# Кібернетика та комп'ютерні технології

УДК 519.67 DOI:10.34229/2707-451X.24.4.8

В.В. ТРЕТИНИК, М.В. ПІНДА

## АНАЛІТИЧНА СИСТЕМА ДЛЯ ВИЗНАЧЕННЯ СТАВЛЕННЯ СТУДЕНТІВ ДО УНІВЕРСИТЕТУ

*Побудовано компонентну модель системи для визначення ставлення студентів до університету. Зібрано відгуки студентів з Telegram-каналів. Проведено сентимент-аналіз, статистичний аналіз даних, аналіз часових рядів, кластерний аналіз. Розроблена система дозволяє автоматизовано отримувати звіт про ставлення студентів до університету на основі запропонованих методів.*

***Ключові слова:*** *аналітична система, відгуки студентів, система для визначення ставлення студентів до університету, аналіз тональності тексту, статистичний аналіз, аналіз часових рядів, кластерний аналіз.*



**Вступ**. В умовах швидкого розвитку вищої освіти та зростаючої конкуренції між навчальними закладами розуміння ставлення студентів до університету стає критично важливим для покращення якості освітніх послуг. Відгуки студентів є цінним джерелом інформації для оцінки ефективності навчального процесу, адміністративних послуг та загальної атмосфери в університеті. Однак традиційні методи збору та аналізу відгуків часто є не автоматизованими, що потребує значних часових і людських ресурсів для обробки великих обсягів текстових даних. Наявні програмні рішення використовують методи обробки та аналізу тональності тексту на основі методів та алгоритмів машинного навчання (наївного байєсового класифікатора, методу опорних векторів, логістичної регресії), а також глибинного навчання (рекурентних нейронних мереж). При цьому більшість наявних програмних рішень не є безкоштовними, що ускладнює їх широке використання в університетах, особливо з обмеженими фінансовими ресурсами. Тому існує потреба в розробці нових рішень, які не лише будуть доступними для використання, але й забезпечать високу точність та ефективність при обробці текстових відгуків.

Мета даної роботи – це розробка аналітичної системи, її математичного та програмного забезпечення для визначення ставлення студентів до університету на основі їх текстових відгуків. Розроблена система повинна забезпечити високу точність та ефективність у роботі з текстовими даними, автоматизуючи процес аналізу відгуків і мінімізуючи затрати людських ресурсів на обробку інформації.

Запропонована компонентна модель системи для визначення ставлення студентів до університету відображає структуру та відношення між елементами через реалізацію відповідних інтерфейсів (рис. 1).

Компонент «Сховище даних» – це сховище, в якому зберігаються набори даних, потрібні для розв'язання задачі визначення ставлення студентів до університету за їхніми відгуками. Збір відгуків можна здійснювати на різних платформах та за допомогою різних інструментів. У даній роботі відгуки студентів збираються з Telegram-каналів, які забезпечують конфіденційність

та не публікують авторів. Зібрані дані зберігаються у форматі .csv.

Компонент «Попередня обробка» призначений для видалення зайвої для розв'язання задачі інформації з даних. У ньому виконуються наступні кроки: видалення нульових значень; перетворення у нижній регістр; видалення дублікатів.


РИС. 1. Модель системи для визначення ставлення студентів до університету за їхніми відгуками. Діаграма компонентів у нотації UML (рисунки належать до статті В.В.Третиник, М.В. Пінди збірника № 4, 2024)

Компонент «Обробка текстів відгуків» призначений для обробки текстових даних за методикою NLP (Natural Language Processing) та визначення емоційного забарвлення відгуків. NLP – це один з напрямів штучного інтелекту, який працює з аналізом людської мови. Використовується популярна бібліотека Python NLTK (Natural Language Toolkit) для всіх етапів обробки текстів: токенізації (розбиття тексту на окремі слова-токени); видалення стоп-слів, тобто слів, які не містять лексично важливої інформації; стемінгу (обрізки слова до кореня). Після наведених кроків проводиться аналіз тональності тексту за допомогою мовної моделі BERT (Bidirectional Representation for Transformers). Основна концепція BERT полягає у тому, щоб аналізувати контекст з обох боків слова, що дозволяє точніше вловлювати його повне значення в конкретному середовищі. Це дає змогу моделі ефективно працювати з неоднозначними та складними мовними явищами.

Компонент «Модель прогнозування» призначений для прогнозування ставлення студентів до університету з плином часу. Для цього використовуються статистичні методи, аналіз часових рядів, методи машинного навчання. Визначаються тренди та сезонність, а також досліджуються класи схожих між собою відгуків.

Компонент «Платформа-застосунок» призначений для генерації звіту, який відображає виявлені загальні тенденції при визначенні ставлення студентів до університету на основі відгуків.

Інтерфейс IDP (Interface Data Processing) – інтерфейс передачі завантажених в оперативну пам'ять сирих наборів даних на вхід процесу попередньої обробки.

Інтерфейс ITP (Interface Text Processing) – інтерфейс передачі вихідних (із компонента попередньої обробки) опрацьованих, очищених даних на вхід процесу обробки текстів відгуків.

Інтерфейс IPP (Interface Prediction Process) – інтерфейс передачі вихідних (із компонента обробки текстів відгуків) опрацьованих даних на вхід процесу прогнозування.

Інтерфейс IAD (Interface Attitude Determination) – інтерфейс передачі результатів роботи системи на вхід процесу виведення результатів для генерації звіту, який відображає виявлені тенденції.

Для навчання використовуються відгуки про столичний ВНЗ. Дані зібрано з Telegram-каналів, де публікуються хороші та погані відгуки, та збережено у форматі .csv. Оскільки відгуки написані двома мовами (українською та російською), відповідні тексти перекладаються англійською за допомогою перекладача Google. Створюється спільний набір даних з текстами відгуків та відповідними датами їх публікації. Необроблені набори даних містять 660 хороших та 7231 поганих відгуків, отже сумарно 7891 записів. Після видалення дублікатів, зайвих пробілів та порожніх рядків кількість записів зменшилася з 7891 до 7395. Тексти переводяться до нижнього регістру. Видаляються смайли та інші символи, що не є текстом.

Для визначення тональності тексту використовується архітектура BERT BASE та фреймворк Torch, у якому ця версія реалізована. Архітектура BERT BASE базується на трансформерній моделі, яка складається з багатошарових блоків енкодерів, а саме: 12 енкодерних шарів, 12 голів уваги, прихований розмір: 768, кількість параметрів: 110 мільйонів. Набір даних розбивається на тренувальну та тестову вибірки у співвідношенні 4:1. Токенізатор BERT попередньо перетворює тексти на числові дані, прийнятні для подачі на вхід моделі. Він розбиває складені слова на прості, додаючи спеціальні мітки, що отримані слова є частинами довшого слова. Також додаються спеціальні токени: [CLS] – використовується в задачах класифікації; [SEP] – використовується як роздільник для позначення меж між різними реченнями чи частинами тексту; [MASK] – використовується для приховування деяких токенів під час попереднього навчання моделі.

На виході токенізатор BERT видає числові ідентифікатори токенів словника ("input_ids"), ідентифікатори, які вказують до якого речення або частини тексту належить кожен токен ("token_type_ids") та мітки, що повідомляють моделі, на які токени потрібно звернути увагу ("attention_mask"). Після проведення токенізації ініціалізується модель BERT та починається навчання. Налаштовується оптимізатор, який використовується для коригування параметрів моделі з метою мінімізації функції втрат. Функція втрат показує, наскільки добре модель передбачає цільові значення. Оскільки набір даних є незбалансованим, для оцінки точності моделі використовується метрика F1-score, яка не залежить від співвідношення класів. Модель навчається на тренувальних даних протягом 10 епох, однак додається умова дострокової зупинки навчання при виявленні перенавчання. Якщо протягом трьох епох втрата (loss) на тестовому наборі даних не покращиться, навчання моделі зупиняється. Навчання завершилося достроково, оскільки протягом перших чотирьох епох втрата (loss) на тестовому наборі не покращувалася. Найкраще значення метрики F1-score виявилося після другої епохи, а саме 0.96. Результати тренування моделі зберігаються в окремому файлі. Після цього модель з найкращими показниками завантажується з файлу, та за її допомогою можна визначати тональність тексту.

Отже, негативних відгуків за той самий період виявилось у приблизно 11 разів більше, ніж позитивних (Telegram-канали було створено одночасно). Для визначення, які відгуки є більш розгорнутими, обчислюється середня кількість символів у відповідних текстах (рис. 2).

Середня довжина позитивних відгуків вийшла більшою: 594.27 символів проти 504.74. Позитивні відгуки часто містять розгорнуті пояснення, що саме сподобалося. Люди схильні конкретизувати позитивні аспекти. Негативні відгуки навпаки часто можуть бути стислими та емоційними, містити короткі, різкі висловлювання про конкретні недоліки.

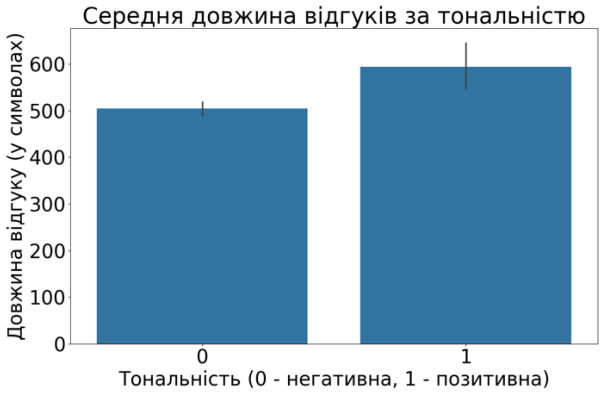


РИС. 2. Середня довжина текстів відгуків

Далі відгуки відсортовуються та групуються: обчислюється кількість позитивних та негативних відгуків на кожну дату. Пропущені дати заповнюються нулями для фіксації, що в ці дні опублікованих відгуків не було. Створюються два окремі часові ряди (з позитивними та негативними відгуками відповідно), та виводиться їх загальна статистика [1]. Для позитивних / негативних відгуків:

- кількість ненульових значень: 2461 / 2461;
- середнє арифметичне: 0.2471 / 2.7578;
- стандартне відхилення: 1.3537 / 6.6229;
- мінімальне значення: 0 / 0;
- 25-й процентиль: 0 / 0;
- 50-й процентиль (медіана): 0 / 0;
- 75-й процентиль: 0 / 2;
- максимальне значення: 39 / 114.

Отже, більша частина днів у наборі даних (75 %) не має жодного позитивного відгуку, а в середньому публікується менше одного відгуку на день. Це свідчить про те, що позитивні відгуки з'являються рідко, проте іноді бувають аномальні сплески (до 39 відгуків на день). Хоча середня кількість негативних відгуків на день більша, ніж позитивних (майже 3 проти 0.25), половина днів (50 %) не містить жодного негативного відгуку. Проте в дні, коли негативні відгуки все ж є, їх кількість може дуже варіюватися, досягаючи до 114 відгуків на день. З огляду на те, що відгуки були зібрані з Telegram-каналів, слід враховувати, що їх збирає та публікує одна людина (або, в окремих випадках, невелика група людей). Тобто присутня деяка похибка між датою написання відгуку студентом та його публікацією. Загалом, негативні відгуки публікуються частіше, ніж позитивні, що легко пояснюється з точки зору психології: негативні події мають більший вплив на наше емоційне благополуччя, ніж позитивні, що може спонукати до більшої активності у вираженні негативних думок.

Коефіцієнт кореляції Пірсона, який відображає лінійну залежність між позитивними та негативними відгуками, дорівнює 0.18. Це вказує на те, що між кількістю позитивних і негативних відгуків майже відсутній лінійний зв'язок – зміна одного параметра майже не впливає на зміну іншого.

Далі послідовність відгуків розглядається як часовий ряд. Часові ряди з негативними та позитивними відгуками перевіряються на стаціонарність (чи зберігаються статистичні характеристики незмінними у часі). Ця властивість допомагає в прогнозуванні майбутніх значень. У даній роботі застосовується тест Дікі – Фуллера. Якщо *p*-значення (імовірність неправильності нульової гіпотези) є меншим за визначену константу (зазвичай, 0.05), то нульова гіпотеза про нестаціонарність ряду відхиляється, і ряд вважається стаціонарним. У нашому випадку для ряду з позитивними відгуками

*p*-value = 0.0, для ряду з негативними відгуками *p*-value = 0.00015. Отже, обидва часові ряди можна вважати стаціонарними.

Часові ряди мають три основні компоненти: тренд, сезонність та шум. Трендом називають довгострокову зміну даних. Тренд може бути висхідним (зріст), низхідним (падіння) або горизонтальним (змін не фіксується). Негативні відгуки демонструють пік близько 2020 року, але їх кількість знижується після 2021 року. Загалом, спостерігається початкове зростання, яке поступово змінюється на спад. Позитивні відгуки з'являються більш епізодично, проте присутня загальна тенденція до зниження їх кількості в останні роки. Загалом з обох графіків видно, що активність публікації відгуків знизилась протягом останніх двох років (рис. 3). Такий спад можна також асоціювати з початком повномасштабного російського вторгнення в Україну. Війна викликає значний емоційний стрес у студентів, що може зменшити їхню здатність або бажання ділитися думками про навчання.

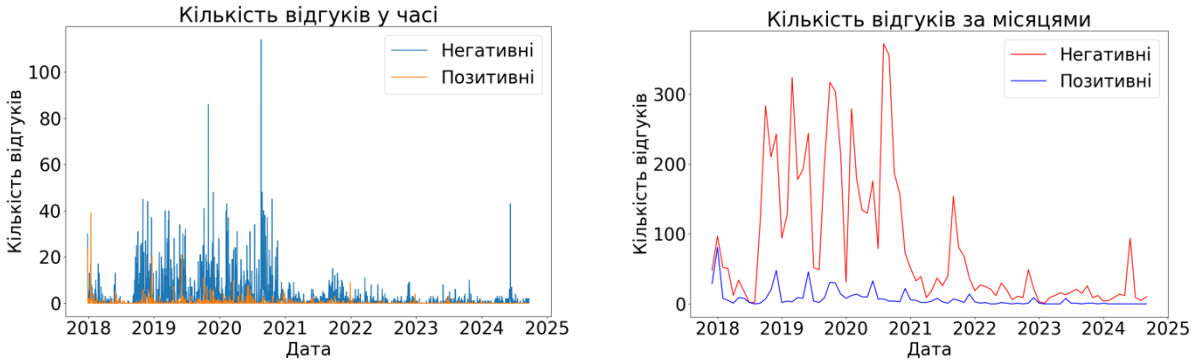


РИС. 3. Кількість позитивних та негативних відгуків у часі

Сезонність може вказати на події (зовнішні фактори), які регулярно впливають на публікації. З графіка негативних відгуків видно, що піки стабільно спостерігаються кожного вересня (крім останніх років, коли загальна активність знизилась), що явно асоціюється з початком навчального року. Піки позитивних відгуків частіше зустрічаються під кінець календарного року, а також у червні, що означає кінець навчального семестру. Перед початком канікул студенти зазвичай публікують відгуки про викладачів, серед яких нерідко зустрічаються позитивні враження.

На графіках видно значні коливання без чіткої періодичності, які можна вважати шумом – випадковими змінами в даних, що не пояснюється трендом або сезонністю. Коливання в кількості позитивних відгуків є менш вираженими та більш стабільними.

Автокореляція та часткова автокореляція допомагають визначити, які лаги є суттєвими, структуру і залежності в даних (рис. 4).

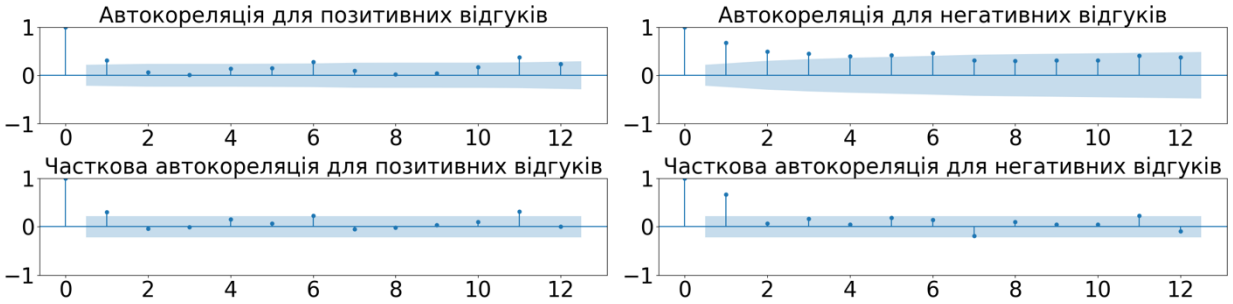


РИС. 4. Автокореляція та часткова автокореляція для часових рядів, агрегованих за місяцями, з лагом 12

На графіку автокореляції для позитивних відгуків спостерігається пік на лагу 1, що вказує на сильну залежність між сусідніми місяцями. Кореляція падає до нуля, а сплески на лагах 6 і 11 свідчать про сезонність (семестр і рік). На графіку часткової автокореляції також помітні ці тенденції. Для негативних відгуків автокореляція поступово зменшується, з найбільш значимими лагами 1 і 2. Часткова автокореляція показує високий зв'язок лише на першому лагу, після чого вплив інших лагів незначний. Отже, наявна тільки короткотривала залежність.

Кількість майбутніх відгуків прогнозується за допомогою моделі ARIMA [2] на місячних даних (рис. 5). Оцінка якості моделі на даних позитивних відгуків вказує на високу точність прогнозу: MSE = 0.556, MAE = 0.689. Прогнозування негативних відгуків виявило більші відхилення через мінливість даних, тому застосовується SARIMA з сезонністю 2 місяці, яка дає дещо кращі результати: MSE = 546.946, MAE = 11.319.

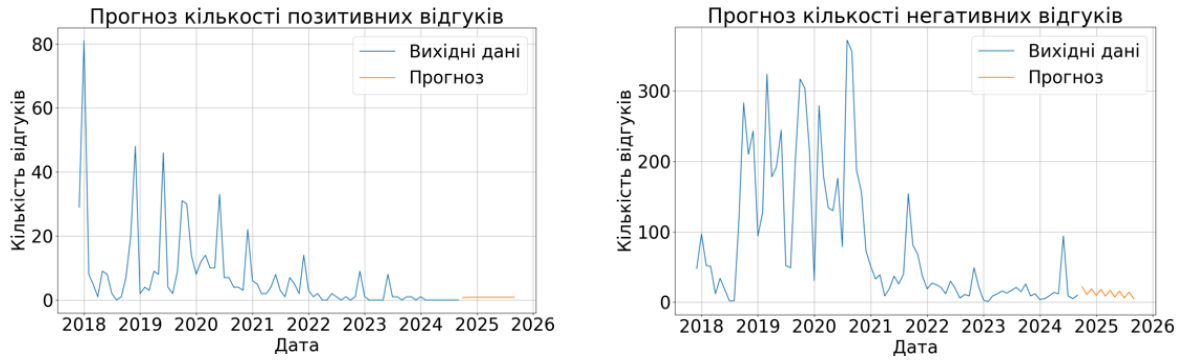


РИС. 5. Прогноз кількості відгуків

Перед проведенням кластеризації тексти відгуків обробляються за методикою NLP та векторизуються за допомогою методу TF-IDF. Кластеризація виконується за алгоритмом K-Means [3], а оптимальна кількість кластерів обирається за методом «ліктя» (рис. 6).

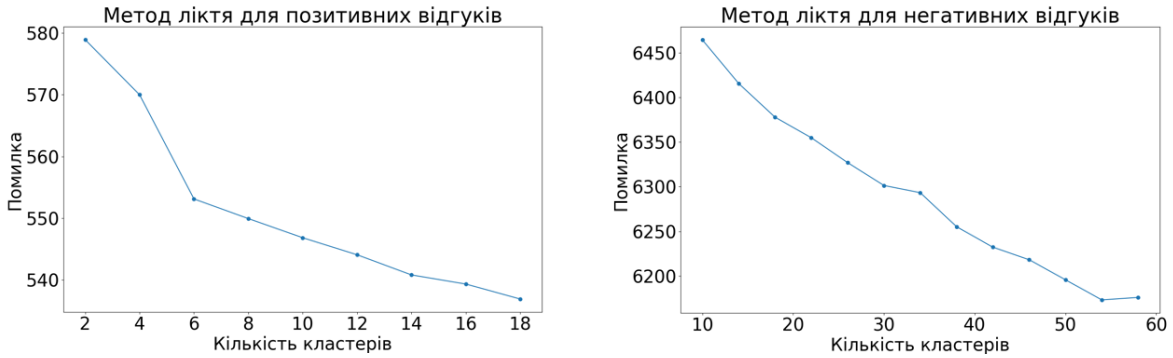


РИС. 6. Визначення оптимальної кількості кластерів

Алгоритм K-Means – це один з найбільш поширених і простих методів кластеризації [3]. Його основна ідея полягає у мінімізації суми квадратів відстаней між кожною точкою і центроїдом кластера, до якого вона належить. Таким чином, об'єкти одного кластера є максимально схожими між собою, але відрізняються від об'єктів з інших кластерів. Ідея методу «ліктя» полягає у виборі кількості кластерів, при якій подальше збільшення кількості кластерів майже не знижує помилку клас-

теризації. Будується графік суми квадратів помилок для різної кількості кластерів, і оптимальна кількість кластерів обирається у точці, де графік змінюється найменш різко (схоже на «лікоть»). Отже, оптимально обрати кількість кластерів для позитивних відгуків 6, а для негативних – 38. Якість кластеризації оцінюється за допомогою коефіцієнту силуету, який показує, наскільки добре схожі зразки об'єднуються в один кластер. Середній коефіцієнт силуету для позитивних відгуків дорівнює 0.028, а для негативних дорівнює 0.013. Це означає, що кластеризація наявних даних вийшла не достатньо чіткою, та багато об'єктів знаходяться на межі між двома кластерами.

Наприкінці проводиться аналіз частоти слів у відгуках всередині кластерів. Для прикладу обирається кластер під номером 1 з позитивними відгуками. Виводяться основні слова у кластері та їх частота у відгуках. Як видно, у даний кластер потрапили відгуки про гарних викладачів (рис. 7).


РИС. 7. Частота слів у кластері

**Висновок**

У роботі розроблено компонентну модель, математичне та програмне забезпечення системи для визначення ставлення студентів до університету. Розроблена система може бути використана для автоматизованого аналізу відгуків студентів. Для цього застосовуються наступні методи та алгоритми: аналіз тональності тексту, статистичний аналіз, аналіз часових рядів, кластерний аналіз. Можливим напрямком модифікації системи є застосування методів аналізу до україномовних текстів без їх попереднього перекладу англійською мовою, що дозволить зберегти тонкощі лексики та зменшить імовірність втратити сенс.

**Список літератури**

1. Yule G. An introduction to the theory of statistics. George Udny Yule. London: C. Griffin and company, limited, 1911. 376 p. https://doi.org/10.1037/13786-000
2. Box G., Jenkins G., Reinsel G., Ljung G. Time Series Analysis: Forecasting and Control, 5th Edition. Hoboken, New Jersey: John Wiley and Sons Inc. 2015. 712 p.
3. Bishop C. Pattern Recognition and Machine Learning. Christopher M. Bishop. Singapore: Springer Science+Business Media, LLC, 2006. 738 p.

Одержано 09.10.2024

**Третиник Віолета Вікентіївна,**
кандидат фізико-математичних наук,
доцент кафедри прикладної математики НТУУ «КПІ ім. Ігоря Сікорського»,
https://orcid.org/0000-0002-3538-8207
viola.tret@gmail.com

**Пінда Марія Володимирівна,**
магістрант кафедри прикладної математики НТУУ «КПІ ім. Ігоря Сікорського».

УДК 519.67

**В.В. Третиник [*], М.В. Пінда**

## Аналітична система для визначення ставлення студентів до університету

*НТУУ «КПІ ім. Ігоря Сікорського»*
*[*] Листування: viola.tret@gmail.com*

**Вступ**. В умовах швидкого розвитку вищої освіти та зростаючої конкуренції між навчальними закладами розуміння ставлення студентів до університету стає критично важливим для покращення якості освітніх послуг. Відгуки студентів – це цінне джерело інформації для оцінки ефективності навчального процесу, адміністративних послуг та загальної атмосфери в університеті. Однак традиційні методи збору та аналізу відгуків часто є не автоматизованими, що потребує значних часових і людських ресурсів для обробки великих обсягів текстових даних. Наявні програмні рішення використовують методи обробки та аналізу тональності тексту на основі методів та алгоритмів машинного навчання (наївного байєсового класифікатора, методу опорних векторів, логістичної регресії), а також глибинного навчання (рекурентних нейронних мереж).

При цьому більшість наявних програмних рішень не є безкоштовними, що ускладнює їх широке використання в університетах, особливо з обмеженими фінансовими ресурсами. Тому існує потреба в розробці нових рішень, які не лише будуть доступними для використання, але й забезпечать високу точність та ефективність при обробці текстових відгуків.

**Мета роботи.** Розробка аналітичної системи, її математичного та програмного забезпечення для визначення ставлення студентів до університету на основі їх текстових відгуків. Розроблена система повинна забезпечити високу точність та ефективність у роботі з текстовими даними, автоматизуючи процес аналізу відгуків і мінімізуючи затрати людських ресурсів на обробку інформації.

**Результати.** Побудовано компонентну модель системи для визначення ставлення студентів до університету. Зібрано відгуки студентів з Telegram-каналів. Проведено сентимент-аналіз, статистичний аналіз даних, аналіз часових рядів, кластерний аналіз. Розроблена система дозволяє автоматизовано отримувати звіт про ставлення студентів до університету на основі запропонованих методів. Виконано програмну реалізацію системи мовою програмування Python.

**Ключові слова:** аналітична система, відгуки студентів, система для визначення ставлення студентів до університету, аналіз тональності тексту, статистичний аналіз, аналіз часових рядів, кластерний аналіз.

UDC 519.67

**Violeta Tretynyk [*], Mariia Pinda**

## The Analytical System for Determining the Attitude of Students to the University

*The National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"*
*[*] Correspondence: viola.tret@gmail.com*

**Introduction.** In the context of the rapid development of higher education and growing competition between educational institutions, understanding students' attitude towards the university becomes critical to improving the quality of educational services. Student feedback is a valuable source of information for assessing the effectiveness of the educational process, administrative services, and the general atmosphere at the university. However, traditional methods of collecting and analyzing feedback are often not automated, requiring significant time and human resources to process large amounts of text data. Existing software solutions use methods for processing and analyzing text tone based on machine learning methods and algorithms (naive Bayesian classifier, support vector machine, logistic regression), as well as deep learning (recurrent neural networks).

At the same time, most of the available software solutions are not free, which makes it difficult to use them widely in universities, especially those with limited financial resources. Therefore, there is a need to develop new solutions that will not only be available for use, but also provide high accuracy and efficiency in processing text reviews.

**The purpose of the article**. The article is aimed at developing an analytical system, its mathematical and software tools for determining students' attitude towards the university based on their textual feedback. The developed system should provide high accuracy and efficiency in working with text data, automating the process of analyzing reviews and minimizing human resources for information processing.

**Results**. A component model of the system for determining the attitude of students to the university was built. Student feedback from Telegram channels was collected. Sentiment analysis, statistical data analysis, time series analysis, and cluster analysis were conducted. The developed system allows to automatically receive a report on students' attitude towards the university based on the proposed methods. The software implementation of the system in the Python programming language has been carried out.

**Keywords:** analytical system, student feedback, system for determining students' attitude towards the university, text tone analysis, statistical data analysis, time series analysis, cluster analysis.