Сучасний стан наукових досліджень та технологій в промисловості. 2023. № 1 (23)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

УДК 004.89;004.94
DOI: https://doi.org/10.30837/ITSSI.2023.23.025

Т. Батюк, Д. Досин

ІМПЛЕМЕНТАЦІЯ ІНТЕЛЕКТУАЛЬНОЇ СИСТЕМИ АНАЛІЗУ ТОНАЛЬНОСТІ ТА КЛАСТЕРИЗАЦІЇ ПУБЛІКАЦІЙ У СОЦІАЛЬНІЙ МЕРЕЖІ TWITTER

Завдяки інтенсивному розвитку соціальних мереж постійно зростає популярність обміну короткими текстовими повідомленнями, тональність яких може бути чутливим індикатором суспільних настроїв і важливих соціальних явищ, цікавих для соціологів, політиків, економістів і фахівців інших галузей. У зв'язку з цим завдання автоматизації опрацювання таких природномовних повідомлень має вагомий науковий і практичний інтерес. **Об'єктом** дослідження є тональність користувацьких публікацій у соціальній мережі *Twitter*. Завдяки широкій популярності цієї соціальної мережі та великій кількості повідомлень, лаконічних за своєю сутністю, можна зручно визначати настрій користувацьких публікацій та об'єднувати їх у кластери відповідно до заданих параметрів інтелектуальної системи. **Предметом** дослідження є методи й алгоритми аналізу тональності великих масивів повідомлень, що містять необхідні ключові слова та стосуються конкретної теми, визначення факторів та розподілів тональності повідомлень, з огляду на вхідний масив даних системи, поділ повідомлень на основні групи та надання оцінок у визначених межах кожній групі, поділ на кластери відповідно до отриманої точки пошуку та відображення отриманих результатів у потрібному форматі. **Мета роботи** – реалізація інтелектуальної системи аналізу тональності та кластеризації публікацій на основі рекурентної нейронної мережі довгої короткочасної пам'яті (LSTM) та алгоритму кластеризації *k-means*. У статті передбачається вирішити такі основні **завдання:** проаналізувати найбільш уживані та найновіші алгоритми, методи, підходи та засоби імплементації завдань аналізу тональності й кластеризації публікацій у соціальних мережах; розробити концептуальну структуру інтелектуальної системи аналізу тональності й кластеризації публікацій; сформувати функціональні завдання до ключових модулів створюваної інтелектуальної системи аналізу тональності й кластеризації публікацій у соціальній мережі *Twitter*; реалізувати інтелектуальну систему аналізу тональності й кластеризації публікацій на основі рекурентної нейронної мережі та алгоритму кластеризації *k-means* і експериментально її перевірити. У роботі застосовано **методи:** рекурентна нейронна мережа довгої короткочасної пам'яті; алгоритм кластеризації *k-means*. Здобуто такі **результати**: проаналізовано, спроєктовано й побудовано загальну структуру інтелектуальної системи аналізу тональності й кластеризації публікацій. Основним завданням створення системи насамперед було покращення рекурентної нейронної мережі довгої короткочасної пам'яті, що, завдяки вдосконаленому алгоритму, суттєво полегшує опрацювання повідомлення обробниками природної мови відповідно до текстових даних певного розміру. Також одночасно використано особливий алгоритм кластеризації, а саме *k-means*, завдяки чому вдалося змінити загальний підхід до кластеризації та створення фінальних кластерів, відповідно до здобутих результатів роботи рекурентної нейронної мережі. **Висновки:** унаслідок застосування комбінації LSTM нейронної мережі та алгоритму кластеризації *k-means* вдалося прискорити процес аналізу тональності й кластеризації публікацій у соціальній мережі *Twitter* на 10...15% порівняно з аналогічними згортковими нейронними мережами та ієрархічною кластеризацією.

Ключові слова: нейронна мережа; LSTM; аналіз тональності публікацій; кластерний аналіз; соціальна мережа *Twitter*.

Вступ

Створення та імплементація інтелектуальної системи аналізу тональності та кластеризації публікацій є актуальним та перспективним завданням, оскільки сьогодні більшість комунікацій між людьми відбувається саме в соціальних мережах. Кожне повідомлення користувача соціальної мережі має конкретну семантику, відображає певні думки автора й аналіз відповідної ситуації та/або є реакцією на певну подію. Сучасні алгоритми та підходи до аналізу даних дають змогу ефективно та відносно просто здійснювати аналіз значних обсягів текстових даних, завдяки чому можна визначати середню тональність реакцій користувачів на певні події і, як наслідок, робити висновки щодо проаналізованого контенту. Таким чином можна зрозуміти відношення різних груп користувачів до певного виду контенту, товарів та інших ринкових пропозицій, і великі бренди та корпорації активно застосовують ці підходи до аналізу інформації шляхом визначення тональності користувацького тексту й подальшого поділу цього тексту й, відповідно, самих користувачів на кластери для подальшої роботи з отриманими кластерами.

Для реалізації такого роду алгоритмів насамперед необхідно визначити, яких саме користувачів ми хочемо дослідити, тобто необхідно мати певні ключові слова, за якими буде здійснюватися пошук,

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2023. No. 1 (23)

геолокація, установлюватися теги тощо. Саме це є початковою точкою поточного дослідження, далі необхідно завантажити потрібний датасет, тобто пул користувацьких повідомлень, збережених відповідно до заданого предиката. Коли надійшов файл зі збереженими повідомленнями, необхідно виконати загальну перевірку файлу на коректність даних і після цього програмно структурувати дані таким чином, щоб можна було оперувати ними максимально точно, тобто потрібно дані відформатувати та привести до єдиної структури. Після цього здійснюється аналіз користувацьких повідомлень. Загальний алгоритм можна поділити на два основні алгоритми: аналіз тональності користувацьких повідомлень, з огляду на певний підхід, де кожне повідомлення має свій рівень тональності та рейтинг, і також здійснити кластеризацію повідомлень для поділу користувачів та їхніх повідомлень на визначену в процесі роботи інтелектуальної системи кількість кластерів. Здобуту інформацію найзручніше показувати у вигляді графіків та діаграм для наочності й розуміння загальної ситуації в результатах аналізу користувацьких публікацій. Ці алгоритми є не новими і містять широкі можливості для модифікації та оптимізації, що й буде здійснено в подальшій роботі.

Аналіз останніх досліджень і публікацій

Нейронні мережі вже практично стали незамінною частиною роботи різноманітних компаній та корпорацій. У статті [1] визначено, яку саме роль виконує *deep learning* в електронній комерції та принципи роботи з користувачами на прикладі мереж онлайн-магазинів. Авторами досліджено відгуки людей щодо певних товарів різної якості та вплив поточних відгуків на продаж товарів іншими користувачами системи в майбутньому. З'ясовано, коли саме й за яких умов люди найбільше чи найменше звертають увагу на відгуки про товар і, відповідно, як саме складений позитивний чи негативний відгук може вплинути на купівельну привабливість товару. У статті [2] автори дослідили коментарі користувачів під відеозаписами в соціальній мережі *YouTube*, був обраний датасет з відеозаписами про вірус COVID-19 і коментарі, досліджено тональність написаних повідомлень, їхню частоту й активність написання користувачами. Відповідно припущено, що люди, які писали негативні коментарі, є ботами. Це пояснюється дуже схожими шаблонами повідомлень і майже однаковим рівнем негативної тональності. Навпаки, автори статті [3] розглянули коментарі відомих особистостей у соціальних мережах і знайшли значну кількість коментарів, що написані в певні моменти часу й мали практично однакову позитивну тональність у межах від 0 до 1. Це також доводило неприродний стан цих повідомлень.

У статті [4] досліджено декілька типів моделей персоніфікації користувачів *Twitter* за допомогою нейронних мереж LSTM. З відкритого API було взято декілька ключових параметрів щодо кожного користувача й розглянуто у вигляді окремого датасету з подальшим поділом користувачів на групи відповідно до їхньої локації, опису та аватара профілю, було навчено відповідну нейронну мережу, що розподіляє користувачів на певні груп. Також створено систему [5], що опрацьовує цитування в статтях і надає змогу аналізувати правильність і коректність поточного тексту й виправляти його за допомогою рекурентної нейронної мережі, основним завданням якої є опрацювання текстових даних і аналіз наступних даних на основі навчання з вчителем. У роботі [6] розглянуто машинне навчання на основі відгуків клієнтів на готелі для розуміння ситуації на цьому ринку та подальшого складання плану розвитку готелів завдяки позитивним і негативним відгукам клієнтів. Для роботи використовувалися згорткові нейронні мережі, оскільки кількість даних була відносно невеликою. Автори статті [7] розглянули інформаційну систему аналізу тональності тексту, розміщеного в оголошеннях товарів електронної комерції. За допомогою згорткової нейронної мережі досліджено оголошення товарів, на яких навчалася нейронна мережа, і визначено, які саме сучасні оголошення викликають найбільш позитивні й, навпаки, негативні враження в потенційних клієнтів. У роботі [8] вивчено основні аспекти ментального здоров'я людей на основі їхньої реакції на конкретні події та явища, залишені під публікаціями в соціальних мережах. Завдяки навчанню з учителем створено дві різні моделі аналізу користувацьких даних, на основі яких в подальшому опрацьовувалися коментарі й публікації людей і визначалися особливості їхнього ментального здоров'я шляхом аналізу тональності тексту з використанням згорткової нейронної мережі.

Виокремлення невирішених частин загальної проблеми. Мета роботи

Метою цієї статті є впровадження інтелектуальної системи аналізу тональності та кластеризації

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2023. №1 (23) ISSN 2524-2296 (online)

публікацій у соціальній мережі *Twitter*. Ідея аналізувати тональність користувацьких повідомлень чи публікацій у соціальних мережах не є новою, адже існує низка практично реалізованих систем, що виконують подібне завдання. На сьогодні важливим завданням є оптимізація та максимально ефективне використання наявних технологій і правильний вибір моделей та алгоритмів для виконання конкретного завдання, що може залежати як від розміру вхідного датасету, так і від розмірів окремих текстових токенів [9] усередині датасету, або навіть параметрів пошуку текстової інформації для користувачів, що застосовують певну систему, або лише її часткового функціоналу для дослідження тексту.

Сам процес аналітики текстових публікацій або коментарів користувачів можна чітко поділити на дві важливі частини: це аналіз тональності тексту й кластеризація текстових даних. Це завдання є нетривіальним і досить складним, оскільки існує безліч параметрів, які потрібно враховувати перед створенням такої інтелектуальної системи для аналізу текстових даних: це і розмір вибірки, і текстові дані, і контекст, якому ці дані належать. Це також користувачі, які пишуть публікації чи повідомлення як реакцію на певну подію чи сукупність подій, що відбувається в певний момент часу. Усе це означає, що створити унікальну інтелектуальну систему, яка зможе врахувати всі параметри й приблизно однаково ефективно аналізувати будь-які текстові дані, не можливо. Так чи інакше, існують завдання, що кожна конкретна система зможе виконувати ефективно та з максимальною точністю, але й завдання, які ефективно чи точно виконати тим самим набором алгоритмів просто не вдасться. Так, найчастіше для аналізу тональності тексту використовують згорткові нейронні мережі, а для кластерного аналізу – ієрархічну кластеризацію. Ці алгоритми є ефективними й перевіреними часом, але можуть виконувати лише аналіз невеликих і середніх за обсягом датасетів чи вибірок даних. Ці алгоритми можна без проблем використовувати і для досить значних обсягів даних, але такий аналіз буде неефективний і менш точний, особливо якщо кількість текстових одиниць даних є великою, а сам токенізований об’єкт не великий [10]. Яскравим прикладом є соціальна мережа *Twitter*, де одне користувацьке повідомлення може мати максимально 280 символів для створення публікації.

Для того щоб найбільш ефективно аналізувати тональність повідомлень та їхню кластеризацію, будемо використовувати нейронну мережу LSTM та алгоритм кластеризації *k-means*, завдяки чому можна досягти на 10...15% більшу ефективність порівняно зі згортковими нейронними мережами та ієрархічною кластеризацією. Перед тим як описувати функціонал системи та основні алгоритми *long-short-term memory* нейронної мережі та алгоритм кластеризації *k-means*, варто звернути увагу на невирішені проблеми, а саме на те, чому в контексті користувацьких публікацій і коментарів у соціальній мережі *Twitter* згорткові нейронні мережі та звичайна ієрархічна кластеризація є неефективними та можуть унаслідок аналізу тональності тексту й подальшої кластеризації показувати неочікувані результати [11]. Це відбувається саме через особливості реалізації цих алгоритмів роботи з текстом та через властивості ваг, що надає згорткова нейронна мережа та які є важливими для врахування, оскільки в разі помилкового прорахунку ваг згенерована модель може бути спотвореною та, відповідно, недійсною.

Згорткову нейронну мережу, або CNN, можна уявити як сукупність матриць, що складаються в одну велику матрицю, у якій векторами можна обирати для навчання як горизонтальні, так і вертикальні множини елементів, сформовані в процесі навчання нейронної мережі для побудови моделі. Під час роботи здійснюється накладання сукупності слів одне на одного за допомогою векторів, оскільки кожне слово є окремим вектором букв, що разом формують певне зображення. За допомогою створених системних предикатів відбувається фільтрація сформованого зображення векторів, що є сукупністю текстових даних, або слів. Оскільки предикат має однакову ширину та довжину перевірки значень, то можна проаналізувати лише частину вектора. Тому для ефективності роботи вектори найчастіше складають у тимчасові матриці для коректної та більш ефективної фільтрації за допомогою заданого предиката [12]. Оскільки це є завданням оброблення природної мови, потрібно зауважити, що фільтри, які ми використовуємо в процесі роботи, стандартно мають ту саму ширину, що й довжина досліджуваного текстового елемента, або окремого слова. Висота ж є більш статичною та зазвичай може змінювати свій розмір від 0 до 5. Оскільки з отриманої текстової інформації потрібно формувати певні послідовності значень, що в подальшому будуть записані в списки значень, відповідно необхідно обмежити кількість

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2023. No. 1 (23)

списків до 5 штук для того, щоб можна було здійснювати ефективне паралельне оброблення текстової інформації та навчання моделі із заданими обмеженнями роботи поточної нейронної мережі [13]. Через ці алгоритмічні обмеження виникає основна проблема – неможливість точного навчання й подальшого аналізу тексту за допомогою згорткових нейронних мереж. Текст з обмеженнями буде проаналізований і результат аналізу його тональності буде коректний, але через обмеження висоти вектора нейронна мережа може навчатися лише на публікаціях і повідомленнях із полярними значеннями тональності, тобто –1, 0 або 1 [14], відповідно, явно негативні, нейтральні чи позитивні коментарі без можливості їхнього розподілу за певним діапазоном, що, з одного боку, не є критичним і може бути корисним для загального розуміння тональності повідомлень, але, з іншого, з допомогою такої згорткової нейронної мережі неможливо здійснювати точний аналіз тональності. Також виникають застереження щодо швидкості роботи, адже така нейронна мережа має невисокі швидкість та ефективність, оскільки для збереження полярних значень векторів використовується алгоритм мемоїзації повторюваних даних.

Якщо говорити про ієрархічну кластеризацію, то вона теж має недоліки у використанні. Зокрема найбільший із них схожий на той, що був описаний раніше щодо згорткової нейронної мережі, а саме – неможливість виконувати складні завдання, а також проблема недостатньої швидкості та ефективності роботи такого алгоритму кластеризації. Для початку варто зазначити, що ієрархічна кластеризація формується на основі деревоподібного графа, який ще відомий під назвою "дендрограма" і, відповідно, у процесі побудови цього графа використовуємо агломеративний підхід роботи з вхідними даними [15]. Маючи дендрограму й застосовуючи агломеративний підхід, спостерігаємо відносно однотонні кластери й паралельно шукаємо можливі зв’язки між ними, на другому кроці послідовно здійснюємо об’єднання кластерів в окремий "зв’язок" на основі необхідних предикатів. Основною перевагою є простота й наочність у використанні, тобто наявність зручної деревоподібної структури, в якій кластери найчастіше за умови правильної генерації моделі можна побачити навіть "на око" [16]. На жаль, недолік цієї простоти полягає в тому, що найчастіше це можна використати лише для незначних обсягів даних і невеликих датасетів. Одним з основних недоліків є складність деревоподібної ієрархії, що кластеризується в часі з алгоритмічною складністю $O\left(n^2 log\, n\right)$, де n – кількість загальних точок даних. Якщо ж на противагу взяти алгоритм *k-means*, у ньому будемо використовувати оптимізацію конкретної цільової функції, наприклад, у межах певного діапазону значень від k до l, тобто ми не маємо фактично цільової функції, а тому набагато ефективніше буде виконання через складність алгоритму $O\left(nKm\right)$, де K – це кількість кластерів, а m – кількість середніх значень [17]. Також проблемою ієрархічної кластеризації є певна статичність, а саме неможливість скасовувати попередні кроки виконання алгоритму. Тобто, якщо ми здійснимо кластеризацію n *-1* точок, а після цього виявиться, що з’єднання між кластерами було неправильним або виникли проблеми у створенні деревоподібного графа на одному із рівнів ієрархії, то ми не можемо скасувати цей крок на програмному рівні під час виконання, адже буде отримано фінальну дендрограму зі спотвореними значеннями або потрібно буде зупиняти роботу програми й запускати систему від початку, що теж є великим недоліком порівняно з роботою алгоритму *k-means*, де є змога перевіряти правильність середніх значень за допомогою заданої умови виконання [18].

З огляду на наведену вище інформацію, можна зробити висновок, що використання згорткової нейронної мережі та ієрархічної кластеризації є неоптимальним для аналізу тональності користувацьких публікацій і повідомлень у соціальній мережі *Twitter*. Натомість для оптимізації та пришвидшення роботи інтелектуальної системи доцільніше використовувати нейронну мережу з архітектурою LSTM та алгоритм кластеризації *k-means*.

Функціонал системи

Говорячи про функціонал створюваної інтелектуальної системи аналізу тональності та кластеризації публікацій у соціальній мережі *Twitter* насамперед варто звернути увагу на два основні алгоритми, що й будуть виконувати основний обсяг роботи в середині системи. Насамперед це нейронна мережа LSTM, за допомогою якої буде здійснюватися ефективний та швидкий аналіз тональності публікацій і коментарів користувачів *Twitter*, й алгоритм кластеризації *k-means*, за допомогою якого будуть виділятися основні кластери та розподілятися користувацький текст за цими кластерами.

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2023. №1 (23) ISSN 2524-2296 (online)

Для початку варто розкрити сутність нейронної мережі, яку було побудовано. Однією з найпопулярніших і найефективніших моделей нейромереж, орієнтованих на опрацювання часових рядів, є модель *Long Short-Term Memory* (LTSM). Вона є ефективною у використанні й набагато точнішою, ніж згадана раніше згорткова нейронна мережа, але водночас вона є складною в реалізації, оскільки це рекурентна нейронна мережа, основна функція якої полягає в прогнозуванні послідовності даних за вхідним датасетом і відповідної проблеми, яку необхідно вирішити. Для кращого розуміння контексту застосування нейронної мережі варто сформулювати, що мається на увазі під аналізом часових рядів, для виконання якого будується нейронна мережа. Сутність полягає в тому, що вхідні дані є конкретними точками інформації, що аналізується у певних часових проміжках, завдяки чому можна створювати та аналізувати закономірності процесів, що відбуваються у конкретних часових проміжках. Оскільки є певні пов'язані між собою точки даних, використовуємо рекурентну нейронну мережу (RNN). В основі рекурентних мереж лежить концепція комірок пам'яті, у яких зберігаються певні точки подання даних. У традиційних RNN є недолік, який полягає в тому, що в якийсь момент роботи мережі кількість точок подання даних може збільшитися настільки, що нові дані неможливо буде запам'ятати, через що ймовірне спотворення кінцевих результатів, адже зрештою нейронна мережа пропускатиме важливий текст і не створить для нього точку подання. На рис. 1 зображено загальну структуру рекурентної нейронної мережі.

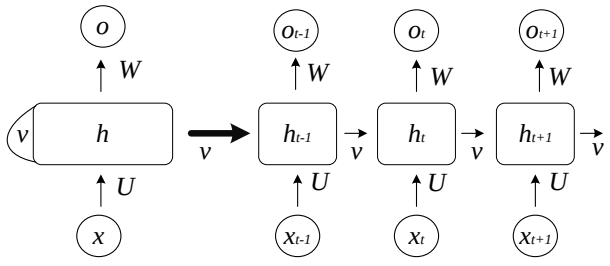

Рис. 1. Загальна структура рекурентної нейронної мережі

Нейронна мережа *Long Short-Term Memory*, або LSTM, – це окремий і особливий випадок рекурентних нейронних мереж, оскільки ця модель є ще ефективнішою, адже може зберігати сукупність інформації протягом тривалішого часу. У нейронній мережі LSTM подолано дві основні рекурентні проблеми: це градієнти, що зникають, і градієнти, які деформуються в процесі роботи інтелектуальної системи. Модель LSTM містить стан комірки пам'яті та три основних проходи. Стан комірки пам'яті зберігає значення комірки протягом певного проміжку часу та є стрічкою, що рухається й лінійно передає дані далі по конвеєру практично без додаткових деформацій. У моделі мережі LSTM є змога додавати, змінювати й видаляти інформацію за допомогою вже згаданих раніше трьох проходів. Проходи допомагають здійснювати регуляцію інформації та є ключовими в архітектурі нейронної мережі LSTM, оскільки завдяки стану комірки пам'яті потік даних формується в певну лінійну структуру й дозволяє здійснювати рівномірний розподіл пам'яті протягом всього часу роботи нейронної мережі в інтелектуальній системі. Нейронна мережа перебуває в трьох основних станах: або дані подаються на вхід, або на вихід, або забуваються через спотвореність чи непотрібність інформації.

Відповідно, робота нейронної мережі реалізована за допомогою згаданих раніше трьох проходів. Першим виконується прохід "забування", що відповідає за видалення інформації, яка через спотвореність чи використаність більше не потрібна для аналізу тональності тексту, отже, її можна видалити та звільнити місце для наступної інформації на вході. Завдяки цьому на кожному кроці роботи нейронної мережі модель стає ефективнішою. Цей прохід має два основні входи: це прихований стан попередньої комірки пам'яті та поточне введення інформації на певному кроці. Ці дані перемножуються зі створеними раніше матрицями ваг, після чого до них додається певний коефіцієнт зміщення. Далі застосовується сигмоїдна функція, що продукує результат від 0 до 1, завдяки чому нейронна мережа "знає", яку інформацію можна "забути", а яку передати далі. Якщо ми маємо значення 0, інформацію про такий стан комірки пам'яті можна видалити, якщо значення наближається до 1, то всю інформацію про комірку потрібно зберегти й передати далі. Векторний вихід сигмоїдної функції, отриманий унаслідок відпрацювання частини моделі нейронної мережі, потрібно перемножити на стан комірок пам'яті, які не було видалено в процесі роботи, а результат передати далі, унаслідок чого формується перший прохід, що виконує важливий функціонал видалення всіх непотрібних для аналізу тональності тексту комірок пам'яті. На рис. 2 зображено основні аспекти архітектури LSTM нейронної мережі.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2023. No. 1 (23)

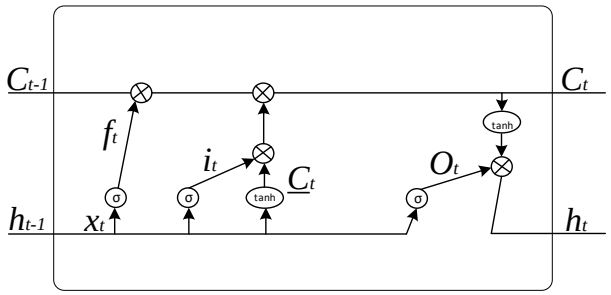

Рис. 2. Архітектура *Long Short-Term Memory* нейронної мережі

Другим іде "вхідний" прохід, який використовується для того, щоб додавати інформацію до стану комірки пам'яті. Спочатку виділені значення, які необхідно додати до комірки регулюються за допомогою сигмоїдної функції, вхідними даними все ще є прихований стан попередньої комірки пам'яті та поточне введення інформації на кроці роботи алгоритму. У процесі роботи створюється вектор, що містить всі або майже всі можливі значення, які необхідно додати до стану комірки пам'яті. Це відбувається з використанням функції *tanh*, відповідно тангенс виводить значення від –1 до 1. Далі необхідно визначити значення сигмоїдної функції, що є регуляторною. Наявні значення регуляторної функції необхідно помножити на створений раніше вектор значень. Уся важлива для інтелектуальної системи інформація з метою аналізу тональності певного тексту додається до стану комірки пам'яті за допомогою операції додавання даних. Завдяки цьому можна бути впевненим, що ще через "вхідний" прохід було додано лише корисну та потрібну інформацію, яка була попередньо відфільтрована та перевірена в процесі роботи мережі.

Останнім іде "вихідний" прохід, який необхідний для того, щоб використовувати поточну інформацію, яка є доступною в конкретний момент часу, й відображати найбільш релевантні результати. Спочатку створюється окремий вектор після застосування функції *tanh* до стану комірки пам'яті, де значення виходу коливається від –1 до 1. Сама ж сигмоїдна функція знову виконує роль регуляторної функції, тобто використовується для регулювання саме тих значень, що далі потрібно вивести з вектора за допомогою двох згаданих раніше входів, а саме прихованого стану попередньої комірки пам'яті та поточного введення інформації на певному кроці. Значення сигмоїдної функції необхідно перемножити на вектор і отриманий результат операції йде як вихідне значення. Також нейронна мережа надсилає результат у прихований стан наступної комірки пам'яті, що відповідно є сучасним рішенням, завдяки чому нейронна мережа LSTM є найефективнішою та найзручнішою в прогнозуванні послідовностей і особливо добре себе показала, виконуючи завдання аналізу тональності публікацій у соціальній мережі *Twitter*.

Також варто уточнити важливість сигмоїдної функції активації саме під час роботи проходу "забування", оскільки замість того, щоб розподіляти значення між –1 та 1, здійснюється розподілення вхідних значень між 0 та 1. Це допомагає вчасно оновлювати змінені дані або "забувати" спотворені, тобто ті, що більше не матимуть жодної користі в процесі аналізу тональності тексту. Розподіл здійснюється саме між 0 та 1 через математичне множення, оскільки будь-яке число, яке помножимо на 0, унаслідок буде 0, так і навпаки, будь-яке число, помножене на 1, залишиться незмінним. Завдяки цьому сигмоїдна функція допомагає ефективно визначити, які дані необхідно "забути" або видалити, які лише оновити на поточне значення тональності, а які зберігаються й передаються в наступний "вхідний" прохід. Отже, на рис. 3 зображено покроковий алгоритм роботи *Long Short-Term Memory* нейронної мережі.

Окрім реалізації нейронної мережі LSTM, важливою частиною інтелектуальної системи аналізу тональності та кластеризації публікацій у соціальній мережі *Twitter* є правильна імплементація кластеризації з використанням алгоритму *k-means*. Як вже було описано раніше, звичайна ієрархічна кластеризація, хоч і є досить популярною та має чітко виражені переваги, все ж не відповідає нашому завданню, а саме для оброблення точних значень тональності тексту й, відповідно, виділення необхідних кластерів. Для початку варто уточнити, що кластеризація – це процес поділу обсягу даних певного розміру на декілька чітко визначених аналогічних за своєю структурою груп таким чином, щоб значення точок даних в одній групі були більш схожими на значення інших точок даних у тій самій групі, ніж інші точки даних, що належать до інших груп. Також варто зазначити, що кластеризація – це алгоритм навчання без учителя через свої особливості роботи.

Сучасний стан наукових досліджень та технологій в промисловості. 2023. №1 (23)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

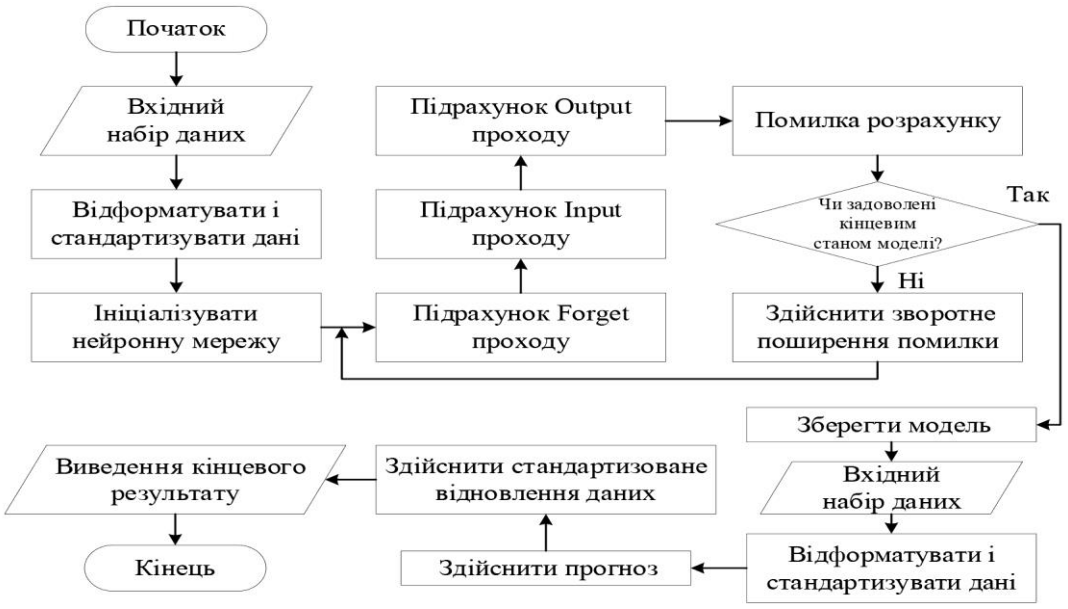

Рис. 3. Алгоритм роботи нейронної мережі LSTM

Кластеризація *k-means* – це алгоритм, що в процесі навчання не потребує позначки вхідних даних, на відміну від алгоритмів навчання з учителем. Кластеризація *k-means* поділяє об'єкти на кластери таким чином, щоб усі об'єкти всередині кластера були схожі один на одний і не схожі на об'єкти в інших кластерах. Латинська літера *k* є позначкою числа, що вказує на кількість кластерів, які необхідно створити. Також сутність алгоритму полягає в тому, щоб знайти, яке найкраще чи оптимальне значення кластерів необхідно використати для ефективної роботи з аналізом тональності. Сама процедура кластеризації *k-means* є досить простою та зрозумілою математичною задачею. Для початку необхідно визначитися із загальною нотацією. Наприклад, у нас є від C_1 до C_k наборів, що мають індекси спостереження в кожному кластері, ці набори задовольняють дві основні властивості: по-перше, кожне спостереження належить хоча б до одного з *k* наявних кластерів, по-друге, кластери не можуть перекрити один одного, тобто один набір спостережень може належати лише одному унікальному кластеру. В основі кластеризації *k-means* лежить ідея, що найкраща кластеризація – це та, де варіація значень усередині кластеру є мінімальною, тобто варіація в межах певного кластеру C_k є мірою величини спостереження кластеру, де один кластер відрізняється від іншого. Це і є проблема, що вирішує кластеризація *k-means* у межах інтелектуальної системи.

Варто наголосити на важливості параметра *k*, що визначає кількість кластерів. Часто його можна визначити певним приблизним значенням, просто прикинувши розмір датасету й дані, які він містить. Для нашого завдання – аналізу тональності тексту – обрано "ліктьовий" метод, що дає змогу точніше визначити необхідну для роботи кількість кластерів за допомогою запуску алгоритму *k-means* із різним набором кластерів з метою емпіричного визначення оптимального значення. "Ліктьовий" метод передбачає пошук певної метрики, щоб оцінити, наскільки результат кластеризації є хороший для різних *k* значень за допомогою знаходження "ліктьової" точки, щоб відділити всі непотрібні подальші значення й обрати оптимальне до визначеної точки. Різкий спад значень на відповідному графіку означає, що значення кластеризації оптимізується, але є момент, коли різкий спад значень припиняє падати й стабілізується. Це і є та сама "ліктьова" точка. Тобто всі значення після "ліктьової" точки потрібно відкинути й залишити лише ті, де є спад середньої суми квадратів значень спостережень усередині кожного кластеру, там, де спад найбільший. Ця точка скоріше за все й буде оптимальним числом кластерів.

Алгоритм кластеризації є, власне, методом поділу індексів спостереження в кожному кластері таким чином, що ціль останнього рівняння виділення *k* кластерів є мінімізованою. Проблема є досить складною, оскільки є k^n способів розділити *n* спостережень на кластери, але існує алгоритм,

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2023. No. 1 (23)

за допомогою якого можна знайти локальний оптимум для проблеми оптимізації *k-means*. Сам алгоритм кластеризації *k-means* має два глобальних кроки. Для початку необхідно обрати випадкове значення від 1 до *k* кожному зі спостережень даних, що були окремо виділені з датасету. Вони є певними початковими значеннями для кластерів. Цю процедуру необхідно повторювати доти, доки значення кластерів не перестануть змінюватися. Для кожного з *k* кластерів необхідно обчислити центроїд кластера. Отриманий центроїд *k*-го кластера є окремим вектором середніх значень індексів спостережень у *k*-му кластері. Кожному кластеру необхідно призначити ідентифікатор спостереження, центроїд якого розташований найближче. Потрібно переконатися, що всі кластери є стабільними і не містять невизначеностей, оскільки це може завадити здійсненню коректного розподілу індексів спостереження щодо кластерів. Алгоритм роботи кластеризації *k-means* зображено на рис. 4.

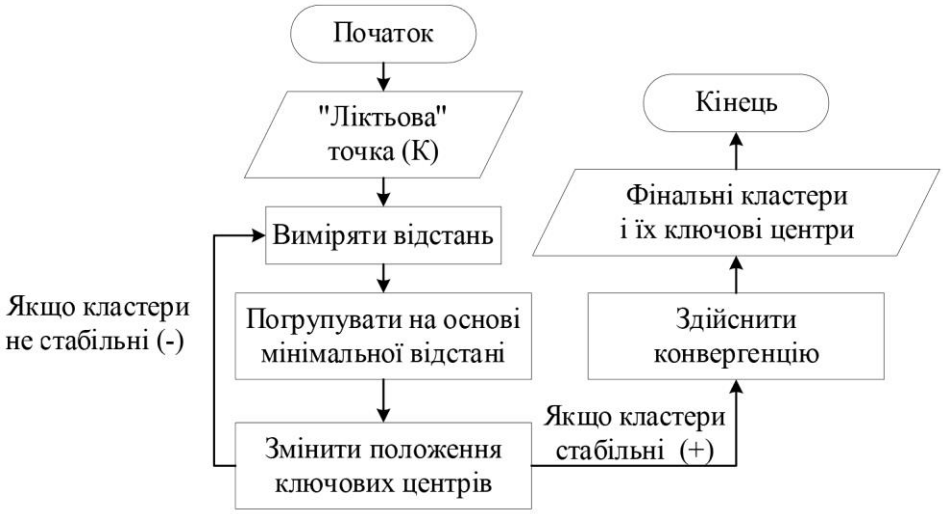

Рис. 4. Алгоритм роботи кластеризації *k-means*

Окресливши логіку та функціонал нейронної мережі LSTM та алгоритму кластеризації *k-means*, стає зрозуміло, як усередині системи буде відбуватися процес аналізу тональності та кластеризації публікацій і повідомлень користувачів соціальної мережі *Twitter*. Крім того, необхідно зрозуміти, як виглядає логіка інтелектуальної системи загалом. Нейронна мережа й алгоритми кластеризації самі по собі виконують лише певну функцію в системі. Сама система містить велику кількість процесів, що не дивно, оскільки ми взаємодіємо з реальним користувачем та відштовхуємося від заданих користувачем параметрів у певний момент часу, коли цей користувач працює із системою.

Найбільш зручною діаграмою для відображення загальної концепції взаємодії користувача, системи та інших елементів структури є діаграма варіантів використання, оскільки здебільшого вона складається з акторів та варіантів використання, які вони здійснюють. В цій інтелектуальній системі аналізу тональності та кластеризації текстової інформації є єдиний актор – "Користувач" – окрема сутність щодо системи й зображений окремо від інших варіантів використання, що є основними аспектами роботи системи. Саму діаграму варіантів використання зображено на рис. 5. На діаграмі інтелектуальну систему зображено всередині прямокутника. Вона містить декілька основних варіантів використання, що відповідають за збереження твітів – публікацій та коментарів користувачів соціальної мережі *Twitter*. Крім того, є варіанти використання, які відповідають за форматування текстової інформації до потрібного вигляду, здійснення LSTM-аналізу тональності текстових даних, виконання кластерного аналізу, виведення всіх необхідних отриманих результатів у форматі, що найбільше відповідає отриманим даним і завершенню роботи системи.

Окрім звичайних варіантів використання, що виконують ключові функції інтелектуальної системи, на діаграмі відображено варіанти включення та розширення, які більш детально пояснюють сутність роботи основних варіантів використання. Так, наприклад, варіант використання, що описує

Сучасний стан наукових досліджень та технологій в промисловості. 2023. №1 (23)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

збереження твітів, містить варіанти включення, які описують задавання параметрів у певному форматі та збереження сформованого датасету в *.csv* файл. Також міститься варіант розширення, який відтворює процес вибору основних ключових слів пошуку. Варіант використання, що описує форматування текстових даних, має лише варіанти розширення, які відтворюють видалення зайвих символів із тексту, стандартизування наявної інформації та видалення обробників тексту, що були завантажені за замовчуванням. Також інший варіант використання, який відповідає за LSTM-аналіз тональності текстових даних, має лише варіанти розширення можливостей, такі як здійснення підрахунку всіх поточних шарів мережі та стану комірки пам'яті і три основних проходи, здійснення прогнозу тональності текстової інформації, тобто публікацій і коментарів користувача *Twitter*, і кінцеве збереження створеної моделі. Далі йде варіант використання кластерного аналізу, який також містить варіанти розширення й включення, такі як пошук "ліктьової" точки, визначення фінальних кластерів і їхніх ключових центрів та імплементація конвергенції.

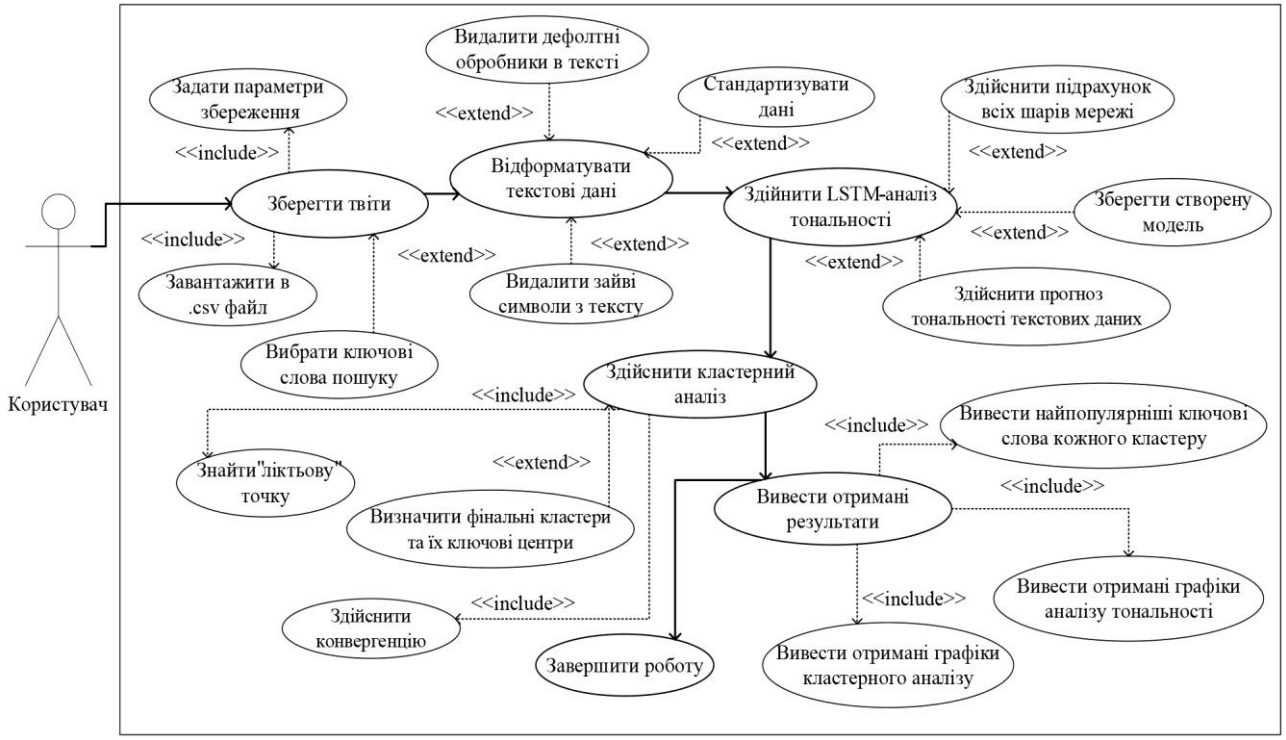

Рис. 5. Діаграма варіантів використання інтелектуальної системи аналізу тональності та кластеризації публікацій

Також варто наголосити на двох останніх варіантах використання, зокрема виведення результатів і завершення роботи інтелектуальної системи. Виведення результатів має лише варіанти включення, зокрема виведення результатів аналізу тональності користувацьких публікацій і коментарів та виведення результатів кластерного аналізу у вигляді текстової інформації, а також графіків і діаграм для більш детального пояснення отриманих даних відповідно до ключових слів пошуку.

Описавши загальну структуру інтелектуальної системи за допомогою діаграми варіантів використання, необхідно більш конкретно описати структуру створеної системи та її функціонал, для цього ідеально підходить діаграма діяльності. Вона дає змогу чітко виділити сутності системи за допомогою доріжок, умови виконання обчислень та розрахунків, чітко показати стан дії та його взаємодію з іншими станами в межах початкового й кінцевого стану інтелектуальної системи аналізу тональності та кластеризації публікацій у соціальній мережі *Twitter*. Також зручно відобразити всі наявні розгалуження потоків у процесі роботи та їхні результати. Побудована діаграма діяльності зображена на рис. 6.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2023. No. 1 (23)

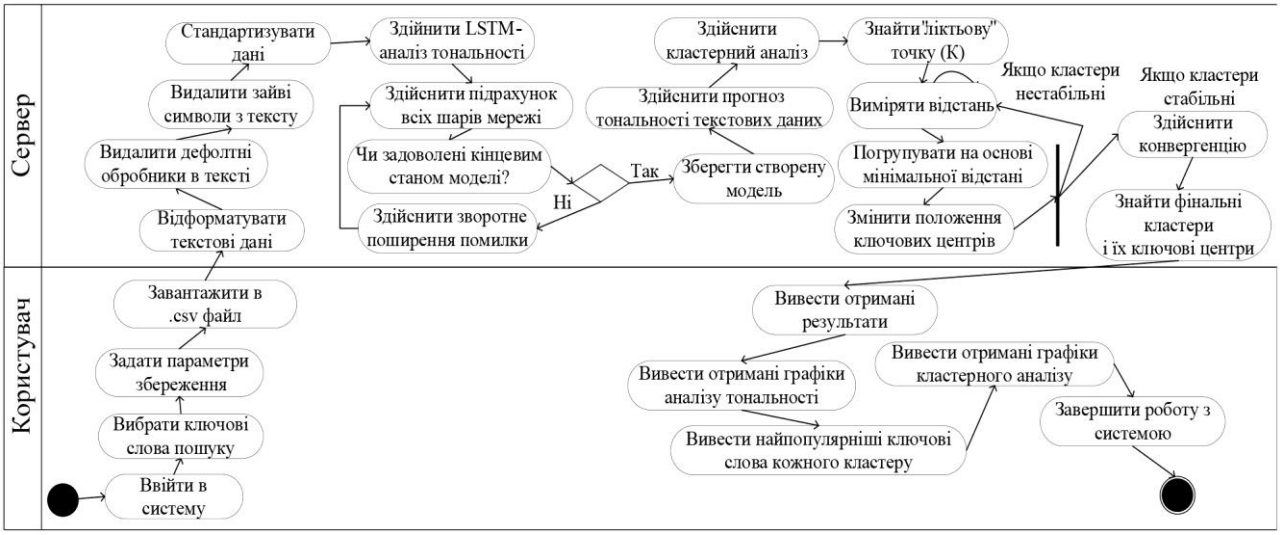

Рис. 6. Діаграма діяльності інтелектуальної системи аналізу тональності та кластеризації публікацій

Для початку варто виокремити дві основні сутності, а саме "Користувач" і "Сервер". На діаграмі діяльності вони подані у вигляді окремих доріжок, відповідно можна описати їхній контекст не лише в просторі, але й у часі, тобто зрозуміти, які процеси в інтелектуальній системі відбуватимуться в конкретний момент часу та яким чином будуть взаємодіяти між собою. Першим іде початковий стан, який описує функціональний початок роботи системи, далі послідовно між собою взаємодіють стани дій, що належать користувачу системи, тобто введення параметрів збереження, задання ключових слів та інші дії, або стани дій, які є повністю інкапсульовані в межах логіки окремого екземпляра класу виконання задачі, тобто кожен стан дії можна розглядати незалежно від інших. Також сутність "Сервер" у процесі роботи перевіряє умову на допустимість збереження створеної моделі чи на необхідність здійснення процесу зворотного поширення помилки. Окрім умов, сутність "Сервер" містить розпаралелення станів дій, де одночасно під час виконання двох різних предикатів зміна ключових центрів може розділитися як на вимір відстаней між кластерами, так і на здійснення конвергенції з уже наявними кластерами внаслідок обчислення центроїда кластеру, що містить розділені індекси спостереження в інтелектуальній системі.

Отже, описавши функціонал системи за допомогою об’єктно орієнтовних діаграм потрібно для більшого розуміння контексту роботи системи описати її основні елементи та взаємозв’язки за допомого функціональної діаграми. Найбільш зручною для виконання цього завдання є діаграма потоків даних, оскільки вона додатково показує систему як сукупність процесів, що взаємодіють між собою під час усього циклу життя системи. Діаграма потоків даних інтелектуальної системи зображена на рис. 7.

Діаграма складається із семи послідовних потоків даних, де кожен потік переходить у наступний і передає певну інформацію в межах запиту. Кожен потік виконує всі необхідні вимоги до транзакцій, тобто інтелектуальна система передає дані без ризику їхньої втрати: якщо транзакція не відбудеться або в процесі роботи буде втрачено чи спотворено інформацію, транзакція просто зупинить роботу поточного потоку й поверне керування минулому потоку, який завершився штатно без спотворення наявної користувацької інформації. Також під номером 1 та під назвою *R* на діаграмі подані сховища даних, у нашому випадку це текстовий файл із розширенням *.csv* для збереження датасету й сервер для динамічного збереження даних між запитами. Бази даних як такої не створюється, оскільки сутність створення інтелектуальної системи полягає в імплементації алгоритмів поточного аналізу тональності та кластеризації публікацій і коментарів у соціальній мережі *Twitter* відповідно до заданих ключових слів.

На підставі виконаного опису основної сутності створюваної системи, її функціоналу, реалізації нейронної мережі LSTM та алгоритму кластеризації *k-means* було проведено експеримент зі створення повноцінного датасету з публікацій і коментарів у соціальній мережі *Twitter* за ключовими словами, заданими користувачем, і проаналізовано тональності окремих повідомлень і відповідний поділ на фінальні кластери.

Сучасний стан наукових досліджень та технологій в промисловості. 2023. №1 (23)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

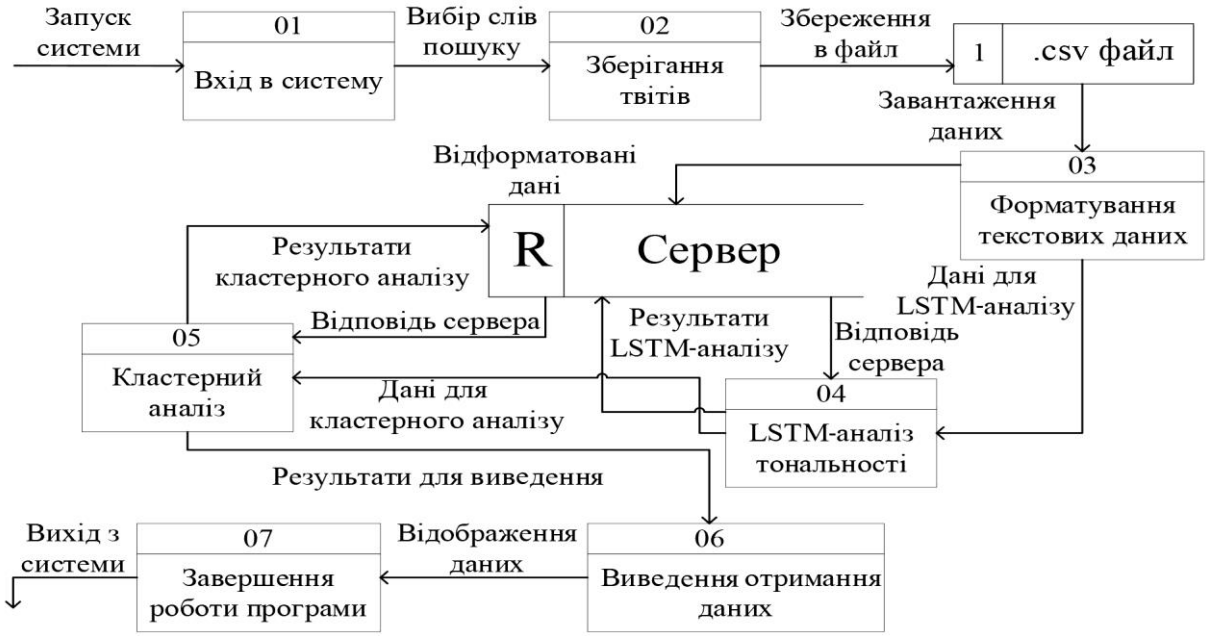

Рис. 7. Діаграма потоків даних інтелектуальної системи аналізу тональності та кластеризації публікацій

Результати досліджень та їхнє обговорення

Проведено експеримент із застосування імплементованої інтелектуальної системи аналізу тональності та кластеризації публікацій в соціальній мережі *Twitter*, перевірено роботу створеної нейронної мережі LSTM і алгоритму кластеризації *k-means*. Для початку потрібно знайти дані для аналізу. Для максимальної актуальності прийнято рішення зробити окремий датасет і завантажити всі користувацькі коментарі та публікації в *Twitter* за останні декілька місяців. Перед тим як завантажувати дані, необхідно було зареєструвати акаунт *Twitter*-розробника та отримати чотири ключі, а саме *consumer key, consumer secret key, access token* та *access token secret*. Усі вони необхідні для використання офіційного API *Twitter* з метою збереження публікацій.

Оскільки в лютому 2022 р. Росія напала на Україну, це стало подією, що викликала відповідну реакцію, як в нашій країні, так і за кордоном. Люди в соціальних мережах, зокрема в *Twitter*, починаючи з лютого і дотепер, активно обговорюють війну та всі теми, пов'язані з нею. Отже, було прийнято рішення проаналізувати тональність публікацій і коментарів українських користувачів щодо сучасних подій. Пошук твітів було здійснено за останні пів року з допомогою геотегу "Ukraine". Пошук почався з липня 2022 р. за публікаціями й коментарями, що містять обов'язкові ключові слова "Ukraine" та "war", а також опціональні слова "missile", "offensive" та "invasion". Ураховуючи обмеження API на завантаження екземплярів публікацій, було вирішено завантажити 2 тис. твітів відповідно до вказаних ключових слів. Для формування датасету необхідні лише конкретні параметри завантажених твітів, тобто потрібно відкинути ті, що не використовуються, за допомогою проходу циклом по всіх завантажених твітах і збереження нікнейму користувача, опису профілю, кількості твітів, підписок і підписників і, що найголовніше, – тексту публікації чи коментаря. Усі дані для початку зберігаються в data-словнику й обробляються до необхідного формату з використанням *DataFrame*-функції для опрацювання даних. З огляду на особливість створеної інтелектуальної системи, визначено, що найбільш зручним форматом для роботи з текстовими даними такого типу за допомогою нейронної мережі та алгоритму кластеризації буде формат *.csv*, оскільки він є універсальним і практично кожна мова програмування чи засіб опрацювання даних має функціонал для роботи з файлами такого типу. Отже, було створено датасет, що опрацьовувався в процесі навчання моделі нейромережі LSTM, і розподілялися дані на кластери. Перед тим виникла необхідність опрацювати всі текстові дані, тобто надати їм одного вигляду для максимально ефективної обробки.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2023. No. 1 (23)

Оскільки ми маємо готовий датасет, можемо здійснювати аналіз тональності, але перед тим необхідно відформатувати й стандартизувати весь текст таким чином, щоб його можна було зручно аналізувати. З тексту було видалено всі хеш-теги, оскільки вони спотворювали зміст публікації чи коментаря. Також було приведено всі слова до нижнього регістру й вилучено всі URL-посилання та всі спеціальні символи, не потрібні для визначення тональності тексту. Під кінець було вилучено всі зайві пропуски та одинарні символи в тексті, а також усі системні символи, додані засобами соціальної мережі *Twitter*. Після цього було здійснено LSTM-аналіз тональності повідомлень, приклад результатів роботи якого зображено на рис. 8.

	text	Sentiment Score	\ Overall Sentiment
0	rt mtracey this woman literally works forus go...	0.000000	Neutral
1	rt smelyansky_igor 6000 branches of ukrposhta ...	0.000000	Neutral
2	snekotron fine with me we all knew they were c...	0.208333	Positive
3	rt euromaidanpress vilnius lithuania protest i...	0.000000	Neutral
4	rt andrew__roth putin is losing the war facing...	0.000000	Neutral

Рис. 8. Приклад результатів визначення тональності тексту засобами LSTM-аналізу

У процесі роботи досить точно визначено значення не тільки загальної тональності слів, а й полярність емоції. Тобто користувач може виявляти певну емоцію з різною силою та інтенсивністю відповідно до ситуації та написаної публікації чи коментаря. Відповідно, ми визначили й тональність текстового екземпляра даних, і емоційний діапазон від –1.0 – максимально негативний текст, до 1.0 – максимально позитивний. Усе що між ними – текст із конкретною тональністю та діапазоном, також є значення 0, що вказує на середньостатистичний нейтральний текст. Аналіз тексту відбувався відповідно до описаного раніше алгоритму запам'ятовування в часі певних моментів з тестової вибірки, на якій навчалася нейронна мережа. Для початку визначаємо мову, якою було написано повідомлення, для зручності обрано лише твіти англійською. Після цього виконуються дві функції *spellcheck* і *correct*, де ми за допомогою базових алгоритмів роботи з текстом перевіряємо правильність написаних слів. Далі відповідно до алгоритму проводимо стематизацію слів, тобто за допомогою методу *definitions* отримуємо список можливих значень до слова згідно з навченою моделлю й обираємо найближче значення. Було згенеровано нейронну мережу, що містить два рівні на 100 елементів, тобто по 50 елементів на кожен рівень. Ми встановлюємо значення просторового спуску, беремо 500 епох для коректного оброблення 2000 повідомлень, 0.5 – значення просторового спуску й, відповідно, аналогічне значення 0.5 задаємо для градієнтного спуску. Було додано значення щільності, що дорівнює 1 та сигмоїдну функцію активації для LSTM-алгоритму й коректного оброблення станів комірок пам'яті. Унаслідок цього наша рекурентна модель складається з вбудованого шару, LSTM-шару й шару щільності, саме в якому сигмоїдна функція відповідає за процес нативної активації. Безпосередньо навчання здійснюється з розміром навчальної партії в 20 елементів і фактором розподілу зі значенням 0.2. Наступним кроком було виділення рівня емоційності тональності тексту для більш широкого аналізу текстової вибірки. Але спочатку було здійснено загальний експеримент на тестовій вибірці з 2000 слів, результати якого наведені на рис. 9.

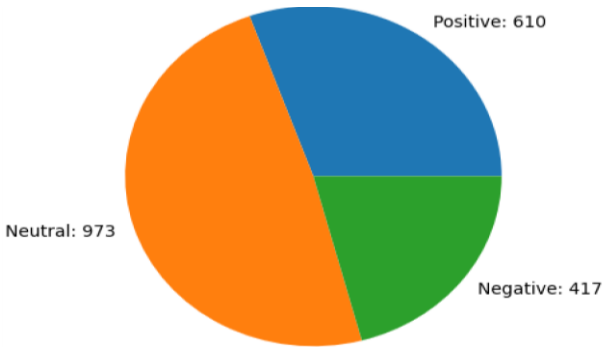

Рис. 9. Загальні результати аналізу тональності тексту

Після отриманого значення початкової тональності здійснюється аналіз, за допомогою якого дізнаємося про ступінь емоційності слів. Сам аналіз певною мірою є складнішим, оскільки

Сучасний стан наукових досліджень та технологій в промисловості. 2023. №1 (23)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

містить одночасно три LSTM-шари. Відповідно, для кожного з них необхідно було запускати свою функцію активації. Для більшої ефективності вся робота виконуватиметься асинхронно, оскільки ми задіяли зовнішні ресурси для виконання коду й очікування результату, що повернеться з функції. Сама модель продовжує виконуватися в межах 500 епох, але замість одного блоку пам'яті маємо три різні, що називають функціональними осередками. Кожен осередок має певний стан, який може змінюватися з огляду на процес навчання й переносити стан моделі. Відповідно, є прихований стан, що є унікальним для кожного осередка й недоступний з інших, так і розподілений шар, який є загальним для всіх трьох осередків і може змінюватися, обираючи оптимальне значення, що здійснюється підбором максимального значення на кожному з кроків навчання суб'єктивної моделі.

Ми здійснимо не тільки загальний аналіз тональності, а й аналіз емоційної суб'єктивної тональності в межах від –1 до 1. У такий спосіб можна зрозуміти, наскільки сильно користувачі соціальної мережі *Twitter* висловили позитивну чи негативну емоцію в межах публікації чи коментаря, тобто реакцію на воєнні дії. На рис. 10 зображено міру тональності проаналізованих текстових екземплярів. Як бачимо, переважну більшість становлять саме публікації та коментарі, що мають нейтральний контекст. Фіолетовим кольором зображено нейтральні текстові екземпляри, оранжевим – позитивні і зеленим – негативні, що добре видно на сформованому графіку. Також спостерігаємо, що оцінка тональності позитивних коментарів коливається в плані значень від 0.1 до 0.4, а оцінка негативного розподілу чисел – від 0.1 до 0.2.

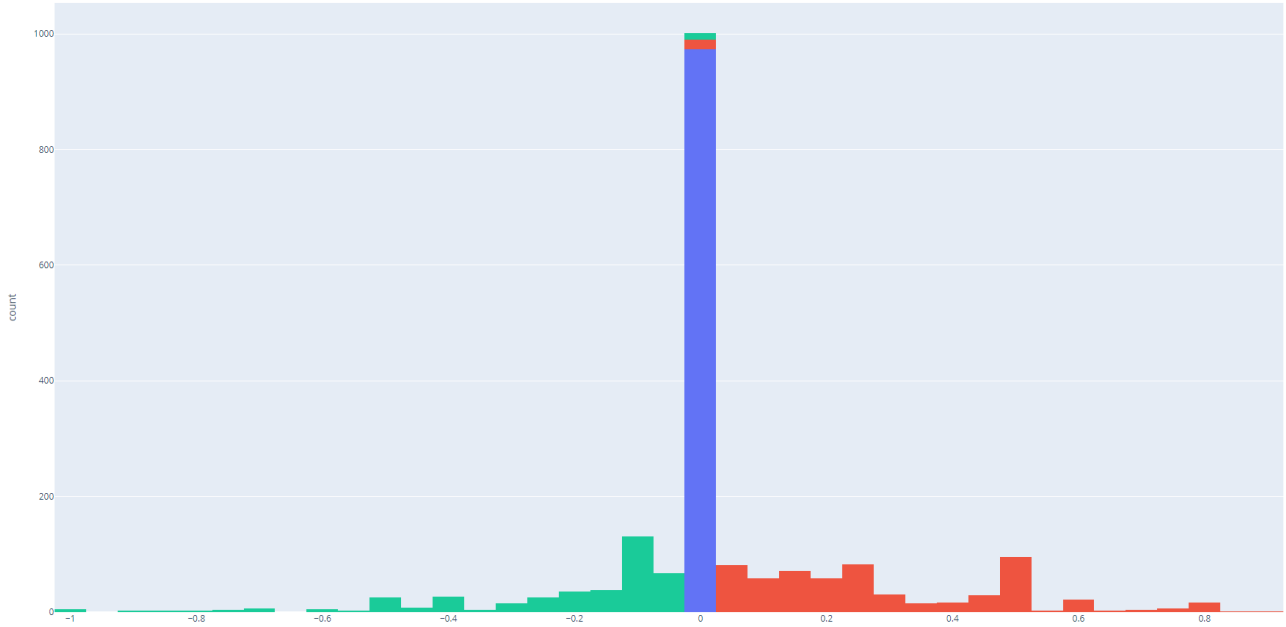

Рис. 10. Міра тональності

Такий нерівномірний розподіл можна пояснити тим, що люди реагують на хороші новини під час війни більш емоційно й імпульсивно. Як видно, на графіку частина коментарів і публікацій сягає 0.8 значення емоційності й, навпаки, негативні коментарі, що пишуть люди після декількох місяців війни, значно менш емоційні. Отже, публікацій, рівень емоційності яких понад 0.4, практично відсутні. Також було сформовано графік суб'єктивності тональності (рис. 11). Завдяки цьому графіку зручно визначити, де саме є найбільша щільність розподілу суб'єктивності публікацій і коментарів з обраного датасету. Як бачимо, найбільша щільність позитивних і негативних емоцій у текстових екземплярах має однакове значення 0.5.

Отже, проаналізувавши тональність користувацьких публікацій і коментарів, чітко побачили, скільки за визначений період часу і з використанням тематичних ключових слів написано позитивних, негативних і нейтральних твітів і який між ними розподіл. Окрім того, у дослідженні здійснено глибший аналіз і визначено розподіл значень тональності тексту та її суб'єктивність.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2023. No. 1 (23)

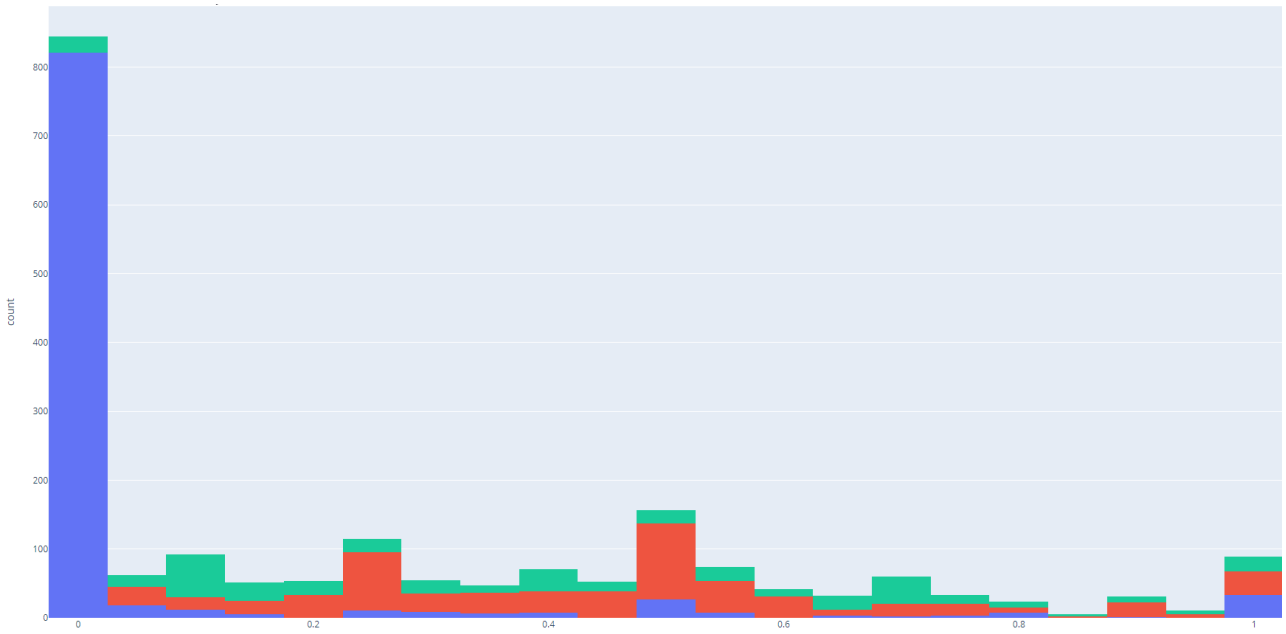

Рис. 11. Суб'єктивність тональності

Далі було здійснено кластеризацію методом *k-means* для того, щоб чітко бачити розподіл текстових повідомлень різних тональностей по кластерах. Сам процес кластеризації має три основних кроки: ініціалізація моделі, налаштування *fit*-параметрів моделі та прогноз для подальшого оброблення й поділу на кластери. Для початку було задано *n* – кількість кластерів, що може становити від 3 до 5, відповідно до кількості текстових елементів, та *random state* – стан процесу кластеризації, що є випадковим числом, необхідним для ініціалізації моделі даних. Для коректної роботи інтелектуальної системи було задано параметри кластеризації, такі як *model.fit* – налаштування моделі, *max iteration* – кількість змін положення центру кластерів, у нашому випадку це 300 змін, *n_init value* – значення, що позначає, скільки разів відбуватиметься кластеризація, і залежить від вибору центрів кластерів, *tol* або *tolerance* – перевірка щодо того, чи центри кластерів рухаються в процесі навчання моделі. Якщо рух менший за одну тисячну, то кластеризація завершується та параметр *verbose*, що позначає проміжну інформацію про процес кластеризації та можливі помилки. Також було задано такі параметри, як *model.labels*, що позначає номер кластеру, якому належить об'єкт, *model.cluster_centers*, що позначає координати центрів кластерів, і *model.predict*, який, відповідно, позначає додавання нового об'єкта до кластеру з найбільшим центром, щоб не робити заново кластеризацію, а використовувати готову модель. Для визначення кількості кластерів було використано "ліктьовий" метод, що зображено на рис. 12, відповідно до результатів на найбільшому спаді було обрано три кластери.

Після визначення кількості кластерів настав час розподілити текстові екземпляри по цих кластерах. Оскільки їх всього три, то зручно розподілити на позитивні, нейтральні й негативні. Щоб це зробити, була використана функція *groupby.mean*, сутність якої полягає у визначенні середнього значення за кожним кластером. Далі здійснюється стандартизація, для якої потрібно від середнього значення вибірки відняти значення кластеру й поділити на середнє відхилення, після чого видаляється середнє значення та масштабується дисперсія одиниць. Розподіл текстових даних користувачів по кластерах зображено на рис. 13 та 14 за допомогою різних типів графіків. Отже, PCA-графік відтворює загальну структуру даних, а TSNE-графік найкраще демонструє відносини між сусідами. Для більш зручного відображення даних, замість усіх можливих об'єктів, на PCA-графік було виведено 40, а на TSNE – 400 текстових екземплярів. Також для зручного об'єднання в кластери текстові одиниці даних на рис. 13 і 14 розміщено в декартовій системі координат, оскільки в ній є змога однозначно визначити кожну точку на площині з використанням пари числових координат та осей абсцис і ординат. Маючи всі точки, у процесі роботи система визначає відстані між ними й розділяє на кластери за схожістю.

Сучасний стан наукових досліджень та технологій в промисловості. 2023. №1 (23)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

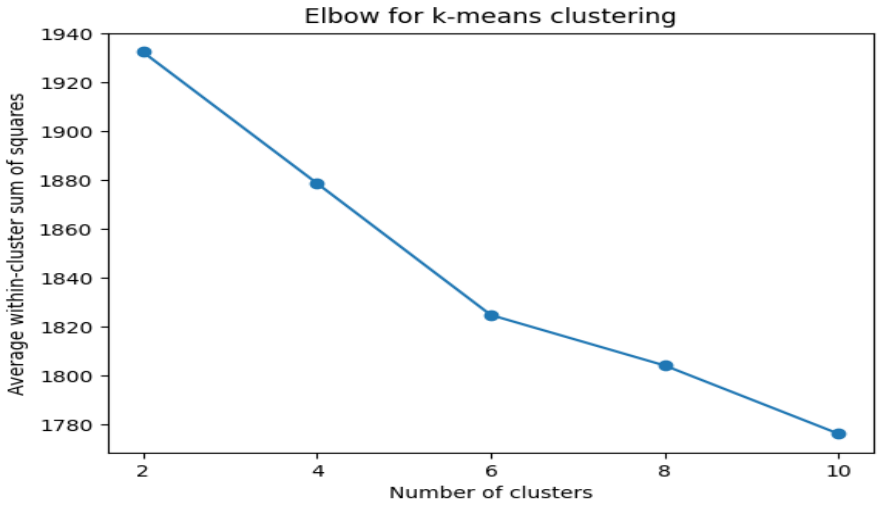

Рис. 12. Результат "ліктьового" методу

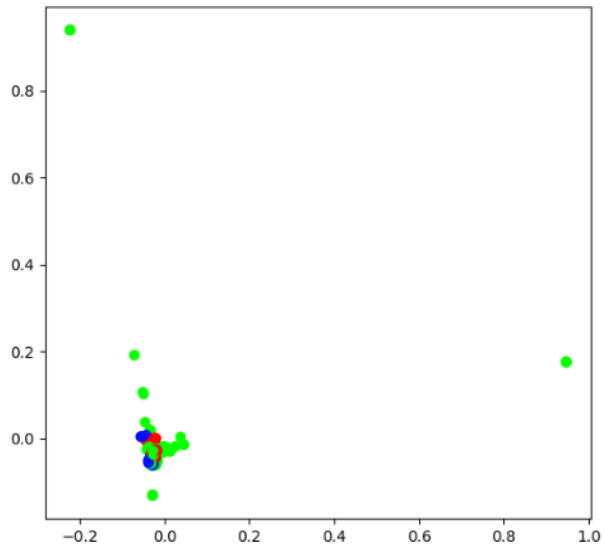

Рис. 13. Графік розподілу кластерів типу PCA

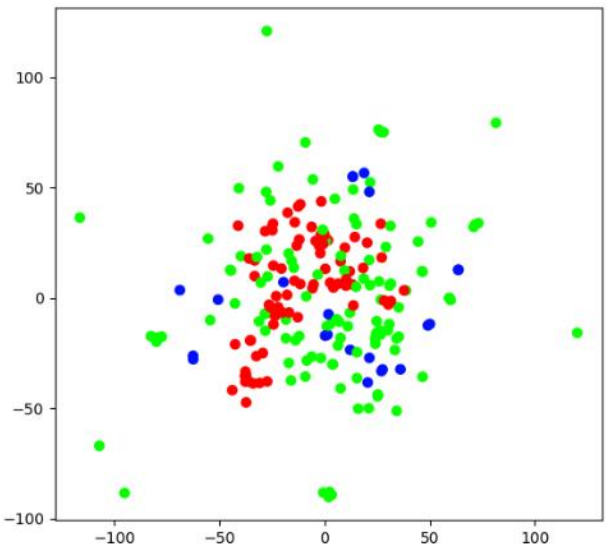

Рис. 14. Графік розподілу кластерів типу TSNE

На останок було виведено найбільш популярні ключові слова до кожного кластеру, з яких можна зробити певні висновки або помітити тенденції. З цією метою було обчислено середнє значення для всіх вимірів створеної моделі, які були розподілені в кожному кластері. Наступним кроком було сортування масивів середніх значень кожного кластеру за спаданням і обрані перші десять елементів, що зображені на рис. 15, де нульовий кластер містить негативні користувацькі публікації та коментарі, перший кластер – позитивні й другий – нейтральні. Також для кращої візуалізації створено об'ємний 3D-графік, що зображений на рис. 16, де кожна точка показує певну текстову одиницю даних і, відповідно, визначаються координати точок *x*, *y* та *z*, що згодом будуть об'єднані в один із трьох кластерів. Інформація, яка раніше була зображена в декартовому просторі, зараз зображена в тривимірному, з чого можна зробити висновок про правильність створених кластерів. На основі наведеного графіка можна підсумувати щодо належності всіх обраних текстових екземплярів до певного кластеру в тривимірному просторі.

Провівши експеримент, можна дійти висновку, що використана комбінація алгоритмів, а саме нейронна мережа для аналізу тональності LSTM та алгоритм кластеризації *k-means*, працює ефективно й дає змогу точніше аналізувати датасети, ніж аналоги, особливо коли датасети сягають значних розмірів і необхідно точно проаналізувати тональність тексту разом із рівнем емоційності й ефективно розподілити на необхідну для подальших досліджень кількість кластерів. На рис. 17, де

ISSN 2522-9818 (print)
ISSN 2524-2296 (online)
Innovative technologies and scientific solutions for industries. 2023. No. 1 (23)

на осі абсцис показано кількість поданих текстових одиниць, а на осі ординат – час у мілісекундах, відображено порівняння роботи комбінацій нейронної мережі CNN та алгоритму ієрархічної кластеризації, які мають складність алгоритму $O(n2)$, та створеної в інтелектуальній системі нейронної мережі LSTM і алгоритму кластеризації *k-means*, що мають складність алгоритму $O(nlogn)$. Відповідно, імплементовані алгоритми в інтелектуальній системі працюють як мінімум на 10…15% швидше та ефективніше, ніж описані раніше згорткова нейронна мережа та ієрархічна кластеризація, а також мають великі обсяги даних і ефективність роботи становить до 20%.

```
Cluster 0
one,country,nato,stop,people,putin,rt,russia,war,ukraine

Cluster 1
apmassaro3,attack,supporter,russia,civilian,today,russian,military,ukraine,rt

Cluster 2
monday,ukraine,oct,10,defense,pic,russian,air,rt,missile
```

Рис. 15. Найпопулярніші ключові слова

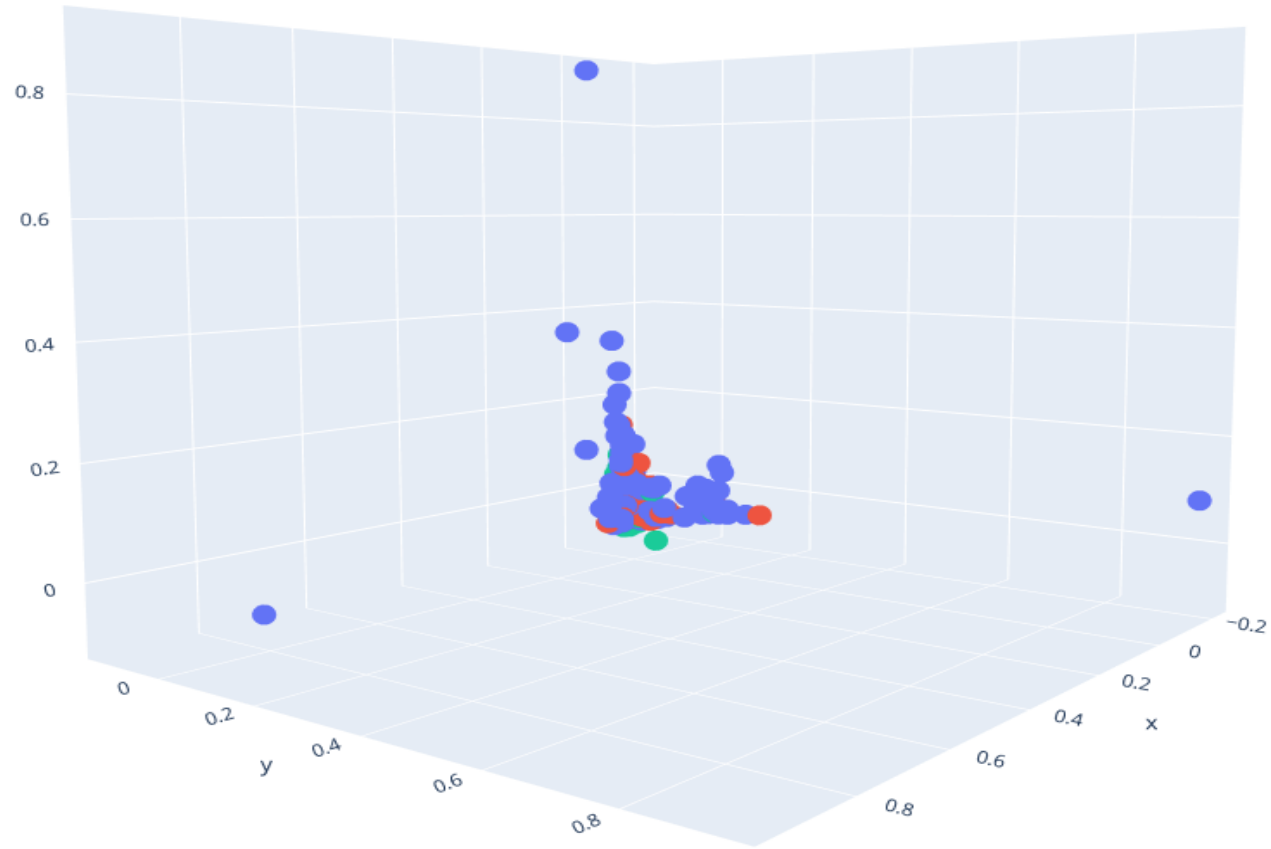

Рис. 16. Об’ємний графік розподілу кластерів

Сучасний стан наукових досліджень та технологій в промисловості. 2023. №1 (23)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

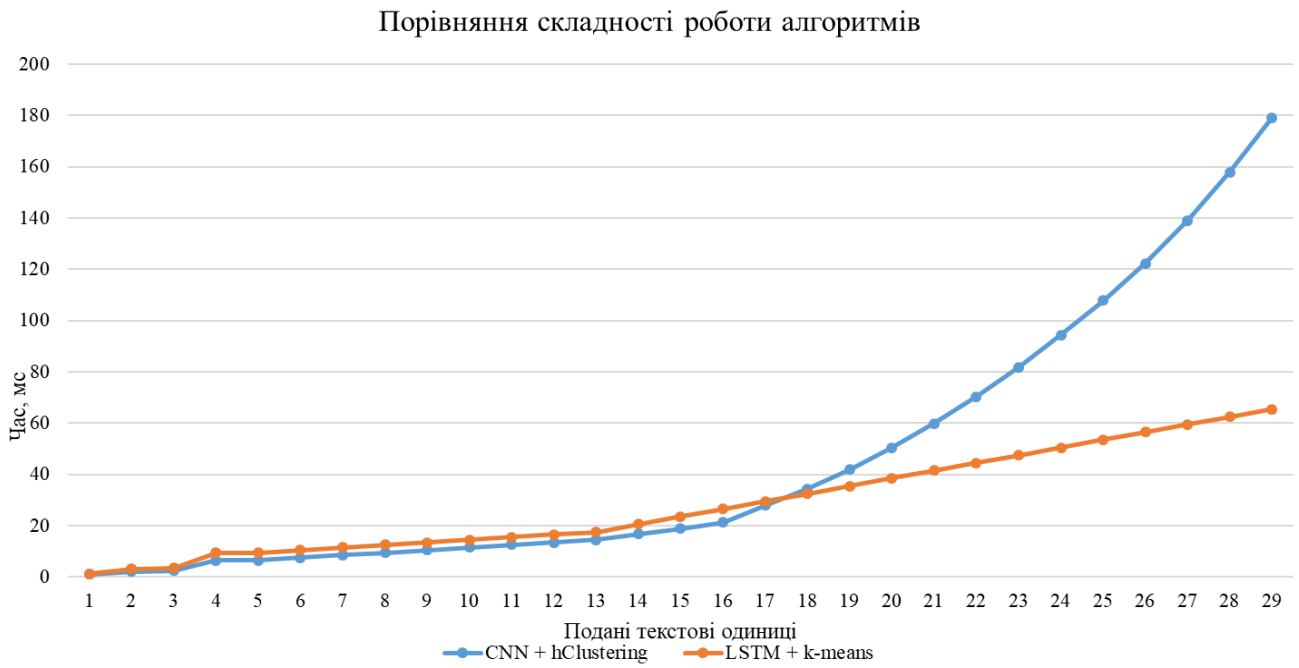

Рис. 17. Порівняння складності роботи алгоритмів

Висновки та перспективи подальшого розвитку

У процесі роботи імплементовано інтелектуальну систему аналізу тональності та кластеризації публікацій у соціальній мережі *Twitter*, за допомогою якої користувач може вводити конкретні параметри й ключові слова, завдяки яким здійснюватиметься завантаження публікацій і коментарів із *Twitter* для аналізу тональності повідомлень, їхньої емоційної оцінки та поділу на кластери. Описано важливість створення такої інтелектуальної системи й проаналізовано останні дослідження для розуміння, що на сьогодні вже реалізовано, а що ще потрібно зробити й, відповідно, повторити переваги й уникнути недоліків. Розглянуто застосування нейронних мереж різного спрямування та алгоритмів кластеризації. Визначено мету роботи та обґрунтовано, чому реалізація сучасної нейронної мережі LSTM та алгоритму кластеризації *k-means* є актуальним завданням для підвищення ефективності роботи системи й покращення точності фінальних результатів. Також описано функціонал системи, а саме сутність роботи її основних алгоритмів і загальну реалізацію ключових функціональних компонентів, і подано за допомогою діаграм, на яких зображено і блок-схеми роботи алгоритмів, і варіанти використання, і взаємодію всіх створених потоків даних інтелектуальної системи загалом.

Здійснено реальний експеримент із використанням створеної інтелектуальної системи, під час якого було задано ключові слова й параметри пошуку публікацій і коментарів у соціальній мережі *Twitter* та завантажено сукупність твітів, які збережені та сформовані в датасет. Після цього виконано форматизацію, лематизацію та стандартизацію текстових екземплярів для коректної роботи алгоритмів. Щодо певного тексту навчено й використано нейронну мережу *Long Short-Term Memory*, за допомогою якої здійснено як і загальний аналіз тональності з розділом тексту на негативний, нейтральний і позитивний, так і аналіз емоційного значення тексту, а саме розподіл від –1.0, тобто максимально можливого негативного, до 1.0 – максимально можливого позитивного, відображено значення і суб'єктивність тональності. Аналогічно здійснено кластерний аналіз тексту з використанням алгоритму кластеризації *k-means*, створено й навчено модель і за допомогою "ліктьового" методу обрано оптимальну кількість кластерів для завантаженого датасету. Після розподілу текстових даних на кластери було графічно зображено сформовані кластери за допомогою PCA- та TSNE-графіків. Обчислено середнє значення для поточних вимірів створеної моделі й відображено по 10 найбільш популярних слів у кожному кластері, виведено зображення текстових елементів кожного кластеру за допомогою тривимірного графіка. Показано

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) *Innovative technologies and scientific solutions for industries. 2023. No. 1 (23)*

різницю в ефективності та складності між, з одного боку, нейронною мережею CNN та ієрархічною кластеризацією і, з іншого боку, нейронною мережею LSTM і алгоритмом кластеризації *k-means*, що були реалізовані в інтелектуальній системі. Здобуті результати можна порівняти із статтею [15], де автори аналізували коментарі користувачів у соціальній мережі *Twitter* щодо вірусу COVID-19. Автори використовували комбінацію нейронної мережі CNN та алгоритму ієрархічної кластеризації, унаслідок цього аналіз тональності через значну кількість даних є неточним і містить прогалини на графіках, також кластеризація здійснювалася лише з окремими виділеними наборами даних, а не з усім датасетом, щоб провести коректну кластеризацію за не дуже тривалий проміжок часу. Створені в інтелектуальній системі нейронна мережа LSTM та алгоритм кластеризації *k-means* вирішують ці проблеми, оскільки вони можуть працювати як і з середніми, так і з великими за обсягом датасетами без втрат у точності отриманих результатів. Також створені алгоритми працюють на 10…15% швидше й, відповідно, дають змогу аналізувати всю наявну вибірку даних за один раз.

Список літератури

1. Almahmood R. J. K., Tekerek A. Issues and Solutions in Deep Learning-Enabled Recommendation Systems within the E-Commerce Field. *Applied Sciences*. 2022. № 12 (21). P. 256–264. DOI: https://doi.org/10.3390/app122111256
2. Xie W., Damiano L., Jong C.-H. Emotional appeals and social support in organizational YouTube videos during COVID-19. *Telematics and Informatics reports*. 2022. № 8 (1). P. 100–128.
3. Abbas A. F., Jusoh A., Mas'od A., Alsharif A. H., Ali J. Bibliometrix analysis of information sharing in social media. *Cogent Business & Management*. 2022. № 9 (1). P. 521–543. DOI: https://doi.org/10.1080/23311975.2021.2016556
4. Villegas-Ch. W., Erazo D. M., Ortiz-Garces I., Gaibor-Naranjo W., Palacios-Pacheco X. Artificial Intelligence Model for the Identification of the Personality of Twitter Users through the Analysis of Their Behavior in the Social Network. *Electronics*. 2022. № 11 (22). P. 381–399. DOI: https://doi.org/10.3390/electronics11223811
5. Malkawi R., Daradkeh M., El-Hassan A., Petrov P. A Semantic Similarity-Based Identification Method for Implicit Citation Functions and Sentiments Information. *Information.* 2022. № 13 (11). P. 546–561. DOI: https://doi.org/10.3390/info13110546
6. Yuan Y., You T., Xu T., Yu X. Customer-Oriented Strategic Planning for Hotel Competitiveness Improvement Based on Online Reviews. *Sustainability*. 2022. № 14 (22). P. 152–199.
7. Yin J. Y. B., Saad N. H. M., Yaacob Z. Exploring Sentiment Analysis on E-Commerce Business: Lazada and Shopee. *Tem journal.* 2022. № 11 (4). P. 1508–1519. DOI: https://doi.org/10.18421/TEM114-11
8. Hinduja S., Afrin M., Mistry S., Krishna A. Machine learning-based proactive social-sensor service for mental health monitoring using twitter data. *International journal of Information Management Data insights*. 2022. № 2 (2). P. 103–124.
9. Bhadamkar A., Bhattacharya S. Tesla Inc. Stock Prediction Using Sentiment Analysis. *Australasian Accounting, Business and Finance journal.* 2022. № 16 (5). P. 52–66. DOI: https://doi.org/10.14453/aabfj.v16i5.05
10. Alhakiem H. R., Setiawan E. B. Aspect-Bas1ed Sentiment Analysis on Twitter Using Logistic Regression with FastText Feature Expansion. *Jurnal resti (Rekayasa sistem dan Teknologi Informasi)*. 2022. № 6 (5). P. 840–846. DOI: https://doi.org/10.29207/resti.v6i5.4429
11. Pawełoszek I. Towards a Smart City—The Study of Car-Sharing Services in Poland. *Energies.* 2022. № 15 (22). P 845–859. DOI: https://doi.org/10.3390/en15228459
12. Huang X., Gong P., Wang S., White M., Zhang B. Machine Learning Modeling of Vitality Characteristics in Historical Preservation Zones with Multi-Source Data. *Buildings.* 2022. № 12 (11). P. 1978–1989. DOI: https://doi.org/10.3390/buildings12111978
13. Li C., Renda M., Yusuf F., Geller J., Chun S. A. Public Health Policy Monitoring through Public Perceptions: A Case of COVID-19 Tweet Analysis. *Information*. 2022. № 13 (11). P. 443–457. DOI: https://doi.org/10.3390/info13110543
14. Vysotska V. Information Technology for Internet Resources Promotion in Search Systems Based on Content Analysis of Web-Page Keywords. *Radio Electronics, Computer Science, Control*. 2021. № 3. P. 133–151.
15. Corti L., Zanetti M., Tricella G., Bonati M. Social media analysis of Twitter tweets related to ASD in 2019–2020, with particular attention to COVID-19: topic modelling and sentiment analysis. *Journal of Big Data*. 2022. № 9 (1). P. 1–17. DOI: https://doi.org/10.1186/s40537-022-00666-4
16. Lampropoulos G., Keramopoulos E. Virtual Reality in Education: A Comparative Social Media Data and Sentiment Analysis Study. *International journal of recent contributions from Engineering, Science & IT*. 2007. № 10 (3). P. 221–235. DOI: https://doi.org/10.3991/ijes.v10i03.34057
17. Liu H. Online review analysis on various networks' consumer feedback using deep learning. *IET networks*. 2022. № 11 (6). P. 234–244. DOI: https://doi.org/10.1049/ntw2.12045
18. Wang Y., Chen Z., Fu C. Synergy Masks of Domain Attribute Model DaBERT: Emotional Tracking on Time-Varying Virtual Space Communication. *Sensors*. 2022. № 22 (21). P. 450–471. DOI: https://doi.org/10.3390/s22218450

Сучасний стан наукових досліджень та технологій в промисловості. 2023. №1 (23)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

References

1. Almahmood R. J. K., Tekerek A. Issues and Solutions in Deep Learning-Enabled Recommendation Systems within the E-Commerce Field. *Applied Sciences*. 2022. № 12 (21). P. 256–264. DOI: https://doi.org/10.3390/app122111256
2. Xie W., Damiano L., Jong C.-H. Emotional appeals and social support in organizational YouTube videos during COVID-19. *Telematics and Informatics reports*. 2022. № 8 (1). P. 100–128.
3. Abbas A. F., Jusoh A., Mas'od A., Alsharif A. H., Ali J. Bibliometrix analysis of information sharing in social media. *Cogent Business & Management*. 2022. № 9 (1). P. 521–543. DOI: https://doi.org/10.1080/23311975.2021.2016556
4. Villegas-Ch. W., Erazo D. M., Ortiz-Garces I., Gaibor-Naranjo W., Palacios-Pacheco X. Artificial Intelligence Model for the Identification of the Personality of Twitter Users through the Analysis of Their Behavior in the Social Network. *Electronics*. 2022. № 11 (22). P. 381–399. DOI: https://doi.org/10.3390/electronics11223811
5. Malkawi R., Daradkeh M., El-Hassan A., Petrov P. A Semantic Similarity-Based Identification Method for Implicit Citation Functions and Sentiments Information. *Information.* 2022. № 13 (11). P. 546–561. DOI: https://doi.org/10.3390/info13110546
6. Yuan Y., You T., Xu T., Yu X. Customer-Oriented Strategic Planning for Hotel Competitiveness Improvement Based on Online Reviews. *Sustainability*. 2022. № 14 (22). P. 152–199.
7. Yin J. Y. B., Saad N. H. M., Yaacob Z. Exploring Sentiment Analysis on E-Commerce Business: Lazada and Shopee. *Tem journal.* 2022. № 11 (4). P. 1508–1519. DOI: https://doi.org/10.18421/TEM114-11
8. Hinduja S., Afrin M., Mistry S., Krishna A. Machine learning-based proactive social-sensor service for mental health monitoring using twitter data. *International journal of Information Management Data insights*. 2022. № 2 (2). P. 103–124.
9. Bhadamkar A., Bhattacharya S. Tesla Inc. Stock Prediction Using Sentiment Analysis. *Australasian Accounting, Business and Finance journal.* 2022. № 16 (5). P. 52–66. DOI: https://doi.org/10.14453/aabfj.v16i5.05
10. Alhakiem H. R., Setiawan E. B. Aspect-Bas1ed Sentiment Analysis on Twitter Using Logistic Regression with FastText Feature Expansion. *Jurnal resti (Rekayasa sistem dan Teknologi Informasi)*. 2022. № 6 (5). P. 840–846. DOI: https://doi.org/10.29207/resti.v6i5.4429
11. Pawełoszek I. Towards a Smart City—The Study of Car-Sharing Services in Poland. *Energies.* 2022. № 15 (22). P 845–859. DOI: https://doi.org/10.3390/en15228459
12. Huang X., Gong P., Wang S., White M., Zhang B. Machine Learning Modeling of Vitality Characteristics in Historical Preservation Zones with Multi-Source Data. *Buildings*. 2022. № 12 (11). P. 1978–1989. DOI: https://doi.org/10.3390/buildings12111978
13. Li C., Renda M., Yusuf F., Geller J., Chun S. A. Public Health Policy Monitoring through Public Perceptions: A Case of COVID-19 Tweet Analysis. *Information*. 2022. № 13 (11). P. 443–457. DOI: https://doi.org/10.3390/info13110543
14. Vysotska V. Information Technology for Internet Resources Promotion in Search Systems Based on Content Analysis of Web-Page Keywords. *Radio Electronics, Computer Science, Control*. 2021. № 3. P. 133–151.
15. Corti L., Zanetti M., Tricella G., Bonati M. Social media analysis of Twitter tweets related to ASD in 2019–2020, with particular attention to COVID-19: topic modelling and sentiment analysis. *Journal of Big Data*. 2022. № 9 (1). P. 1–17. DOI: https://doi.org/10.1186/s40537-022-00666-4
16. Lampropoulos G., Keramopoulos E. Virtual Reality in Education: A Comparative Social Media Data and Sentiment Analysis Study. *International journal of recent contributions from Engineering, Science & IT*. 2007. № 10 (3). P. 221–235. DOI: https://doi.org/10.3991/ijes.v10i03.34057
17. Liu H. Online review analysis on various networks' consumer feedback using deep learning. *IET networks*. 2022. № 11 (6). P. 234–244. DOI: https://doi.org/10.1049/ntw2.12045
18. Wang Y., Chen Z., Fu C. Synergy Masks of Domain Attribute Model DaBERT: Emotional Tracking on Time-Varying Virtual Space Communication. *Sensors*. 2022. № 22 (21). P. 450–471. DOI: https://doi.org/10.3390/s22218450

Received 16.01.2023

Відомості про авторів / About the Authors

Батюк Тарас Миронович – Національний університет "Львівська політехніка", аспірант кафедри "Інформаційні системи та мережі", вул. Степана Бандери, 12, Львів, Україна; e-mail: taras.m.batiuk@lpnu.ua; ORCID ID: https://orcid.org/0000-0001-5797-594X

Досин Дмитро Григорович – доктор технічних наук, старший науковий співробітник, Національний університет "Львівська політехніка", професор кафедри "Інформаційні системи та мережі", вул. Степана Бандери, 12, Львів, Україна; e-mail: dmytro.h.dosyn@lpnu.ua; ORCID ID: https://orcid.org/0000-0003-4040-4467

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) *Innovative technologies and scientific solutions for industries. 2023. No. 1 (23)*

Batiuk Taras – Lviv Polytechnic National University, Postgraduate Student of the Information Systems and Networks Department, Lviv, Ukraine.

Dosyn Dmytro – Doctor of Sciences (Engineering), Lviv Polytechnic National University, Professor of Information Systems and Networks Department, Lviv, Ukraine.

IMPLEMENTATION OF THE INTELLECTUAL SYSTEM OF SENTIMENT ANALYSIS AND CLUSTERIZATION OF PUBLICATIONS IN THE TWITTER SOCIAL NETWORK

Thanks to the intensive development of social networks, the intensity of exchange of short electronic text messages is constantly increasing, the tone of which can serve as a sensitive indicator of public mood and important social phenomena, interesting for sociologists, politicians, economists, and specialists in other fields. In this regard, the task of automating the processing of such natural language messages is of significant scientific and practical interest. The **object** of this study is the sentiment of user publications in the Twitter social network. Due to the great popularity of the social network itself and the large number of user messages, which are short in nature, it is possible to conveniently determine the mood of user posts and combine them into clusters according to the given parameters of the intelligent system. The **subject** of the study is methods and algorithms for analysing the sentiment of large arrays of messages containing the necessary keywords and relating to a certain specific topic, determining the factors and distributions of the sentiment of messages based on the input array of system data, dividing messages into main groups and providing estimates within certain defined limits in to each group, division into clusters according to the obtained search point and display of the obtained results in the desired format. The **purpose** of the work is to implement an intelligent system of sentiment analysis and clustering of publications based on a recurrent neural network of long short-term memory (LSTM) and the k-means clustering algorithm. The following main **tasks** are solved in the work: 1. To analyse the most used and newest algorithms, methods, approaches and means of implementing tasks of sentiment analysis and clustering of publications in social networks. 2. To develop a conceptual structure of an intellectual system of sentiment analysis and clustering of publications. 3. To form functional tasks for the key modules of the created intelligent system of sentiment analysis and clustering of publications in the Twitter social network. 4. Implement an intelligent system of sentiment analysis and clustering of publications based on a recurrent neural network and the *k-means* clustering algorithm and conduct experimental verification. Among the **methods** used for this purpose are the recurrent neural network of long short-term memory; k-means clustering algorithm. The following **results** were obtained: the general structure of the intellectual system of sentiment analysis and clustering of publications was analyzed, designed and built. The main task of creating the system, first of all, was to improve the recurrent neural network of long-short-term memory, which, thanks to the improved algorithm, significantly facilitates text processing by natural language processors according to text data of a certain size. Also, a special clustering algorithm, namely k-means, was used in parallel, thanks to which it was possible to change the general approach to clustering and the creation of final clusters, in accordance with the obtained results of the work of the recurrent neural network. **Conclusions:** As a result of applying a combination of LSTM neural network and k-means clustering algorithm, it was possible to speed up the process of sentiment analysis and clustering of posts in the Twitter social network by 10...15% compared to similar convolutional neural networks and hierarchical clustering.

Keywords: neural network; LSTM; sentiment analysis of publications; cluster analysis; social network Twitter.

Бібліографічні описи / Bibliographic descriptions

Батюк Т. М., Досин Д. Г. Імплементація інтелектуальної системи аналізу тональності та кластеризації публікацій у соціальній мережі Twitter. *Сучасний стан наукових досліджень та технологій в промисловості*. 2023. № 1 (23). С. 25–44. DOI: https://doi.org/10.30837/ITSSI.2023.23.025

Batiuk, T., Dosyn, D. (2023), "Implementation of the intellectual system of sentiment analysis and clusterization of publications in the Twitter social network", *Innovative Technologies and Scientific Solutions for Industries*, No. 1 (23), P. 25–44. DOI: https://doi.org/10.30837/ITSSI.2023.23.025