УДК 004.9

ТЕХНОЛОГІЯ СЕНТИМЕНТ-АНАЛІЗУ ВІДГУКІВ КОРИСТУАЧІВ СИСТЕМ Е-КОМЕРЦІЇ НА ОСНОВІ МАШИННОГО НАВЧАННЯ

Тчинецький С. А. – магістр кафедри «Інформаційні системи та мережі», Національний університет «Львівська політехніка», Львів, Україна.

Поліщук Б. О. – магістр кафедри «Інформаційні системи та мережі», Національний університет «Львівська політехніка», Львів, Україна.

Висоцька В. А. – канд. техн. наук, доцент, доцент кафедри «Інформаційні системи та мережі», Національний університет «Львівська політехніка», Львів, Україна.

АНОТАЦІЯ

Актуальність. Взаємодія між компанією та цільовою аудиторією досліджується вже століттями. З самого початку комерційних відносин, стосунки надавача послуг та отримувача цінувалися чи не понад усе. Торгівля побудована на довірі та повазі. Імідж підприємця часто є важливішим ніж товар, який він продає. За багато сотень років, взаємини торговця і покупця, підприємця та клієнта не втратили важливості і в час масової диджиталізації якість відносин компанії та цільової аудиторії різного розміру та професійна підтримка зворотного зв'язку з клієнтами часто визначають успіх е-бізнесу. Для цього необхідні додаткові інструменти та інформаційні технології для допомоги бізнесменам слідкувати за можливостями розвитку е-бізнесу в певній локації, а також встановлювати зворотній зв'язок з користувачами за допомогою соціальних мереж та ЗМІ. Такі інструменти допоможуть суттєво розширити бачення ринкових можливостей для е-бізнесу, з'ясує – в які з них є сенс інвестувати, а на які не варто витрачати час. Також побачити, яка ідея має майбутнє і яку бізнес-модель потрібну реалізувати/підтримувати/розвивати для стрімкого розвитку територіального/ міжрегіонального е-бізнесу. Також допоможе розібратися, які важелі мають найбільший ефект для зміни політики бізнесу: що не чіпати, а що змінити, щоб забезпечити високу швидкість в реалізації задуму на основі аналізу відповідних результатів досліджень, наприклад, отримувати: прямий фідбек від клієнтів, динаміку зміни загальної задоволеності або зацікавленості цільової аудиторії та переваги/недоліки від користувачів за допомогою NLP-аналізу; підтримку розвитку е-бізнесу відносно локацій знаходження їхнього підприємства та найкращі напрями розвитку; – графіки розвитку бізнесу (покращення/погіршення) залежно від змісту коментарів.

Метою дослідження є розробка інформаційної технології підтримки розвитку е-бізнесу за допомогою аналізу локацій знаходження бізнесу, опрацювання фідбеку від користувачів, аналізу та класифікації відгуків клієнтів в режимі реального часу з соціальних мереж: Twitter, Reddit, Facebook та інші за допомогою методів глибокого навчання та Natural Language Processing українсько- та англомовних текстів.

Метод. Для аналізу відгуків користувачів та клієнтів використано NLP-методи. Серед методів реалізації основних функцій класифікації англомовних новин використані такі методи машинного навчання, як: наївний Баєсів класифікатор, логістична регресія та метод опорних векторів. Для класифікації україномовних відгуків від користувачів використано алгоритм Наївного Байєса, оскільки він добре показує себе на малих обсягах даних, простий у тренуванні та експлуатації та добре працює з текстовими даними. Наївний класифікатор Байєс є дуже хорошим варіантом для нашої системи і з розрахунку того, що кількість відгуків у датасеті є меншою порівняно з середніми показниками.

Результати. Розроблено модель машинного навчання для аналізу та класифікації україномовних та англомовних відгуків від користувачів систем е-комерції.

Висновки. Створена модель показує відмінні результати класифікації на тестових даних. Загальна точність сентиментальної моделі для аналізу україномовного контенту є доволі задовільною, 92.3%. Найкраще з завданням аналізу впливу англомовних новини на фінансовий ринок впорався метод логістичної регресії, який показав точність 75,67%. Безперечно, це не є бажаним результатом, проте це найбільший показник із усіх розглянутих. Дещо гірше зі завданням впорався метод опорних векторів (SVM), який показав точність 72,78%, що є дещо гіршим результатом за той, який було отримано завдяки методу логістичної регресії. І найгірше зі завданням впорався метод наївного баєсового класифікатора, який отримав точність 71,13%, що є меншою за отриману у двох попередніх методах.

КЛЮЧОВІ СЛОВА: NLP, text pre-processing, сентимент-аналіз, відгук, коментар, е-комерція, е-бізнес, машинне навчання, контент аналіз.

АБРЕВІАТУРА

БД – база даних;
ЗМІ – засоби масової інформації
ІС – інтелектуальна система;
ІТ – інформаційна технологія;
ШІ – штучний інтелект;
ПО – предметна область;
ML – machine learning;
NLP – natural language processing.

НОМЕНКЛАТУРА

S – система аналізу та класифікації відгуків;
I – множина вхідних даних;
O – множина вихідних даних;
R – основні правила опрацювання вхідних даних;
U – параметри опрацювання вхідних даних;
N – машинне навчання;
α – оператор скачування вхідних даних;
β – оператор опрацювання вхідних даних;
γ – оператор збереження вхідних даних;
μ –оператор видалення шуму в даних;
χ – оператор пошуку ключових слів;
ω – оператор машинного навчання ІС на достовірних текстових даних;

DOI 10.15588/1607-3274-2023-3-11

λ – оператор класифікації відгуків;
i_1 – множина даних ідентифікації;
i_2 – множина вхідного текстового контенту;
i_3 – множина шаблонів/правил NLP;
i_4 – множина відфільтрованих відгуків;
o_1 – маркований/тегований текст відгуків;
o_2 – колекція пропозицій аналізу відгуків;
o_3 – множина класифікованих відгуків;
r_1 – правила алгоритму взаємодії;
r_2 – NLP-правила;
r_3 – правила алгоритму машинного навчання;
r_4 – правила алгоритму класифікації відгуків;
r_{5j} – правила алгоритму *j*-тої NLP-задачі;
r_6 – правила алгоритму подання результатів;
u_1 – множина рівнів доступу;
u_2 – множина вимог доступу;
u_3 – множина NLP-вимог;
u_4 – множина метрик машинного навчання;
u_5 – множина вимог класифікації відгуків;
u_{6j} – множина вимог розв'язку *j*-тої NLP-задачі;
α_1 – оператор збору текстового тематичного контенту з Google за певний період часу;
α_2 – оператор збору текстового тематичного контенту з Twitter за певний період часу;
α_3 – оператор збору текстового тематичного контенту з Facebook за певний період часу;
α_4 – оператор збору текстового тематичного контенту з Reddit за певний період часу;
α_5 – оператор завантаження власних даних;
ϕ_1 – оператор пошуку за ключовими словами;
ϕ_2 – оператор сентимент аналізу відгуків;
ϕ_3 – оператор розрахунку рівня популярності запитів;
ϕ_4 – оператор узагальнення тексту;
ϕ_5 – оператор пошуку оптимальних локацій.

ВСТУП

Бізнес відіграє ключову роль в економіці кожної країни. Так в Україні малий та середній бізнес забезпечує близько 64% доданої вартості, 81,5% зайнятих працівників у суб'єктів господарювання та 37% податкових надходжень в 2021 році [1]. Із-за війни в України велика частина малого та середнього бізнесу біла або ліквідована (особливо на окупаційних територіях), або переїхала, або перейшла на чатової/повністю в сферу електронної торгівлі. Великою проблемою е-бізнесу є те, що вони не мають достатньої інформації про можливості розвитку у певних локаціях та не мають зворотного зв'язку з їхніми споживачами. Або ця інформація надходить із запізненням або неповна, або з надлишковим шумом. В умовах війни варто також говорити не тільки про розвиток е-бізнесу, а й про його відновлення, адже багато підприємств зупиняються або взагалі руйнуються у зв'язку з війною. В таких умовах необхідні додаткові інструменти та інформаційні технології для допомоги бізнесменам слідкувати за можливостями розвитку е-бізнесу в певній локації, а також встановлювати зворотній зв'язок з користувачами за допомогою соціальних мереж та ЗМІ. Такі інструменти допоможуть суттєво розширити бачення ринкових можливостей для е-бізнесу, з'ясує – в які з них є сенс інвестувати, а на які не варто витрачати час. І врешті решт, побачити, яка ідея має майбутнє і яку бізнес-модель потрібну реалізувати/підтримувати/розвивати для стрімкого розвитку територіального/міжрегіонального е-бізнесу. Також допоможе розібратися, які важелі мають найбільший ефект для зміни політики бізнесу: що не чіпати, а що змінити, щоб забезпечити високу швидкість в реалізації задуму на основі аналізу відповідних результатів досліджень, наприклад, отримувати:

– прямий фідбек від клієнтів, динаміку зміни загальної задоволеності або зацікавленості цільової аудиторії та переваги/недоліки від користувачів за допомогою NLP-аналізу.

– підтримку розвитку е-бізнесу відносно локацій знаходження їхнього підприємства та найкращі напрями розвитку.

– графіки розвитку бізнесу (покращення/погіршення) залежно від змісту коментарів.

Метою дослідження є розробка інформаційної технології аналізу україномовних та англомовних відгуків користувачів-клієнтів на сайтах е-комерції, дописів та новин в соцмережах та ЗМІ на основі методів опрацювання природної мови та технології машинного навчання для просування, адаптації та подальшого розвитку відповідного е-бізнесу.

Для досягнення поставленої мети необхідно вирішити такі завдання:

– дослідження та порівняння аналогів;

– порівняння та дослідження сучасних методів NLP як лематизація і стемінг, вилучення ключових слів, аналіз настроїв, узагальнення тексту, мішок слів та токенізація;

– розробити модель системи класифікації відгуків клієнтів та новин з достовірних джерел для ідентифікації емоційне забарвлення тексту українсько-англійською мовами на основі класифікатора Naive Bayes;

– здійснити експериментальну апробації розробленої системи сентимент аналізу інформаційного простору як зворотна реакція цільової аудиторії для підтримки е-бізнесу в Україні.

Об'єкт дослідження – процеси аналізу емоційного забарвлення текстового контенту відгуків цільової аудиторії на товари/послуги е-комерції.

Предмет дослідження – методи та засоби сентимент-аналізу англомовного та україномовного текстового контенту відгуків користувачів.

1 ПОСТАНОВКА ПРОБЛЕМИ

Необхідно розробити таку систему, яка покликана спростити спілкування клієнтів та компаній, особливо для тих компаній, які не можуть собі дозволити повноцінний центр підтримки. Особливість цієї системи полягатиме у використанні NLP-алгоритмів

DOI 10.15588/1607-3274-2023-3-11

для скорочення витрат на обслуговування клієнтів за рахунок скорочення кількості активних працівників в компанії. На заміну людській силі прийде алгоритм штучного інтелекту, який сам класифікуватиме відгуки та скарги клієнтів і визначатиме потрібні дії для них. Систему аналізу тональності інформаційного простору як зворотна реакція цільової аудиторії для підтримки та розвитку е-бізнесу подано кортежем:

$$S = < I, O, R, U, N, \alpha, \beta, \gamma >,$$

де $I = \{i_1, i_2, i_3, i_4\}$, $O = \{o_1, o_2, o_3\}$, $R = \{r_1, r_2, r_3, r_4\}$, $U = \{u_1, u_2, u_3, u_4, u_5\}$.

Основними процесами ІС є «Збір відгуків», «NLP відгуків», «Машине навчання» та «Класифікація відгуків». Процес збору відгуків із соціальних мереж опишемо суперпозицією:

$$C_{AU} = \mu \circ \beta \circ \alpha,\ C_{AU} = \mu(\beta(\alpha(i_1, i_2, i_4), r_1, u_1), u_2).$$

Процес «NLP відгуків» ІС граматичної корекції опишемо суперпозицією: $C_{CU} = \chi \circ \beta \circ \alpha$, тобто

$$C_{CU} = \chi(\beta(\alpha(C_{AU}, i_2, i_3, i_4), r_1, u_3), r_2).$$

Процес машинного навчання на достовірних даних ІС граматичної корекції опишемо суперпозицією:

$$C_{UL} = \omega \circ \gamma \circ \beta \circ \alpha,\ C_{UL} = \omega(\gamma(\beta(\alpha(C_{CU}, i_2), i_3), u_4), r_3).$$

Процес «Класифікація відгуків» ІС на основі машинного навчання опишемо суперпозицією:

$$C_{US} = \lambda \circ \gamma \circ \beta \circ \alpha,\ C_{US} = \lambda(\gamma(\beta(\alpha(C_{US}, i_2), i_4), u_5), r_4).$$

2 АНАЛІЗ ЛІТЕРАТУРНИХ ДЖЕРЕЛ

Взаємодія між компанією та цільовою аудиторією досліджується вже століттями. З самого початку комерційних відносин, стосунки надавача послуг та отримувача цінувалися чи не понад усе. Торгівля побудована на довірі та повазі. Імідж підприємця часто є важливішим ніж товар, який він продає. За багато сотень років, взаємини торговця і покупця, підприємця та клієнта не втратили важливості і в час масової диджиталізації якість відносин компанії та цільової аудиторії різного розміру та професійна підтримка зворотного зв'язку з клієнтами часто визначають успіх е-бізнесу [1].

Взаємодія між компаніями та клієнтами – це складні стосунки, які дуже необхідно підтримувати в хороших тонах для компаній. Саме тому, що більше пів століття тому почали відкриватися центри підтримки клієнтів з арміями агентів, які допомагали покупцям. Проте час не стояв на місці, і вже зараз ці величезні центр не є ні корисними, ні вражаючими. Кожна компанія повинна тепер мати свій центр підтримки клієнтів. Проте, такі центри коштують дорого і, в часи стартапів і компаній, які з'являються і зникають однаково швидко, створювати домашній центр підтримки клієнтів з найманим персоналом невигідно. Зараз, в час глобальної диджиталізації та ще більшого прискорення руху життя, мати центри підтримки клієнтів, який оперує на базі агентів – невигідно. Адже швидкості бізнесу зростають, а з нею зростає й кількість нових клієнтів. Проте більше клієнтів – це не лише збільшення прибутку.

З іншого боку на сьогодення соціальні мережі займають велике, можливо навіть надто велике, місце в житті пересічної сучасної людини, потенційного клієнта конкретного е-бізнесу. Та швидкість, з якою новина може розлетітися по соціальних мережах – захоплює та лякає водночас. І саме в такому середовищі компаніям доводиться спілкуватися з клієнтами. Ціна поганого обслуговування клієнтів, в тому числі і підтримка, може бути надто велика. Саме тому важливо мати якісний, ефективний центр підтримки клієнтів. Саме центри підтримки клієнтів часто визначають ставлення загальної маси до компанії. Відношення компанії до цільової аудиторії збільшує не лише утримання е-бізнесу, а ще й слугує як безкоштовна реклама: якщо клієнту сподобалось товар/послуга та обслуговування – він скоріше порекомендує його бізнес іншим або залишить коментар/відгук в соціальній мережі.

Підтримка клієнтів – один з найважливіших аспектів багатьох підприємств та компаній. Проте, це не так вже і легко. Для ефективного центру підтримки клієнтів необхідно багато витрат – зарплата агентів, їхні робочі місця, інструктаж агентів. Це все – витрати. І для багатьох компаній ці витрати стають надто великими. Все більше компаній віддають перевагу фірмам-посередникам, які спеціалізуються на спілкуванні з цільовою авдиторією конкретного е-бізнесу. Це також вимагає певних витрат та часу на співпрацю та навчання персоналу під конкретний-е-бізнес. В сучасний час диджиталізації саме заміна таких call-центрів та фірм-посередників інструментом у вигляді інформаційної системи взаємодії з клієнтами та аналізу коментарів та новин на основі методів машинного навчання та NLP може стати успішним рішенням ведення бізнесу. NLP дозволяє застосовувати алгоритми машинного навчання для тексту та мови. Наприклад, ми можемо використовувати NLP, щоб створювати системи на кшталт розпізнавання мовлення, узагальнення документів, машинного перекладу, виявлення спаму, розпізнавання іменованих сутностей, відповіді питання, автодоповнення, предиктивного введення тексту тощо [4]. Завдяки новітнім і/або класичним алгоритмам, наприклад, тест Тюрінга [5], система може конкурувати з провідними компаніями на ринку аутсорсингу та, потенційно, змінити правила взаємодії з клієнтами. Тоді й невеликі компанії зможуть запросто утримувати лише кількох агентів, проте мати таку ж якість підтримки, як і гіганти їх індустрії з багатократними бюджетами, наприклад, на основі технології моделювання, синтезування та розпізнання мовлення [6]. Зараз також є дуже актуальною проблема вирішення задач NLP для

DOI 10.15588/1607-3274-2023-3-11

слов'янських мов, особливо української мови на фоні війни в Україні (наприклад, для ідентифікації фейків та пропаганди навіть актуально для е-бізнесу – приклад буде чи не буде війна в Тайвані змінює цінову політику на всі цифрові девайси), яка б дозволила слов'янським країнам якісно користуватись такими NLP рішеннями як: генерація тексту; аналіз настроїв; узагальнення тексту; та інші.

Аутсорсинг є стратегічним рішенням компанії для зниження витрат та підвищення ефективності бізнесу шляхом найму фізичної/юридичної особи для виконання відповідних завдань [7]. Аутсорсинг підтримки клієнтів досить поширена практика (наприклад, Sykes [8], Sensee [9], Serco [10], Teleperformance [11]), тому ринок аутсорсингових компаній, що спеціалізуються на спілкуванні з клієнтами досить обширний. В ньому знайдеться рішення практично для будь-якого е-бізнесу. Проте, якщо створити стартап як аналог виконання хоча б частини завдань відповідних аутсорсингових компаній, що буде більш заощадливим або ефективнішим, то це сильно підірве уже встановлений ринок. Проаналізувавши різні компанії та послуги, розроблено набір характеристик та критеріїв оцінювання для системи взаємодії з клієнтами:

– Доступ підтримки цілодобово – оцінюється наявність/відсутність підтримки цілодобового зв'язку;

– Швидкість зворотного зв'язку – скільки в середньому між усіма каналами годин потрібно для надання першої відповіді клієнту;

– Конфіденційність, ціна та кількість мов;

– Кількість агентів – значення кількості агентів не повинне бути надто високим та не надто низьким;

– Розташування та розміри офісу – розташування офісу повинне дозволяти охоплення якнайбільшу кількість клієнтів, розміри офісу повинні забезпечувати робоче місце для усіх агентів компанії;

– Кількість доступних каналів зв'язку;

– Можливості вхідного/вихідного зв'язку, телемаркетингу, активного збору відгуків;

– Глибина підтримки – на скільки агенти можуть допомогти клієнту тут і зараз.

Ще одним напрямом збору інформації та настроїв, яка впливає на розвиток е-бізнесу певного сектору, є платформи відстежування світових/регіональних медіа та друкованих ЗМІ, соціальних, онлайнових, цифрових і телерадіокомпаніях як Carma Media Monitoring, Repustate. Patient Voice [12–13], Siri [14], Grammarly [15], Klevu Smart Search [16] тощо. Зазвичай продукти, які використовують NLP в бізнесі є дуже зручними, але обмеженість в функціоналі не дає користувачам повністю покрити свої потреби. Тому в розроблюваному продукті потрібно залучити всі переваги аналогових продуктів, розширити функціонал продукту, який би покривав всі потреби клієнтів та основне виправив би недоліки продуктів-аналогів. Найкращим аналогом є Repustate, саме він має бути основним конкурентом, якого потрібно обійти. Даний продукт залучає велику кількість NLP методів, як і передбачається в розроблюваному продукті. Всі інші продукти розглянуті вище, зроблені за допомогою методів NLP та є лідерами у своїх сферах, тож маючи їх досвід можна залучити їхні підходи у якості розширення функціоналу для розроблюваного продукту, що зробить його лідером на ринку продуктів, які залучають NLP.

3 МАТЕРІАЛИ ТА МЕТОДИ

Для створення серйозного та процвітаючого е-бізнесу у будь-якій галузі із взаємодією з клієнтом необхідно приділяти час і увагу обслуговуванню цих клієнтів. Зрештою, команди служб обслуговування клієнтів щодня напряму взаємодіють із потенційно вашими клієнтами [2]. Це може принести як найбільше вигоди, так і найбільших збитків. Коли обслуговування клієнтів є пріоритетним, компанії отримують масу переваг: більше лояльних клієнтів, більше позитивних відгуків і більше доходу. Ось чому так важливо зосередитися на обслуговуванні клієнтів. Надання підтримки клієнтам може зайняти багато часу та енергії, тому традиційне обслуговування клієнтів часто розглядається як центр витрат. Керівники компаній знають, що їм потрібно надавати послуги, але вони бачать це як «витрати на ведення бізнесу». Проте спілкування з клієнтами може бути настільки ж прибутковим, як і розробка самого продукту. Обслуговування клієнтів – це не лише вартість ведення бізнесу. Це важлива частина загального досвіду клієнтів. Проте хороша підтримка клієнтів може привести до великих витрат що ніколи не є добре, особливо для менших компаній, чи таких, що лише починають комерційний шлях. Саме тому, все більше компаній [3] починають передавати проблеми з організацією та підтримки хорошого та ефективного сервісного центра іншим, аутсорсинговим компаніям чи стартапам. Отже є актуальним аналіз напрямів побудови інформаційної технології підтримки розвитку е-бізнесу України за допомогою аналізу локацій знаходження бізнесу, опрацювання фідбеку від користувачів, аналізу та класифікації відгуків клієнтів в режимі реального часу з соціальних мереж: Twitter, Reddit, Facebook та інші за допомогою методів глибокого навчання та NLP українсько- та англомовних текстів (рис. 1).

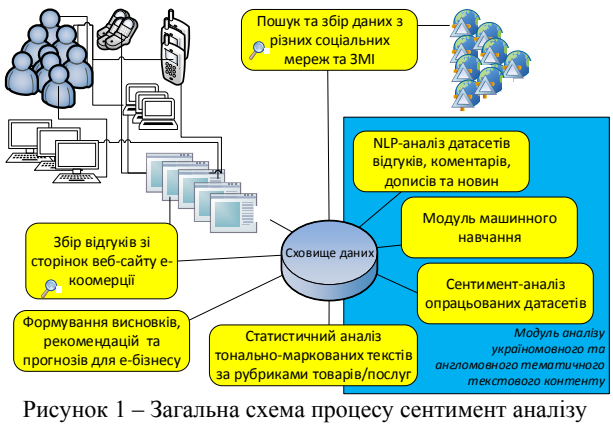

Рисунок 1 – Загальна схема процесу сентимент аналізу інформаційного простору

DOI 10.15588/1607-3274-2023-3-11

Тобто аналіз реакцій на товари-послуги через аналіз коментарів, відгуків на них на сайтах, в профілях соцмереж та паралельно новин в пабліках на подібні товари-послуги тощо. Вхідні дані – україномовний та англомовний контент з власних сайтів конкретного е-бізнесу, з профілів соцмереж постійних клієнтів та профілю самої фірми і паралельно з достовірних джерел ЗМІ, де можливі новини щодо цих подібних товарів-послуг, наприклад будівництва тощо. Необхідно розробити підхід для аналізу зворотної реакції цільової аудиторії для українського е-бізнесу, бо в сучасних умовах більше виживає на території України саме електронна торгівля. Тобто воюємо з часом – швидко та оперативно автоматично зібрати та проаналізувати реакцію цільової аудиторії для можливості скерувати бізнес. У час війни, постійного відключення в тому числі не за графіком світла, бізнес мусе адаптуватися швидко без використання стандартних для мирного часу інструментів та технік, в тому числі збору даних наприклад для прогнозування що буде актуальніше та краще реалізовуватися. Для якої саме аудиторії (вік, стать, регіон тощо). Це має бути технологія опрацювання вже зібраних даних з достовірних джерел для витягування певних реакцій (сентимент-аналіз, тональність відгуку позитивна, нейтральна чи негативна наприклад на товар). Коментарі зазвичай на сайтах українці пишуть або українською або російською. Другу не розглядали принципово. Українська не лише складна та багатогранна. Просто користувачі часто або не грамотні, або випадково пишуть з помилками, або застосовують суржик у залежності від регіону користувача в тому числі англійські слова вставляють англійською або транслітерацією. Та ще з помилками. А ось ті ж самі користувачі особливо молодь в соцмережах часто пишуть відгуки англійською. Тому і комбінація двома мовами та з різних джерел. Це подібно як бот збирає дані з достовірних джерел та фільтрує, а потім формує датасет (це детально в статі не описано, бо багато є подібних публікацій в тому числі авторів). В статті акцент надано лише на процес опрацювання датасетів двома мовами на основі NLP та машинного навчання. І акцент більше на сентимент-аналіз, що витягнути примітивну емоцію в тексті на конкретний або товар, або вид товару або послуги від наприклад конкретних користувачів. Щоб далі можна було аналізувати та прогнозувати на основі рекомендацій та зібраної статистики наприклад реакцію загальну на категорію товарів від певного класу цільової аудиторії.

NLP поєднує обчислювальну лінгвістику зі статистичними моделями, моделями машинного навчання та глибокого навчання. Разом ці технології дозволяють комп'ютерам опрацьовувати людську мову у вигляді тексту або голосових даних і «розуміти» її повне значення, враховуючи наміри та настрої мовця чи письменника [17]. NLP стало важливим бізнес-інструментом для виявлення прихованих даних із каналів соціальних мереж. Аналіз настроїв може аналізувати мову, що використовується в публікаціях у соціальних мережах, відповідях, оглядах тощо, щоб витягти ставлення та емоції у відповідь на продукти, рекламні акції та події – інформацію, яку компанії можуть використовувати в дизайні продуктів, рекламних кампаніях тощо. Також доречно NLP використати для класифікації відгуку клієнта. Єдина зовнішня дія яка потрібна для запуску роботи системи – це написання клієнтом відгуку. Цей відгук може бути написаний на будь-якій платформі: від соціальних мереж до Google Maps. Специфіка кількості та яких саме платформ узгоджує компанія, яка використовує систему. Після того, як клієнт написав свій відгук, система стягує цей відгук з визначеної платформи до власного сховища. Таким чином будується банк відгуків, які можна використати в подальших ітераціях моделі системи. Коли відгук стягнуто та записано до сховища, система проводить операцію класифікації відгуку. Це означає, що система визначає чи новий відгук позитивний чи негативний, перевіряє чи потрібна якась дія стосовно цього відгуку та яке слово з відгуку найточніше описує відгук загалом. Після успішної класифікації, залежно від результатів, система зберігає відгук в ще одне сховище для ведення архіву та передає інформацію далі до агентів, якщо це потрібно.

Оскільки ресурс планується бути онлайновим, для взаємодії з користувачем використовуватиметься доступні йому девайси. Коли користувач заходить на якусь з визначених платформ, він повинен натиснути відповідну кнопку, щоб залишити відгук. Після того, як користувач надіслав свій відгук, відгук автоматично стягується контролером системи (рис. 2).

Контролер передає цей відгук до Сховища, яке виконує зберігання сирого відгуку. Після того, як Сховище провело збереження, він надсилає статус відгука назад до Контролера для логування. Тоді, коли Контролер отримав зворотне повідомлення від Сховища, він надсилає відгук до Класифікатора. Класифікатор проводить класифікацію відгуку. Тоді, вже класифікований відгук надсилається назад Класифікатору, який, в свою чергу, надсилає інформацію про класифікований відгук Сховищу, щоб воно знову зберегло відгук, проте уже в опрацьованому вигляді. Після збереження відгуку, Сховище знову надсилає статус збереження Контролеру системи, де той продовжує потік, а саме, надсилає відгук Агенції. Агенція, залежно від того, що передбачив Класифікатор, або надсилає відгук далі до агентів, або закінчує шлях цього відгуку.

Система постійно моніторить доступні платформи на наявність нових відгуків. Цикл перевірки нових відгуків триває доти, поки не буде знайдений хоча б один новий відгук на будь-якій платформі. Якщо ж буде знайдено новий відгук, система вибивається з циклу та починає активну роботу.

Спочатку новий відгук зберігається у сховище. У сховище поступає будь-який відгук, який пройшов попередній етап, тому, можливо, що в сховищі можуть бути наявні однакові, або близькі за значенням та структурою відгуки. В будь-якому разі,

DOI 10.15588/1607-3274-2023-3-11

коли новий відгук поступає у систему і він записується в сховище – система передає новий відгук вниз по воронці та повертається до моніторингу нових відгуків. Завдяки цьому нові відгуки не будуть накопичуватись, що є важливим для швидкості опрацювання усіх відгуків. Після збереження та передачі відгука далі, йде найбільш затратний дія з усієї системи – класифікація. Тут відбуваються усі основні обрахунки системи, через що це є критична точка для ефективності системи. Важливо оптимізувати цю діяльність. Після класифікація, залежно від результатів, відгук або передається агентам для подальших дій, або відправляється в сховище для можливого подальшого використання, такого як аналіз, архівування, покращення та ітерація моделей класифікації. Якщо система вирішила, що відгук потребує дії, вона відправляє його агентам. Агенти повинні вирішити проблему, яку піднімає відгук так, як тільки це можливо.

Людська мова вражаюче складна і різноманітна. Люди висловлюються нескінченними способами, як усно, так і письмово. Існують не тільки сотні мов і діалектів, але й у кожній мові є унікальний набір граматичних і синтаксичних правил, термінів і сленгу. Коли люди пишуть, то часто роблять помилки, скорочують слова або пропускають розділові знаки. Також є регіональні акценти, бурмотіння, заїкання та запозичені терміни з інших мов, в тому числі в українській [18]. Усі бізнес дані містять багато корисної інформації, ідей, а NLP може швидко допомогти компаніям отримати їх. Інструменти NLP опрацьовують дані в режимі реального часу, 24/7, і застосовують одні й ті самі критерії до всіх даних, тому отримані результати є точними – і не містять невідповідностей. Після того, як інструменти NLP зможуть зрозуміти, про що йдеться в тексті, і навіть виміряти такі речі, як настрої, компанії можуть почати розставляти пріоритети та організовувати свої дані таким чином, щоб відповідати їхнім потребам [19]. Перед застосуванням методів машинного навчання будь-який текст англійською чи українською, або їх суміш має пройти попереднє опрацювання методами NLP, зокрема або частково в залежності від мети та типу задачі з врахуванням особливостей методу:

1) Тематичний аналіз – витягування сенсу із тексту шляхом визначення повторюваних тем [20];

– Тематичне моделювання може виводити шаблони та групувати подібні вирази без необхідності визначати теги теми або навчати дані заздалегідь;

– Класифікація тексту або виділення теми з тексту.

2) Аналіз настроїв – визначення того, чи є текст позитивним, негативним чи нейтральним на основі інших методів NLP і машинного навчання, щоб призначити зважені оцінки настроїв об’єктам, темам, темам і категоріям у реченні або фразі [21];

3) Виявлення намірів використовує машинне навчання та NLP, щоб пов’язувати слова/вирази з певним наміром. Наприклад, модель машинного навчання може дізнатися, що слова купити або придбати асоціюються з наміром придбати [22];

4) Вилучення ключових слів – техніка аналізу тексту, яка автоматично витягує з тексту найбільш вживані та найважливіші слова/вирази [23–24];

5) Лематизація – групування різних флективних форм слова для подальшого аналізу як єдиного елементу та на відмінну від стемінгу привносить контекст до слів, тобто зв’язує слова зі подібними значеннями в одне слово; використовують позиційні аргументи як вхідні дані, наприклад, чи є слово прикметником, іменником чи дієсловом [25–26];

6) Стемінг використовується для видалення суфіксів зі слів і в кінцевому підсумку отримати так звану основу слова, що дозволяє стандартизувати слова до їхньої основи незалежно від їх перегинів наприклад для кластеризації або класифікації тексту та пошуку [25–26];

7) Токенізація – це спосіб поділу фрагмента тексту на менші одиниці (токени) та використовують у традиційних методах NLP (Count Vectorizer), так і в архітектурах на основі розширеного глибинного навчання (Transformers); маркери можуть бути словами, символами або підсловами (n-грам) [27];

8) Машинний переклад – завдання автоматичного перетворення однієї природної мови в іншу, збереження значення введеного тексту та створення вільного тексту мовою виведення [28–29];

9) Узагальнення тексту – семантичне скорочення тексту, видаливши неважливий текст і перетворивши той самий текст у меншу семантичну текстову форму без видалення семантичної структури тексту [30]; визначення важливих фраз у документі та використання їх для виявлення відповідної інформації для додавання в резюме є критичною роботою для узагальнення на основі вилучення [31].

Для навчання моделей використовується дані, зібрані з відгуків у Google Maps на різного роду закладів: ресторани, готелі, кафе, магазини, тощо. В датасет входить відгук, записаний у вигляді стрічки, до якого класу відноситься цей відгук до класу позитивних відгуків, чи класу негативних відгуків, а також до якого класу відноситься цей відгук стосовно потреби у допомозі/діях. Загалом, у датасеті наявні три показники. Відгуки написані українською мовою, що суттєво ускладнює завдання. Також, загалом, в датасеті наявно приблизно 500 рядків даних. В рамках функцій цього процесу використаний словник української мови користувача GitHub DICT_uk, де зібрані більше мільйону українських слів, значення, приналежність до частин мови та більше [32]. Для класифікації опрацьованого тексту виберемо [33–34]:

– Naive Bayes Classifier – це група дуже простих алгоритмів класифікації, які базуються на теоремі Байєса; всі атрибути датасету є незалежні і що жодна з них не впливає на будь-яку іншу; є швидким та потребує мало даних для тренування, також, має хорошу тенденцію роботи з текстами при NLP;

– Підтримка векторних машин (SVM) – алгоритм, що використовується для класифікації та для

DOI 10.15588/1607-3274-2023-3-11

регресійних задач; розбиває дані в дві півплощини з найкращим можливим результатом, тобто, знаходить таку лінію на площині даних, що ділить ці дані на два класи; є швидкість тренування, висока точність та велика кількість можливих застосувань;

– Decision Tree – алгоритм розбиває датасет на малі підсети даних та вибудовує асоціативне дерево рішень для кожного з них; використовується для побудови моделі для передбачення цільових значень, де правила передбачення вибудовуються на основі попередніх даних; є простий та легкий в розумінні та реалізації з здатністю пояснення складних моделей за допомогою чітких візуалізацій, проте, легко піддається до оверфітингу та погано працює з нечисловими значеннями, також показує погані результати з малою кількістю даних.

Для розробки використаємо мову програмування Python [35–36] та його бібліотеки та фреймворки Flask [37], FastAPI [38] та NLTK [39], також для інтерфейсу буде застосовуватись javascript та його бібліотека React. Для ілюстраціїзміну даних на екрані, а не чекав повного опрацювання застосуємо message-broker Kafka [40]. Для створення кваліфікуючої частини системи класифікації відгуків, використаємо мову Python та середовище програмування Jupyter Notebook. Для реалізації алгоритму використаємо sklearn, а саме sklearn.naive_bayes.GausianNB. В рамках проекту також використовуються бібліотеки Python [35–36]: Numpy (робота моделей), Pandas (зберігання та трансформації даних), Re (маніпуляція з стрічками) та NLTK (з tokenize функція TreebankWordTokenizer для токеназації слів у реченнях) та Sklearn (машинне навчання).

4 ЕКСПЕРИМЕНТИ

Систему сентименту-аналізу відгуків користувачів подано кортежем:

$$S = < I, O, R, U, N, \alpha_i, \beta, \gamma, \mu, \omega, \chi, \phi_j, \lambda >,$$
$$i=[1,5], j=[1,5],$$

де $I = \{i_1, i_2, i_3, i_4\}$, $O = \{o_1, o_2, o_3\}$, $R = \{r_1, r_2, r_3, r_4, r_{5j}, r_6\}$, $U = \{u_1, u_2, u_3, u_4, u_5, u_{6j}\}$.

Основними процесами ІС є «Збір відгуків з різних джерел», «NLP відгуків», «Машине навчання» та «Класифікація відгуків за параметрами та ключовими словами». Додатковими процесами ІС є «Пошук за ключовими словами», «Сентимент аналіз», «Популярність запитів», «Узагальнення тексту», «Пошук оптимальних локацій», «Збір відгуків з Google», «Збір відгуків з Reddit», «Збір відгуків з Twitter», «Збір відгуків з Facebook», «Виведення отриманих результатів в вигляді звіту», «Виведення оттрманих результатів в вигляді графіку».

Процес формування колекції тематичних відгуків користувачів із різних джерел з попереднім опрацюванням (видалення дублів, інформаційного шуму, форматуванням за шаблоном) опишемо суперпозицією: $C_{AU} = \gamma \circ \lambda \circ \gamma \circ \mu \circ \beta \circ \alpha_i$, $i=[1,5]$, тобто

$$C_{AU} = \gamma(\lambda(\gamma(\mu(\beta(\alpha_i(i_1, i_2, i_4), r_1, u_1), u_2)), r_4)).$$

Дані отримуються в режимі реального часу, тому вгадати з препроцисінгом майже нереально, але можна покращити отримані від користувачів дані, наприклад з соціальної мережі Twitter інтегруються так звані сирі дані із великим обсягом сміття, не потрібного для дослідження (багато юнікод символів). Для цього на основі пакету re треба очистити дані від юнікоду. Регулярний вираз є послідовністю символів для визначення шаблону пошуку в тексті, наприклад, для операцій типу "знайти" або "знайти і замінити" над рядками або для перевірки введених даних [50]. При очищенні даних, тобто видалення інформаційного шуму з тестовго контенту, для кожного отриманого посту замінимо юнікод символи за допомогою RegEx та патерну "[^\x00-\x7F]+".

Для покращеної роботи з даним текстом токенізуємо пости за допомогою wordtokenization або regexptokenization (regexptokenization працює краще, бо вилучає зайві розділові знаки). за допомогою лемматизації на англомовного текстового контенту та основі стемінгу для україномовного контенту.

Процес «NLP відгуків» ІС для інтегрованого тематичного контенту опишемо суперпозицією:

$$C_{CU} = \gamma \circ \phi_j \circ \chi \circ \beta \circ \alpha_i, \ i=[1,5], j=[1,5], \text{ тобто}$$
$$C_{CU} = \gamma(\phi_j(\chi(\beta(\alpha_i(C_{AU}, i_2, i_3, i_4), r_1, u_3), r_2), r_{5j}, u_{6j})).$$

Методи NLP доцільно реалізувати за допомогою Python та відповідних бібліотек: Nltk (завантаження датасетів та імпорт класів Tokenizer), Re (regexp), SentimentIntensityAnalyzer (сентимент аналіз), WordNetLemmatizer (лемматизація речень на слова), PorterStemmer (стеммінг), Stopwords (словник стоп слів), Heapq.nlargest (визначає список n найбільших елементів в датасеті).

Процес машинного навчання та тренування на достовірних корпусах текстів для вдосконалення аналізу тематичних відгуків опишемо суперпозицією:

$$C_{UL} = \omega \circ \gamma \circ \phi_j \circ \beta \circ \alpha_i, \ i=[1,5], j=[1,5], \text{ тобто}$$
$$C_{UL} = \omega(\gamma(\phi_j(\beta(\alpha_i(C_{CU}, i_2), i_3), r_{5j}, u_{6j}), u_4), r_3).$$

Процес «Виведення отриманих результатів» ІС на основі машинного навчання опишемо суперпозицією:

$$C_{US} = \upsilon \circ \gamma \circ \beta \circ \alpha_i, \ C_{US} = \upsilon(\gamma(\beta(\alpha_i (C_{US}, i_2), i_4), u_5), r_6).$$

Опис розгорнутого сценарію прецеденту за стандартом RUP (рис. 2–3) [41–49]:

DOI 10.15588/1607-3274-2023-3-11

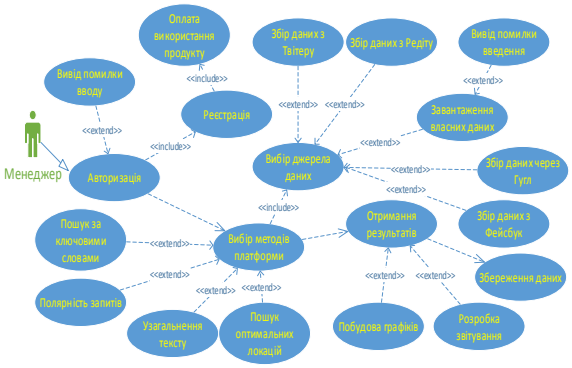

Рисунок 2 – Діаграма варіантів використання

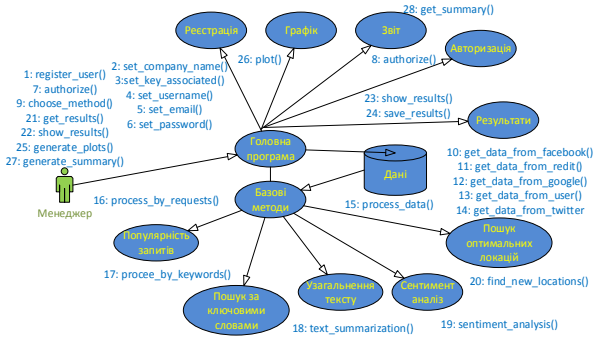

Рисунок 3 – Діаграма кооперацій

1) Менеджер отримує результат сентимент-аналізу множини відгуків відповідно до тематичного запиту на категорію товарів/послуг тощо;

2) Основний користувач системи сентимент-аналізу множини відгуків – менеджер е-комерції.

3) Передумови прецеденту (preconditions): підтримка збору тематичного контенту з Google та соціальних мереж за вимогою та їх опрацювання;

4) Основний успішний сценарій Менеджера: входить в систему → реєструється/авторизується → якщо вперше або закінчився термін, оплачує підписку → вибирає методи/ джерела → отримує результати;

5) Альтернативні потоки (рис. 3–4), коли Менеджер:

– не може ввійти в систему: повідомлення про помилку → повернення користувача на початок.

– задав некоректні дані: отримує повідомлення про помилку, що дані введені некоректно → повторне надсилання даних надсилають в систему.

– вибирає методи (рис. 3): Пошук за ключовими словами; Сентимент аналіз; Популярність запитів; Узагальнення тексту; Пошук оптимальних локацій.

– обирає джерела: Google; Reddit; Twitter; Facebook;

– вибирає розширені результати через Побудову графіків та Розробку звітування.

6) Postconditions: Менеджер отримав результати;

7) Спеціальні системні вимоги – це забезпечити надійність передачі даних, зручним інтерфейсом, цілодобову підтримку та швидке опарцювання запиту.

8) Список необхідних технологій: веб-платформа з підтримкою візуального відображення результатів.

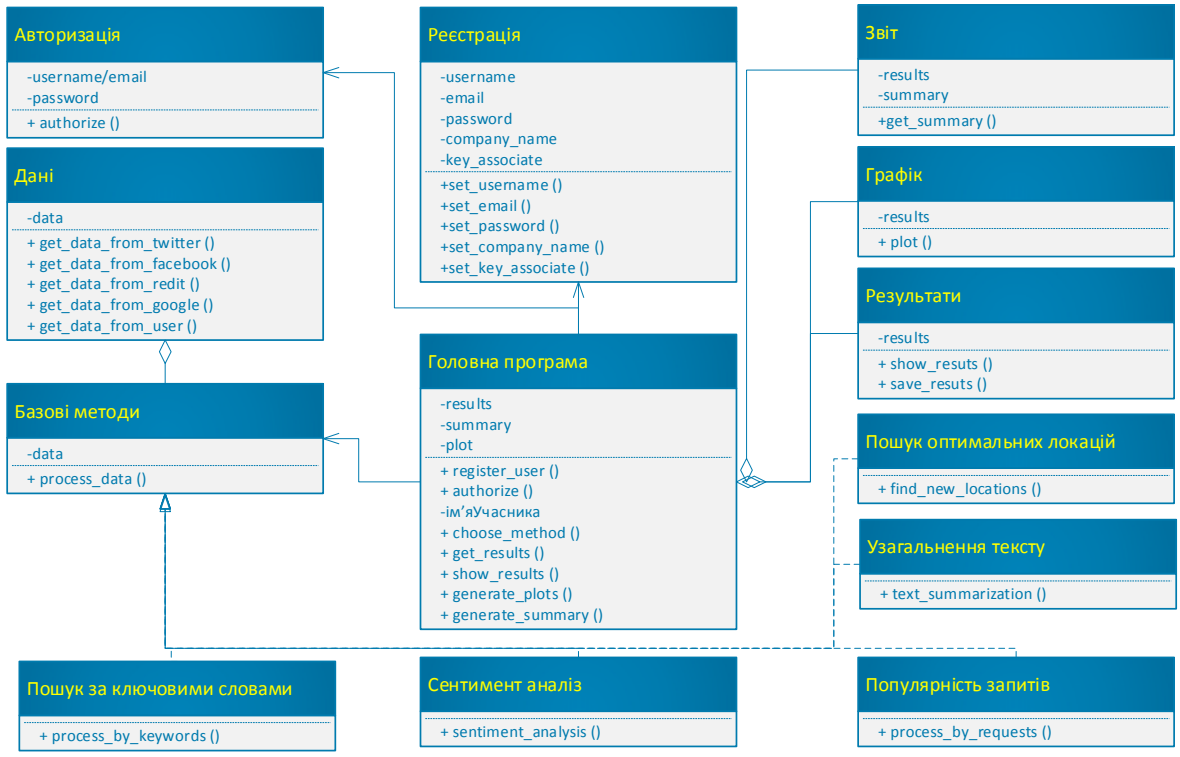

Рисунок 4 – Діаграма класів

DOI 10.15588/1607-3274-2023-3-11

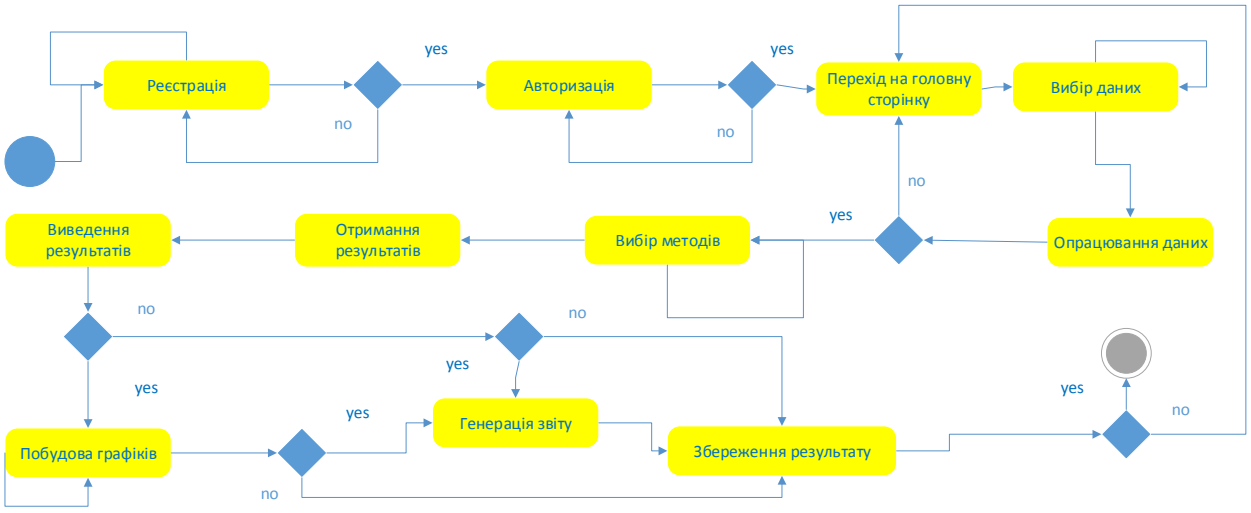

Рисунок 5 – Діаграма діяльності

На рис. 6 зображені компоненти, з яких складається розроблювана програма. Опишемо взаємодію між компонентами програми:

1. Main.py – даний компонент виконує роль керівника серед компонентів.

2. Authorization.py – компонент, який виконує роль авторизації, даний компонент ділиться на SignIn – логування в системі, та SignUp – реєстрація в системі;

3. DataGathering.py – компонент, який виконує роль збирання та опрацювання даних, включаючи:

– Get_data_from_twitter – дані з Twitter;
– Get_data_from_Reddit – дані з Reddit;
– Get_data_from_Google – дані з Google;
– Get_data_from_Facebook – дані з Facebook;
– Get_data_from_user – дані від користувача;

4. Methods.py – компонент, який містить різноманітні NLP методи та інші для аналізу даних:

– SentimentAnalysis – сентимент аналіз;
– Search_by_keywords – пошук за словом;
– Popularity_of_requests – популярність запиту;
– Text_summarization – узагальнення тексту;
– Look_for_new_locations – пошук нових локацій

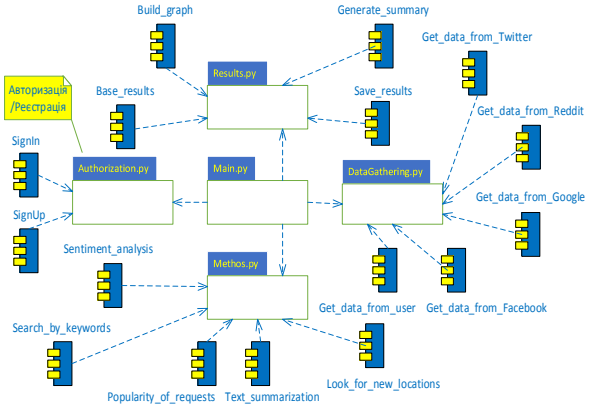

Рисунок 6 – Діаграма компонентів

5. Results.py – компонент, який відповідає за формування результатів, отриманих під час аналізу:

– Base_results – формує базові результати аналізу;
– Build_graph – побудова графіків результатів;
– Generate_summary – генерація результатів звіту.
– Save_results – збереження результатів.

5 РЕЗУЛЬТАТИ

При сентимент-аналізі користувач вводить ключ, згідно до якого він хоче отримати оцінку настроїв, наприклад назву компанії, товару, категорії товару тощо. Далі дані завантажуються з Twitter та опрацьовуються за допомогою regexp. Після чого ініціалізується об'єкт SentimentIntensityAnalyzer, також змінні для визначення кількості позитивних, негативних та нейтральних постів (рис. 7). Для кожного посту визначаємо оцінку настрою та за допомогою compound визначаємо до якої групи належить пост. Якщо ≤–0.05, то негативний, а ≥ 0.05 – позитивний, інакше пост можна вважати нейтральним. Після чого формуємо відсотковий розподіл та відправляємо дані клієнту.

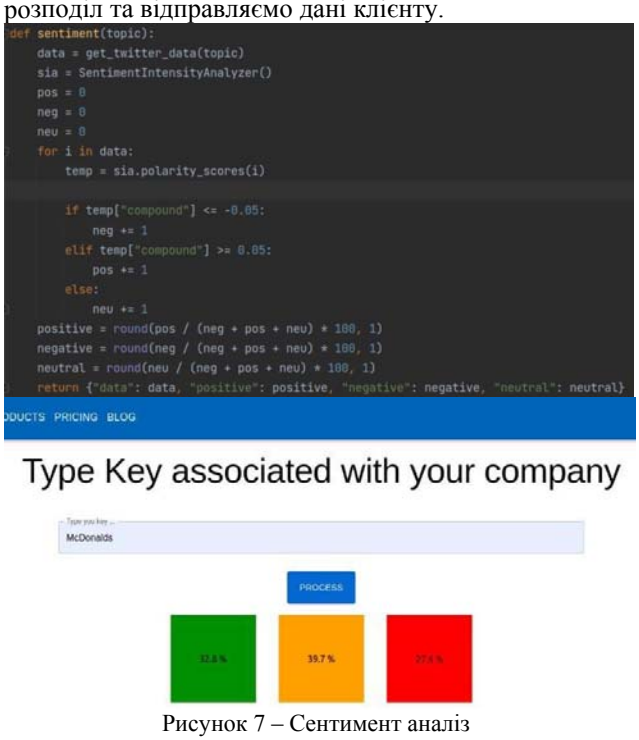

Рисунок 7 – Сентимент аналіз

DOI 10.15588/1607-3274-2023-3-11

При лематизації англомовного тексту застосовуємо RegexpTokenizer та WordNetLemmatizer (рис. 8). Для кожного посту реалізується токенізація, а для кожного токена застосовують Лемматайзер, після чого формуєтьмся множина результату лематизації.

Рисунок 8 – Лематизація

При стеммінгу ініціалізуються PorterStemmer, RegexpTokenizer та змінна results як масив (рис. 9). При узагальненні тексту постів їх об’єднують в однин масив, маркують стоп-слова датасету, ініціалізують та застосовюють RegexpTokenizer (рис. 10).

Рисунок 9 – Стеммінг

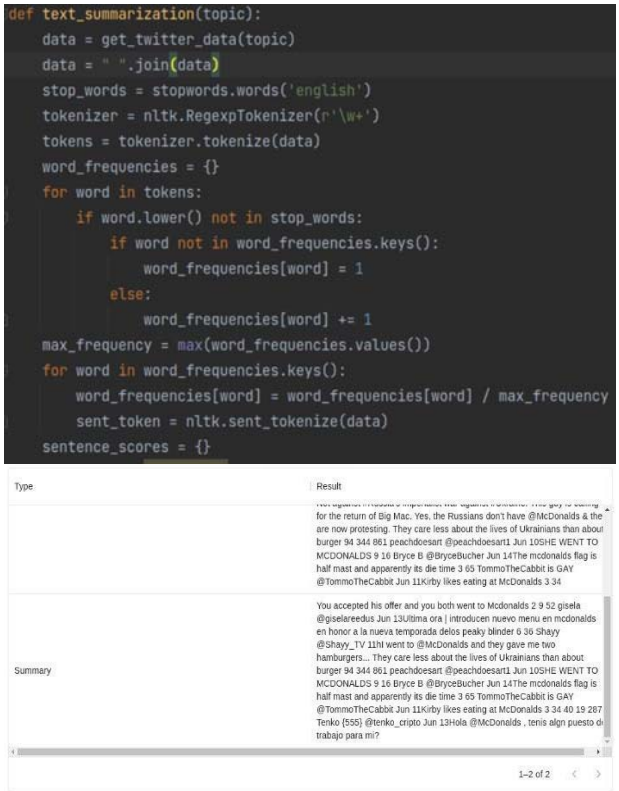

Рисунок 10 – Узагальнення тексту

Для всіх токенів, які не відносяться до стоп слів, створюємо частотний словник, а згодом нормалізуємо частоту на основі найбільшої знайденої частоти. Для кожного речення збираємо частоту появи слів в інших реченнях, після чого за допомогою nlargest алгоритму формуємо узагальнення та об’єднуємо в одне ціле.

При Pos Tagging слів ініціалізуємо RegexpTokenizer та змінну result як масив, застосовуємо алгоритм nltk.pos_tag (рис. 11).

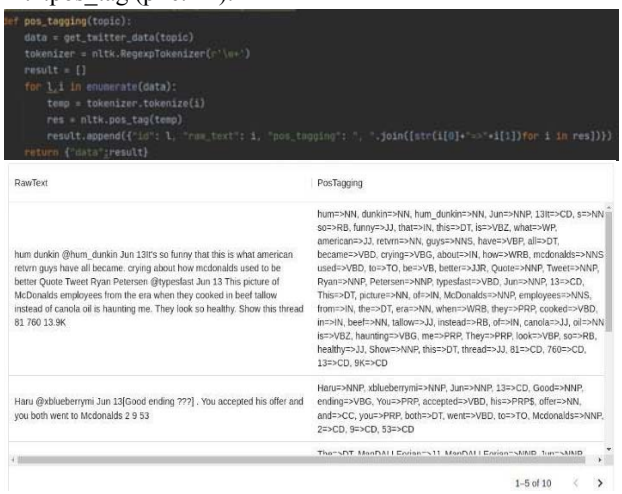

Рисунок 11 – Pos Tagging слів та Токенізація

DOI 10.15588/1607-3274-2023-3-11

При токенізації ініціалізуємо RegexpTokenizer та змінну result як масив (рис. 12).

```
def tokenization(topic):
    data = get_twitter_data(topic)
    tokenizer = nltk.RegexpTokenizer(r'\w+')
    result = []
    for l, i in enumerate(data):
        temp = tokenizer.tokenize(i)
        result.append({"id": l, "raw_text": i, "tokenization": ", ".join(temp)})
    return {"data": result}
```

RawText	Tokenization
hum dunkin @hum_dunkin Jun 13It's so funny that this is what american retvrn guys have all became. crying about how mcdonalds used to be better Quote Tweet Ryan Petersen @typesfast Jun 13 This picture of McDonalds employees from the era when they cooked in beef tallow instead of canola oil is haunting me. They look so healthy. Show this thread 81 761 13,9K	hum, dunkin, hum_dunkin, Jun, 13It, s, so, funny, that, this, is, what, american, retvrn, guys, have, all, became, crying, about, how, mcdonalds, used, to, be, better, Quote, Tweet, Ryan, Petersen, typesfast, Jun, 13, This picture, of, McDonalds, employees, from, the, era, when, they, cooked, in, beef, tallow, instead, of, canola, oil, is, haunting, me, They, look, so, healthy, Show, this, thread, 81, 761, 13, 9K
Haru @xblueberrymi Jun 13[Good ending ???] . You accepted his offer and you both went to Mcdonalds 2 9 53	Haru, xblueberrymi, Jun, 13, Good, ending, You, accepted, his, offer, and, you, both, went, to, Mcdonalds, 2, 9, 53
The ManDALLEorian @ManDALLEorian Jun 12Max Rebo working at McDonalds 15 132 1,013	The, ManDALLEorian, ManDALLEorian, Jun, 12Max, Rebo, working, at, McDonalds, 15, 132, 1, 013
Julia Davis @JuliaDavisNews Jun 12If only the Russians came out en masse to protest the war in Ukraine, but no, this is just the re-opening of McDonalds by another name. From 490 1,273 4,695	Julia, Davis, JuliaDavisNews, Jun, 12If, only, the, Russians, came, out, en, masse, to, protest, the, war, in, Ukraine, but, no, this, is, just, the, re, opening, of, McDonalds, by, another, name, From, 490, 1, 273, 4, 695
Inna Sovsun @InnaSovsun Jun 12The Russian protests. Do you know against what? Not against #Russia's imperialist war against #Ukraine. This	Inna, Sovsun, InnaSovsun, Jun, 12The, Russian, protests, Do, you, know, against, what, Not, against, Russia, s, imperialist, war, against, Ukraine,

1–5 of 10

Рисунок 12 – Реалізація та результат токенізації

Продемонструємо тепер роботу системи з україномовними текстами за допомогою бібліотеки Pandas (рис. 13). Датасет відгуків записаний у вигляді tsv-фалу (наявна пунктуація). Після завантаження тестових також потрібно створити масив стоп- слів (не несуть ніякого змісту або є надлишковим шумом).

```
data = pd.read_csv('data.tsv', delimiter = '\t', quoting =3)
slova = pd.read_csv('base.lst')
```

Рисунок 13 – Завантаження даних для навчання

Найпоширенішими службовими стоп-словами в українськомовних постах є ‘я’, ‘ти’, ‘там’, ‘де’, тощо (рис. 14). Також, стоп-словом є і ‘не’, проте виключимо це слово з масиву, оскільки воно достатньо сильно впливає на значення відгуку. Як частина класифікації система визначає найголовніше слово у відгуку на основі того, на скільки часто слово з’являється в українській мові.

```
stop_words = ['я',
 'мій',
 'та',
 'сам',
 'ми',
 'наш',
 'самі',
 'ти',
```

```
def most_import(review):
    d = {'words': review, 'freq': [0]*len(review)}
    df = pd.DataFrame(data = d)
    for word in review:

        if word not in set(stop_words) and word != 'не':
            for i in range(len(word)):

                for each in freq['word']:
                    if each.startswith(word[:i+1]) and len(each) <= len(word):
                        refr = each

                df.loc[df['words'] == refr, 'freq'] = freq['freq'][freq['word'] == refr].iloc[0]

    mst_word = df[df['freq'] == df['freq'].max()]['words'].iloc[0]

    return mst_word
```

Рисунок 14 – Масив стоп-слів української мови та визначення найголовнішого слова у відгуку

```
def Ukr_stem(review):
    stemmed = []

    for word in review:
        det = 0
        bord = 0
        word_len = len(word)
        found = False

        if word not in set(stop_words):

            if word_len <= 3 and not found:
                found = True
                stemmed.append(word)
            elif word_len == 4:
                for each in slova['word']:
                    if each == word:
                        found = True
                if found:
                    stemmed.append(word)
                else:
                    stemmed.append(word[:-1])

            else:

                root = word
                for i in range(len(word)):
                    for each in slova['word']:
                        if i != 0:
                            if each.startswith(word[:-i]) and i < (len(word)-3):
                                if len(word[:-i]) < len(root):
                                    root = word[:-i]

                stemmed.append(root)

    return stemmed
def ukr_stem2(review):
    stemmed = []
    review = [word for word in review if word not in stop_words]
    for word in review:
        root_len = len(word)-1 if word[-1:] in set(let_1) and len(word) > 2 else 0
        root_len = len(word)-2 if word[-2:] in set(let_2) and len(word) > 3 else root_len
        root_len = len(word)-3 if word[-3:] in set(let_3) and len(word) > 4 else root_len
        root_len = len(word)-4 if word[-4:] in set(let_4) and len(word) > 5 else root_len
        if root_len == 0:
            root = word
            for i in range(len(word)):
                for each in slova['word']:
                    if i != 0:
                        if each.startswith(word[:-i]) and i < (len(word)-3):
                            if len(word[:-i]) < len(root):
                                root = word[:-i]

            stemmed.append(root)
        else:
            root = word[:root_len]
            stemmed.append(root)
    return stemmed
```

Рисунок 15 – Функції стемінгу Ukr_stem та ukr_stem2

Функція перевіряє перші букви слів і знаходить таке слово, яке найближче підходить, ітеруючи стільки раз, скільки є букв у слові. З кожною ітерацією кількість перших букв збільшується і в кінці записується те слово зі словника, яке мало найбільшу кількість збігів. Далі, для оптимальної класифікації відгуків, необхідно їх підготувати, перед тим як навчати моделі на їх основі. Для цього потрібно провести ряд операцій: видалення пунктуації; переведення усіх літер в нижній регістр; токенізація; стемінг. Видалення та пониження регістру здійснюємо за допомогою імпортованої бібліотеки Re, яка призначена для роботи з регулярними висловами, та за допомогою функції lower() відповідно. Токенізацію здійснюємо за допомогою функції TreebankWordTokenizer. Стемінг – скорочення слів до найменшої можливої форми, коли зміст слова зберігається. Стемінг є ключовим у будь-яких NLP алгоритмах, оскільки грамотно проведене скорочення слів дозволяє оптимізувати роботу пізніших моделей. Розроблена функція стемінгу для української мови Ukr_stem (рис. 15).

Розроблена функція ukr_stem2 є другою ітерацією функції стемінгу українських слів. Функція Ukr_stem є повільною та неповорткою, але точною. Головна ідея Ukr_stem полягала у порівнянні кожного слова

DOI 10.15588/1607-3274-2023-3-11

відгуку з словами із словника, що займало багато часу. Функція ukr_stem2 є кращим варіантом для отримання швидко оперативних даних. Вона перевіряє закінчення слів та підбирає найкраще скорочення для нього. Для цього також створено масиви, що містять найпопулярніші закінчення слів в українській мові. За зразок взято дерево закінчень усіх можливих українських слів, розроблений на основі GNU Aspell Сеником Миколою (рис. 16) [51].

Рисунок 16 – Масиви закінчень слів української мови

```
def prepro(data):
    tokenizer = TreebankWordTokenizer()

    corpus = []
    most_import_word = []

    for rev in data:
        review = re.sub("[^а-яА-Я-іїґєЄ']", ' ', rev)
        review = review.lower()
        review = tokenizer.tokenize(review)
        most_import_word.append(most_import(review))
        review = ukr_stem2(review)
        review = ' '.join(review)
        corpus.append(review)
    df = pd.DataFrame(data = corpus)
    df['import_word'] = most_import_word
    return df
```

```
proc_reviews = prepro(data['review'])
proc_reviews.columns = ['review', 'import_word']
```

```
cv = CountVectorizer(max_features = 250)
x = cv.fit_transform(proc_reviews['review']).toarray()
y1 = data.iloc[:, 1].values
y2 = data.iloc[:, 2].values
```

Рисунок 17 – Функція prepro, створення Bag of Words

Як видно з рисунків, нова функція є набагато коротшою в плані коду, а також має набагато менше порівнянь, що скорочує час опрацювання відгуків, оскільки саме порівняння є найдовшою операцією в плані часу. Процес попереднього опрацювання відгуків подано функцією prepro (рис. 17).

Результати попереднього опрацювання текстів відгуків використаємо для навчання моделей та створюємо модель Bag of Words [52], на основі якої і навчатимуться моделі. У цій моделі текст (наприклад, речення або документ) подається як пакет (мультинабір) його слів, нехтуючи граматикою і навіть порядком слів, але зберігаючи множинність. Для класифікаторів наївного Байєса така модель підходить найкраще. Розбиваємо датасет на тренувальний та тестовий сети (рис. 18). Та проводимо навчання обох моделей. Тестувальний сет дорівнюватиме лише 2% від всього датасету для максимізації точності навчання моделей.

```
x_train, x_test, y_train1, y_test1 = train_test_split(x, y1, test_size = 0.02, random_state = 15)
x_train, x_test, y_train2, y_test2 = train_test_split(x, y2, test_size = 0.02, random_state = 15)
```

```
model1 = GaussianNB()
model1.fit(x_train, y_train1)

model2 = GaussianNB()
model2.fit(x_train, y_train2)
```

```
GaussianNB()
```

Рисунок 18 – Розподіл датасету, тренування моделей

Дослідимо точність роботи натренованих моделей на основі маркерів визначення сентименту (негативний/позитивний) відгука (рис. 19). Загальна точність сентиментальної моделі помилок для тестового сету є доволі задовільною (92,3%). Навчена модель добре класифікує позитивні відгуки, але має деякі проблеми з негативними (рис. 19а). Проблеми можуть бути через те, що негатив у відгуку людина, особливо українці, не передають на пряму, частіше нейтральними висловами, або сарказмом. Щодо другої моделі, то результати не настільки хороші при класифікації дії, з цим маємо проблеми (рис. 19б). Загальна точність моделі не є надто високою (61.5%). З матриці помилок випливає, що найбільше помиляється з тими відгуками, які не потребують дії.

```
cm1 = confusion_matrix(y_test1, predict1)
print(cm1)
accuracy_score(y_test1, predict1)*100
```

```
[[7 1]
 [0 5]]

92.3076923076923
```

а

```
cm2 = confusion_matrix(y_test2, predict2)
print(cm2)
(accuracy_score(y_test2, predict2))*100
```

```
[[5 4]
 [1 3]]

61.53846153846154
```

б

Рисунок 19 – Результат роботи сентимент- моделі та точність роботи сентимент-моделі

Проведемо тестування “живими” відгуками. Дамо класифікувати моделям щойно взятий відгук з інтернету та ще один, який написаний авторами. Обидва відгуки є позитивними і не потребують ніяких додаткових дій (рис. 20).

```
rev = ["Чудовий заклад! Приємний персонал, дуже смачна їжа, коктейлі і кальян) На 14 лютог
       "улюблений ресторан, кращого в світі не знайдеш"]
prot_test(rev)
```

```
Головне слово відгука: лютого;  Позитивне/негативне: 1; Потрібна допомога: 0
Головне слово відгука: світі;   Позитивне/негативне: 1; Потрібна допомога: 0
```

Рисунок 20 – Результат тестування «живих» відгуків

Також система визначила найголовніші слова відгуків. Для класифікації двох відгуків їй знадобилося 37 секунд, що означає приблизно 18.5 секунд на один відгук. Звісно, довжина відгуку сильно впливає на тривалість класифікації, оскільки найбільше часу все ще займає стемінг.

6 ОБГОВОРЕННЯ

Здійснено аналіз на основі методів машинного навчання (рис. 21): наївний Баєсів класифікатор (точність передбачення 71,13%), логістична регресія (точність передбачення 75,67%) та опорних векторів (точність передбачення 72,78%).

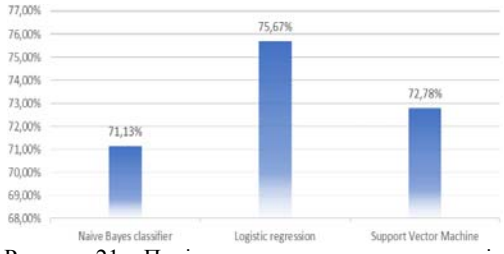

Рисунок 21 – Порівняння отриманих результатів

DOI 10.15588/1607-3274-2023-3-11

	precision	recall	f1-score	support
negative	0.52	0.54	0.53	128
neutral	0.76	0.87	0.81	575
positive	0.66	0.46	0.54	267
accuracy			0.71	970
macro avg	0.65	0.62	0.63	970
weighted avg	0.70	0.71	0.70	970

71.1340206185567

	precision	recall	f1-score	support
negative	0.70	0.52	0.59	128
neutral	0.78	0.88	0.83	575
positive	0.70	0.60	0.65	267
accuracy			0.76	970
macro avg	0.73	0.67	0.69	970
weighted avg	0.75	0.76	0.75	970

75.6701030927835

	precision	recall	f1-score	support
negative	0.80	0.26	0.39	128
neutral	0.72	0.97	0.82	575
positive	0.77	0.42	0.55	267
accuracy			0.73	970
macro avg	0.76	0.55	0.59	970
weighted avg	0.74	0.73	0.69	970

72.78350515463917

Рисунок 22 – Аналіз методом класифікатора Баєса, логістичної регресії та опорних векторів

На рис. 22 подані classifiaction report модулів для вимірювання якості прогнозів за алгоритмом класифікації (порівняння правдивих та хибних передбачень).Відповідно для прогнозування показників в classifiaction report використовують істинно позитивні TP, хибно позитивні FP, істинно негативні TN та хибно негативні FN показники передбачень [53–55]. Зокрема, коли випадок (подія) є при TN – негативним і прогнозовано негативним; TP – позитивним і прогнозовано позитивним; FN – позитивним, але передбачався негативним; FP – негативним, але передбачався позитивним.

ВИСНОВКИ

Описано застосування сентимент аналізу коментарів, відгуків, запитів та новин для підтримки та розвитку е-бізнесу. Проаналізовані аналоги дали можливість розробити інформаційну технологію для розв'язку NLP-задач е-бізнесу, адаптовану для української цільової аудиторії. Розроблена загальна типова структура інформаційної системи підтримки та розвитку е-комерції за рахунок аналізу зворотної реакції цільової аудиторії на основі технології машинного навчання та методів опрацювання природної мови. Серед методів реалізації основних функцій використані такі методи машинного навчання, як: наївний Баєсів класифікатор, логістична регресія та метод опорних векторів. Здійснено розробку програмного забезпечення та описано його структуру. Здійснено огляд звітів виконання методів машинного навчання. Це дало змогу краще переглянути та проаналізувати отримані результати. Опісля чого здійснено статистику виконання програми, описано її та проаналізовано отримані результати. А саме побудовано графік порівняння отриманих результатів. Також у ході роботи створена презентація про розроблений проект і написана стаття, у якій двома мовами, а саме українською та англійською, описано процес роботи над проектом. Найкраще з завданням аналізу впливу новини на фінансовий ринок впорався метод логістичної регресії, який показав точність 75,67%. Безперечно, це не є бажаним результатом, проте це найбільший показник із усіх розглянутих. Дещо гірше зі завданням впорався метод опорних векторів (SVM), який показав точність 72,78%, що є дещо гіршим результатом за той, який було отримано завдяки методу логістичної регресії. І найгірше зі завданням впорався метод наївного Баєсового класифікатора, який отримав точність 71,13%, що є меншою за отриману у двох попередніх методах. Звісно ж, що отримані результати далекі від ідеалу і демонструють точність у проміжку від 71 % до 76 %. Що означає те, що вони потребують удосконалення. На кінець хотілось би зазначити, що дана тема є неабияк популярною та актуальною, а аналогів на даний момент не існує.

ЛІТЕРАТУРА

1. Kuzminov M. Modern Development of Small Business in Ukraine / M. Kuzminov // Sciences of Europe. – 2022. – Vol. 107. – P. 29–31. DOI: 10.5281/zenodo.7479719
2. Definition of customer support. – Access mode: https://www.helpscout.com/helpu/definition-of-customer-support.
3. Edvardsson I. R. Strategic outsourcing in SMEs / I. R. Edvardsson, S. Durst, G. K. Oskarsson // Journal of small business and enterprise development. – 2020. – Vol. 27(1). – P. 73–84. DOI: 10.1108/JSBED-09-2019-0322
4. Sarkar D. Text analytics with Python: a practitioner's guide to natural language processing / D. Sarkar. – Bangalore : Apress, 2019. – 674 p. DOI: 10.1007/978-1-4842-4354-1
5. Eisenstein J. Introduction to natural language processing / J. Eisenstein. – Cambridge : MIT press, 2019. – 536 p.
6. Goldberg Y. A Primer on Neural Network Models for Natural Language Processing / Y. Goldberg. – Access mode: https://jair.org/index.php/jair/article/view/11030/26198
7. Britannica dictionary. – Access mode: https://www.britannica.com/topic/outsourcing
8. Sykes. – Access mode: https://www.sykes.com
9. Sensee. – Access mode: https://www.sensee.co.uk/index.html
10. Serco. – Access mode: https://www.serco.com
11. Teleperformance. – Access mode: https://www.teleperformance.com/en-us
12. Repustate. Using NLP for business success. – Access mode: https://www.repustate.com/blog/using-nlp-for-business-success/
13. Repustate. How can sentiment analysis help you with Patient Voice? – Access mode: https://www.repustate.com/patient-voice/
14. SkywellSoftware. How does Siri work: technology and algorithm. – Access mode: https://skywell.software/blog/how-does-siri-worktechnology-and-algorithm/
15. Grammarly. How Grammarly uses Natural Language Processing and Machine Learning to identify the main points in a message. – Access mode: https://www.grammarly.com/blog/engineering/nlp-mlidentify-main-points/

DOI 10.15588/1607-3274-2023-3-11

16. Klevu. Smart Search Overview. – Access mode: https://www.klevu.com/smart-search/
17. IBM. Natural Language Processing (NLP). What is natural language processing? – Access mode: https://www.ibm.com/cloud/learn/natural-language-processing#tocwhat-is-na-jLju4DjE
18. SaS. Natural Language Processing (NLP). What it is and why it matters. – Access mode: https://www.sas.com/en_us/insights/analytics/what-is-natural-languageprocessing-nlp.html
19. MonkeyLearn, What is NLP. – Access mode: https://monkeylearn.com/blog/what-is-natural-language-processing/
20. Topic Analysis: The Ultimate Guide. – Access mode: https://monkeylearn.com/topic-analysis/
21. Sentiment Analysis Explained. – Access mode: https://www.lexalytics.com/technology/sentiment-analysis/
22. MonkeyLearn. Intent Classification: How to Identify What Customers Want. – Access mode: https://monkeylearn.com/blog/intentclassification/
23. MonkeyLearn. Keyword Extraction. – Access mode: https://monkeylearn.com/keyword-extraction/
24. Edia. What is Keyword Extraction? – Access mode: https://www.edia.nl/keyword-extraction
25. Stemming vs. Lemmatization in NLP. – Access mode: https://towardsdatascience.com/stemming-vslemmatization-in-nlp-dea008600a0
26. Analytics steps. What is Stemming and Lemmatization in NLP? – Access mode: https://www.analyticssteps.com/blogs/what-stemming-andlemmatization-nlp
27. What is Tokenization in NLP?. – Access mode: https://www.analyticsvidhya.com/blog/2020/05/what-is-tokenizationnlp/
28. Stanford. Machine Translation. – Access mode: https://nlp.stanford.edu/projects/mt.shtml
29. Data Science UA. Machine Translation. – Access mode: https://data-science-ua.com/wiki/natural-language-processingnlp/machine-translation/
30. Text Summarization in NLP. – Access mode: https://www.topcoder.com/thrive/articles/text-summarization-in-nlp
31. What Is Text Summarization in NLP? – Access mode: https://www.analyticssteps.com/blogs/what-text-summarization-nlp
32. Dict_uk Github repository. – Access mode: https://github.com/brown-uk/dict_uk/tree/master/data
33. Advantages and disadvantages of different classification models. Access mode: https://www.geeksforgeeks.org/advantages-and-disadvantages-of-different-cl assification-models/
34. Naive Bayes Classifier. – Access mode: https://www.upgrad.com/blog/naive-bayes-classifier/
35. Coursera. What Is Python Used For?. – Access mode: https://www.coursera.org/articles/what-is-python-used-fora-beginners-guide-to-using-python
36. Python.org. Executive Summary. – Access mode: https://www.python.org/doc/essays/blurb/
37. PymBook. Introduction to Flask. – Access mode: https://pymbook.readthedocs.io/en/latest/flask.html
38. FastApi. – Access mode: https://fastapi.tiangolo.com/
39. NLTK. Natural Language Toolkit. – Access mode: https://www.nltk.org/
40. AWS. What is Apache Kafka? – Access mode: https://aws.amazon.com/ru/msk/what-is-kafka/
41. Tutorialspoint. System Analysis and Design – Overview. – Access mode: https://www.tutorialspoint.com/system_analysis_and_design/system_ana lysis_and_design_overview.htm
42. Lonnie D. Bentley. System Analysis and Design for the Global Enterprise / D. Lonnie. – Columbus : McGraw-Hill Education Ltd, 2007. – 747 p.
43. WeyBackMachine. System Analysis. – Access mode: https://web.archive.org/web/20070822025602/http://pespmc1.vub.ac.be/ASC/SYSTEM_ANALY.html
44. Ritchey T. Analysis and synthesis: on scientific method-based on a study by Bernhard Riemann / T. Ritchey // Systems research. – 1991. – Vol. 8(4). – P. 21–41. DOI: 10.1002/sres.3850080402
45. Booch G. Unified Modeling Language User Guide / G. Booch, J. Rumbaugh, I. Jacobson. – Boston : AddisonWesley. 2005. – 391 p.
46. Iso.org. ISO/IEC 19501:2005 – Information technology – Open Distributed Processing – Unified Modeling Language (UML) Version 1.4.2. – Access mode: https://www.iso.org/standard/32620.html
47. Iso.org. ISO/IEC 19505–1:2012 – Information technology – Object Management Group Unified Modeling Language (OMG UML) – Part 1: Infrastructure. – Access mode: https://www.iso.org/standard/32624.html
48. WeyBackMachine. Basic UML. – Access mode: https://web.archive.org/web/20121214050605/http://ooad.asf.ru/Files/U ML.djvu.zip
49. Chen F. Linguistic tone and non-linguistic pitch imitation in children with autism spectrum disorders: A cross-linguistic investigation / F. Chen, C. C. H. Cheung, G. Peng // Journal of Autism and Developmental Disorders. – 2022. – Vol. 52(5). – P. 2325–2343. DOI: 10.1007/s10803-021-05123-4
50. Robertson A. Emoji skin tone modifiers: Analyzing variation in usage on social media / A. Robertson, W. Magdy, S. Goldwater // ACM Transactions on Social Computing. – 2020. – Vol. 3(2). – P. 1–25. DOI: 10.1145/3377479
51. Tree of endings in Ukrainian language. – Access mode: http://www.senyk.poltava.ua/projects/ukr_stemming/ukr_endings.html
52. Ishihara S. Score-based likelihood ratios for linguistic text evidence with a bag-of-words model / S. Ishihara // Forensic Science International. – 2021. – Vol. 327. – P. 110980. DOI: 10.1016/j.forsciint.2021.110980
53. Understanding the Classification report through sklearn. – Access mode: https://muthu.co/understanding-the-classification-report-in-sklearn/
54. Emotion recognition system project of English newspapers to regional E-business adaptation / [O. Markiv, V. Vysotska, L. Chyrun et al.] // Computer science and information technologies : IEEE 17th International conference, Lviv, Ukraine, 10–12 November 2022 : proceedings. – Lviv: IEEE, 2022. – P. 392–397. https://doi.org/10.1109/CSIT56902.2022.10000527
55. Sentiment analysis technology of English newspapers quotes based on neural network as public opinion influences identification tool / [V. Vysotska, O. Markiv, S. Voloshyn et al.] // Computer science and information technologies : IEEE 17th International conference, Lviv, Ukraine, 10–12 November 2022 : proceedings. – Lviv: IEEE, 2022. – P. 83–88. https://doi.org/10.1109/CSIT56902.2022.10000627

Стаття надійшла до редакції 05.05.2023.
Після доробки 15.08.2023.

DOI 10.15588/1607-3274-2023-3-11

UDC 004.9

SENTIMENT ANALYSIS TECHNOLOGY FOR USER FEEDBACK SUPPORT IN E-COMMERCE SYSTEMS BASED ON MACHINE LEARNING

Tchynetskyi S. – PhD student of Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine.

Polishchuk B. – PhD student of Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine.

Vysotska V. – PhD, Associate Professor of Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine.

ABSTRACT

Context. The interaction between a company and its target audience has been studied for centuries. From the very beginning of commercial relations, the relationship between the service provider and the recipient has been valued almost above all else. Trade is built on trust and respect. The image of an entrepreneur is often more important than the product he sells. For hundreds of years, the relationship between the merchant and the buyer, the entrepreneur and the client has not lost its importance, and in the era of mass digitalization, the quality of the relationship between the company and the target audience of different sizes and professional feedback support with clients often start the success of e-business. To provide these additional tools and information technologies to help businessmen monitor e-business development opportunities in a specific location, as well as establish feedback with users through social networks and mass media. Obtaining such tools will significantly expand the vision of market opportunities for e-business, it will clarify which of them make sense to invest in, and which ones are not worth paying time for. Also see what idea has the future and what business model needs to be implemented/maintained/developed for the rapid development of territorial/interregional e-business. It will also help to understand which levers have the greatest effect for business changes: what not to touch, and what policies to change to ensure high speed in the implementation of the plan based on the analysis of relevant research results, for example, to receive: direct feedback from customers, the dynamics of changes in overall satisfaction or interest of the target audience and advantages/disadvantages from users using NLP analysis; support for the development of e-business in relation to the location of their enterprise and the best directions; – graphs of business development (improvement/deterioration) depending on the content of comments.

Objective of the study is to develop information technology to support the development of e-business by analyzing business locations, processing feedback from users, analyzing and classifying customer feedback in real time from social networks: Twitter, Reddit, Facebook and others using deep learning and Natural methods. Language Processing of Ukrainian-speaking and English-speaking texts.

Method. NLP-methods were used to analyze the opinions of users and customers. Among the methods of implementing the main functions of English-language news classification, the following machine learning methods are used: naive Bayesian classifier, logistic regression, and the method of support vectors. The Naive Bayes algorithm was used to classify Ukrainian-language user feedback, as it performs well on small amounts of data, is easy to train and operate, and works well with text data. Naive Bayes classifier is a very good option for our system and considering that the number of responses in the dataset is smaller compared to the averages.

Results. A machine learning model was developed for the analysis and classification of Ukrainian- and English-language reviews from users of e-commerce systems.

Conclusions. The created model shows excellent classification results on test data. The overall accuracy of the sentimental model for the analysis of Ukrainian-language content is quite satisfactory, 92.3%. The logistic regression method coped best with the task of analyzing the impact of English-language news on the financial market, which showed an accuracy of 75.67%. This is certainly not the desired result, but it is the largest indicator of all considered. The support vector method (SVM) coped somewhat worse with the task, which showed an accuracy of 72.78%, which is a slightly worse result than the one obtained thanks to the logistic regression method. And the naïve Bayesian classifier method did the worst with the task, which achieved an accuracy of 71.13%, which is less than the two previous methods.

KEYWORDS: NLP, text pre-processing, sentiment analysis, feedback, comment, e-commerce, e-business, machine learning, content analysis.

REFERENCES

1. Kuzminov M. Modern Development of Small Business in Ukraine, *Sciences of Europe,* 2022, Vol. 107, pp. 29–31. DOI: 10.5281/zenodo.7479719
2. Definition of customer support. Access mode: https://www.helpscout.com/helpu/definition-of-customer-support.
3. Edvardsson I. R., Durst S., Oskarsson G. K. Strategic outsourcing in SMEs, *Journal of small business and enterprise development,* 2020, Vol. 27(1), pp. 73–84. DOI: 10.1108/JSBED-09-2019-0322
4. Sarkar D. Text analytics with Python: a practitioner's guide to natural language processing. Bangalore, Apress, 2019, 674 p. DOI: 10.1007/978-1-4842-4354-1
5. Eisenstein J. Introduction to natural language processing. Cambridge, MIT press, 2019, 536 p.
6. Goldberg Y. A Primer on Neural Network Models for Natural Language Processing. Access mode: https://jair.org/index.php/jair/article/view/11030/26198
7. Britannica dictionary. Access mode: https://www.britannica.com/topic/outsourcing
8. Sykes. Access mode: https://www.sykes.com
9. Sensee. Access mode: https://www.sensee.co.uk/index.html
10. Serco. Access mode: https://www.serco.com
11. Teleperformance. Access mode: https://www.teleperformance.com/en-us
12. Repustate. Using NLP for business success. Access mode: https://www.repustate.com/blog/using-nlp-for-business-success/
13. Repustate. How can sentiment analysis help you with Patient Voice? Access mode: https://www.repustate.com/patient-voice/
14. SkywellSoftware. How does Siri work: technology and algorithm. Access mode: https://skywell.software/blog/how-does-siri-worktechnology-and-algorithm/
15. Grammarly. How Grammarly uses Natural Language Processing and Machine Learning to identify the main points in a message. Access mode:

DOI 10.15588/1607-3274-2023-3-11

https://www.grammarly.com/blog/engineering/nlp-mlidentify-main-points/
16. Klevu. Smart Search Overview. Access mode: https://www.klevu.com/smart-search/
17. IBM. Natural Language Processing (NLP). What is natural language processing? Access mode: https://www.ibm.com/cloud/learn/natural-language-processing#tocwhat-is-na-jLju4DjE
18. SaS. Natural Language Processing (NLP). What it is and why it matters. Access mode: https://www.sas.com/en_us/insights/analytics/what-is-natural-languageprocessing-nlp.html
19. MonkeyLearn, What is NLP. Access mode: https://monkeylearn.com/blog/what-is-natural-language-processing/
20. Topic Analysis: The Ultimate Guide. Access mode: https://monkeylearn.com/topic-analysis/
21. Sentiment Analysis Explained. Access mode: https://www.lexalytics.com/technology/sentiment-analysis/
22. MonkeyLearn. Intent Classification: How to Identify What Customers Want. Access mode: https://monkeylearn.com/blog/intentclassification/
23. MonkeyLearn. Keyword Extraction. Access mode: https://monkeylearn.com/keyword-extraction/
24. Edia. What is Keyword Extraction? Access mode: https://www.edia.nl/keyword-extraction
25. Stemming vs. Lemmatization in NLP. Access mode: https://towardsdatascience.com/stemming-vslemmatization-in-nlp-dea008600a0
26. Analytics steps. What is Stemming and Lemmatization in NLP?. Access mode: https://www.analyticssteps.com/blogs/what-stemming-andlemmatization-nlp
27. What is Tokenization in NLP? Access mode: https://www.analyticsvidhya.com/blog/2020/05/what-is-tokenizationnlp/
28. Stanford. Machine Translation. Access mode: https://nlp.stanford.edu/projects/mt.shtml
29. Data Science UA. Machine Translation. Access mode: https://data-science-ua.com/wiki/natural-language-processingnlp/machine-translation/
30. Text Summarization in NLP. Access mode: https://www.topcoder.com/thrive/articles/text-summarization-in-nlp
31. What Is Text Summarization in NLP? Access mode: https://www.analyticssteps.com/blogs/what-text-summarization-nlp
32. Dict_uk Github repository. Access mode: https://github.com/brown-uk/dict_uk/tree/master/data
33. Advantages and disadvantages of different classification models. Access mode: https://www.geeksforgeeks.org/advantages-and-disadvantages-of-different-cl assification-models/
34. Naive Bayes Classifier. Access mode: https://www.upgrad.com/blog/naive-bayes-classifier/
35. Coursera. What Is Python Used For? Access mode: https://www.coursera.org/articles/what-is-python-used-fora-beginners-guide-to-using-python
36. Python.org. Executive Summary. Access mode: https://www.python.org/doc/essays/blurb/
37. PymBook. Introduction to Flask. Access mode: https://pymbook.readthedocs.io/en/latest/flask.html
38. FastApi. – Access mode: https://fastapi.tiangolo.com/
39. NLTK. Natural Language Toolkit. Access mode: https://www.nltk.org/
40. AWS. What is Apache Kafka? Access mode: https://aws.amazon.com/ru/msk/what-is-kafka/
41. Tutorialspoint. System Analysis and Design – Overview. Access mode: https://www.tutorialspoint.com/system_analysis_and_design/system_ana lysis_and_design_overview.htm
42. Lonnie D. Bentley. System Analysis and Design for the Global Enterprise. Columbus, McGraw-Hill Education Ltd, 2007, 747 p.
43. WeyBackMachine. System Analysis. Access mode: https://web.archive.org/web/20070822025602/http://pespmc1.vub.ac.be/ASC/SYSTEM_ANALY.html
44. Ritchey T. Analysis and synthesis: on scientific method-based on a study by Bernhard Riemann, Systems research, 1991, Vol. 8(4), pp. 21–41. DOI: 10.1002/sres.3850080402
45. Booch G., Rumbaugh J., Jacobson I. Unified Modeling Language User Guide. Boston, AddisonWesley, 2005, 391 p.
46. Iso.org. ISO/IEC 19501:2005 – Information technology – Open Distributed Processing – Unified Modeling Language (UML) Version 1.4.2. Access mode: https://www.iso.org/standard/32620.html
47. Iso.org. ISO/IEC 19505-1:2012 – Information technology – Object Management Group Unified Modeling Language (OMG UML), Part 1, Infrastructure. Access mode: https://www.iso.org/standard/32624.html
48. WeyBackMachine. Basic UML. Access mode: https://web.archive.org/web/20121214050605/http://ooad.asf.ru/Files/U ML.djvu.zip
49. Chen F., Cheung C. C. H., Peng G. Linguistic tone and non-linguistic pitch imitation in children with autism spectrum disorders: A cross-linguistic investigation, *Journal of Autism and Developmental Disorders*, 2022, Vol. 52(5), pp. 2325–2343. DOI: 10.1007/s10803-021-05123-4
50. Robertson A., Magdy W., Goldwater S. Emoji skin tone modifiers: Analyzing variation in usage on social media, *ACM Transactions on Social Computing*, 2020, Vol. 3(2). – pp. 1–25. DOI: 10.1145/3377479
51. Tree of endings in Ukrainian language. Access mode: http://www.senyk.poltava.ua/projects/ukr_stemming/ukr_endings.html
52. Ishihara S. Score-based likelihood ratios for linguistic text evidence with a bag-of-words model, *Forensic Science International,* 2021, Vol. 327, P. 110980. DOI: 10.1016/j.forsciint.2021.110980
53. Understanding the Classification report through sklearn. – Access mode: https://muthu.co/understanding-the-classification-report-in-sklearn/
54. Markiv O., Vysotska V., Chyrun L., Voloshyn S., Dyyak I., Panasyuk V. Emotion recognition system project of English newspapers to regional E-business adaptation, *Computer science and information technologies, IEEE 17th International conference,* Lviv, Ukraine, 10–12 November 2022, proceedings. Lviv, IEEE, 2022, pp. 392–397. https://doi.org/10.1109/CSIT56902.2022.10000527
55. Vysotska V., Markiv O., Voloshyn S., Dyyak I., Budz I., Schuchmann V. Sentiment analysis technology of English newspapers quotes based on neural network as public opinion influences identification tool, *Computer science and information technologies, IEEE 17th International conference, Lviv, Ukraine, 10–12 November 2022, proceedings*. Lviv, IEEE, 2022, pp. 83–88. https://doi.org/10.1109/CSIT56902.2022.10000627

DOI 10.15588/1607-3274-2023-3-11