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781v3.* 2013. 12 p. https://doi.org/10.48550/arXiv.1301.3781
26. Mikolov T., Le Q. V., Sutskever I. Exploiting Similarities among Languages for Machine Translation. *arXiv:1309.4168v1.* 2013. 10 p. https://doi.org/10.48550/arXiv.1309.4168
27. Rong X. word2vec Parameter Learning Explained. *arXiv:1411.2738v4.* 2016. 21 p. https://doi.org/10.48550/arXiv.1411.2738
28. Rumelhart D.E., Hinton G.E., Williams R.J. Learning representations by back-propagation errors. *Nature.* 1986. P. 533–536. https://doi.org/10.1038/323533a0

Одержано 11.07.2023

Щьоголєв Максим Олегович,
асистент кафедри прикладної математики
Національного технічного університету України
«Київський політехнічний інститут імені І. Сікорського», Київ,
https://orcid.org/0000-0002-3267-6617
shchoholiev.maksym@gmail.com

Андрійчук Олег Валентинович,
кандидат технічних наук, старший дослідник
Інституту проблем реєстрації інформації НАН України, Київ.

УДК 004[023+62] + 519.7

М.О. Щьоголєв [1]*, О.В. Андрійчук [2]

Математичні методи обробки природної мови у системі оперативного визначення рівня напруженості в суспільстві

[1] Національний технічний університет України «Київський політехнічний інститут ім. І. Сікорського», Київ
[2] Інститут проблем реєстрації інформації Національної академії наук України, Київ
** Листування: shchoholiev.maksym@gmail.com*

Вступ. Найголовнішими чинниками, що спричиняють підвищення рівня напруженості в суспільстві, є кризові явища та інформаційні операції. Сьогодні основним способом визначення рівня напруженості, що виникає у деякої групи людей у зв'язку з певною подією, вважаються соціологічні опитування. Однак цей спосіб не дозволяє отримувати детальну інформацію про динаміку зміни напруженості, пов'язаної з певними новинними подіями, та про вплив цих новинних подій за загальний рівень напруженості в суспільстві, що ускладнює процес прийняття рішень державними посадовими особами у кризових ситуаціях.

Мета роботи – підвищення ситуаційної обізнаності представників державних установ щодо поточного рівня соціальної напруженості, спровокованої кризовими явищами, новинними подіями чи інформаційними операціями. Отримана інформація допомагатиме державним посадовим особам приймати швидкі рішення для подолання цих кризових явищ та протидії дезінформації.

Основна задача дослідження – розробити архітектуру та математичне забезпечення системи оперативного визначення рівня напруженості в суспільстві за даними з соціальних мереж.

Результати. Розроблено архітектуру та математичне забезпечення системи оперативного визначення рівня напруженості в суспільстві. Продемонстровано приклад застосування цієї системи для визначення рівня напруженості, яку провокує одна новинна публікація. Визначено основні переваги, недоліки розробленої системи, а також напрямки подальших досліджень.

Висновки. Розроблена система оперативного визначення рівня напруженості в суспільстві допомагає швидко визначати новинні події та новинні публікації, які найбільше впливають на підвищення рівня соціальної напруженості у масштабах країни у певні конкретні момент часу. Використання системи з

опорою на соціальні мережі дозволяє будувати на основі актуальних даних оцінки, за якими можна досліджувати динаміку зміни соціальної напруженості, пов'язаної з певною новинною подією чи новинною публікацією.

Ключові слова: рівень напруженості в суспільстві, соціальні мережі, сентимент-аналіз, TF-IDF, Word2vec, нейронні мережі.

MSC 68U15

Maksym Shchoholiev [1]*, Oleh Andriichuk [2]

Mathematical Methods of Natural Language Processing in the System of Operative Determination of the Level of Tension in Society

[1] *National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv*
[2] *Institute for Information Recording of National Academy of Sciences of Ukraine, Kyiv*
* *Correspondence: shchoholiev.maksym@gmail.com*

Introduction. The most important factors causing an increase in the level of tension in society are crisis phenomena and information operations. Today, sociological surveys are considered the main way to determine the level of tension that arises in some group of people in connection with a certain event. However, this method does not allow obtaining detailed information about the dynamics of changes in tension associated with certain news events and the impact of these news events on the general level of tension in society, which complicates the decision-making process by government officials in crisis situations.

The purpose of the work is to increase the situational awareness of representatives of state institutions regarding the current level of social tension provoked by crisis phenomena, news events or information operations. The information obtained will help government officials to make quick decisions to overcome these crisis phenomena and counter disinformation.

The main task of the research is to develop the architecture and mathematical support of the system of operative determination of the level of tension in society based on data from social networks.

Results. The architecture and mathematical support of the system of operative determination of the level of tension in society were developed. An example of the application of this system to determine the level of tension provoked by one news publication is demonstrated. The main advantages and disadvantages of the developed system, as well as directions for further research, are determined.

Conclusions. The developed system of operative determination of the level of tension in society helps to quickly identify news events and news publications that have the greatest impact on increasing the level of social tension across the country at certain specific moments of time. The use of a system based on social networks makes it possible to build on the basis of current data such assessments, which can be used to study the dynamics of changes in social tension associated with a certain news event or news publication.

Keywords: level of tension in society, social networks, sentiment analysis, TF-IDF, Word2vec, neural networks.