DOI: 10.36910/6775-2524-0560-2020-40-21
УДК: 004.4 + 004.6
Мироненко Сергій Сергійович, аспірант
https://orcid.org/0000-0001-7561-8305
Онищенко Єлизавета Андріївна, студент
https://orcid.org/0000-0001-6502-0406
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна

ПОРІВНЯЛЬНИЙ АНАЛІЗ МЕТОДІВ ДЛЯ ВИРІШЕННЯ ЗАДАЧІ СЕНТИМЕНТ АНАЛІЗУ ТЕКСТУ

Мироненко С. С., Онищенко Є. А. Порівняльний аналіз методів для вирішення задачі сентимент аналізу тексту. В даній статті розглядається підхід навчання з вчителем (supervised learning) для вирішення проблеми, пов'язаної з Natural Language Processing (NLP), а саме сентимент-аналіз текстових даних. В ході роботи було реалізовано 4 різних класифікатори на одній й тій самій виборці даних та порівняно їх ефективність за часом навчання, тестування та точності класифікації. В результаті роботи було визначено, що найкращий метод серед реалізованих – 3D CNN модель, яка використовує BERT токенізатор для попередньої обробки тексту. Саме завдяки використанню BERT для препроцессінгу тексту цей метод показав кращі результати.
Ключові слова: класифікація тексту, сентимент-аналіз, машинне навчання, Natural Language Processing, поглиблене навчання.

Мироненко С. С., Онищенко Е. А. Сравнительный анализ методов для решения задачи сентимент анализа текста. В данной статье рассматривается подход обучения с учителем (supervised learning) для решения проблемы, связанной с Natural Language Processing (NLP), а именно сентимент-анализ текстовых данных. В ходе работы было реализовано 4 различных классификатора на одной и той же выборке данных и сопоставленна их эффективность по времени обучения, тестирования и точности классификации. В результате работы было определено, что лучший метод среди реализованных - 3D CNN модель, которая использует BERT токенизатор для предварительной обработки текста. Именно благодаря использованию BERT для препроцессинга текста этот метод показал лучшие результаты.
Ключевые слова: классификация текста, сентимент-анализ, машинное обучение, Natural Language Processing, углубленное обучение.

Myronenko Serhii, Onyshchenko Yelyzaveta. Comparative analysis of methods for solving the problem of sentiment text analysis. This article discusses a supervised learning approach for sentiment analysis of text data, using NLP. The research was conducted, using four methods on the same dataset and their efficiency was compared by using the following characteristics: training time, testing time and accuracy of classification. A 3D CNN model using BERT Tokenizer for text preprocessing was selected as the best method of this comparative study, because of its text preprocessing algorithm.
Keywords: Text classification, Sentiment analysis, Machine learning, Natural Language Processing, Deep learning.

Постановка наукової проблеми. Під час прийняття рішень про купівлю певного продукту, споживач збирає та аналізує доступну для нього інформацію таку як поради консультантів в магазині або за відгуками друзів, які вже мають досвід у використанні цього продукту. А з розвитком інформаційних технологій та Інтернету можна легко прочитати сотні або навіть тисячі відгуків про товар або послугу, не виходячи з дому. Люди все більше діляться своїми думками, ставленнями, оцінками в соціальних мережах, блогах, форумах, в той самий час як інші користувачі читають ці відгуки та формують свою думку про продукт або послугу. В зв'язку з цим, менеджерам продуктів, маркетологам потрібно постійно відстежувати статистику товару. З метою економії часу та для ефективної роботи менеджерам спочатку потрібно швидко класифікувати відгук за настроєм для подальшого аналізу. Та з кожним днем кількість загальнодоступної інформації лише зростає і її неможливо обробити вручну, тому не дивно що аналіз людських емоцій щодо певної проблеми є перспективною областю дослідження, яка називається сентимент-аналіз.

Аналіз тональності тексту або як варіант сентимент-аналіз це, у першу чергу, процес виділення та встановлення емоційного забарвлення тексту, його категорійна приналежність [1]. Найчастіше, думка автора тексту щодо деякої теми, послуги, події, тощо відноситься до однієї з наступних категорій: позитивна або негативна. Також можливо додавання ще одної категорії – нейтральної, що ускладнює задачу. В цій роботі буде розглянуто лише бінарну проблему класифікації. Ця галузь стала активно розвиватися з 2000-х років з декількох причин. Перша причина – поява великої кількості комерційних додатків, що є великою мотивацією для досліджень. Друга – подальше вивчення цієї області пропонує велику кількість дослідницьких проблем, які раніше не розглядалися [1].

Отже, дослідження в напрямку сентимент – аналізу впливає не тільки на Natural Language Processing, що вивчає можливості розшифрування та осмислення людських мов, але й на політичні, соціальні та управлінські науки, бо вони всі певним чином залежать від людських емоцій та оцінок.

Аналіз досліджень. Наукові дослідження у вирішенні задачі сентимент аналізу тексту є доволі ґрунтовними. Так Н. М. Кобець та Т. В. Ковалюк [2] дослідили задачі аналізу тональності тексту як фундаментальна основа аналізу думок користувачів соціальних мереж. У науковій роботі здійснено дослідження методів, задач та математичної моделі контент-аналізу. Здійснено опис методу аналізу повідомлень та думок, а також алгоритм категоризації аспектів та визначення рейтингу думок.

Н. І. Мазниченко [3] наводить аналіз можливості застосування сучасних програмних засобів обчислювальної техніки для змістовного аналізу текстів електронних документів. Автором розглянуті можливі напрямки застосування відповідних комп'ютерних програм.

Методи автоматичного аналізу тональності контенту у соціальних мережах для виявлення інформаційно-психологічних впливів розкрили Д. В. Шингалов, Є. В. Мелешко, Р. М. Минайленко, В. А. Резніченко [4]. Дослідження включає в себе огляд лексемного методу та методів машинного навчання, окремо розглянуто процедуру попередньої обробки тексту перед аналізом та категорії словників, що використовуються для аналізу.

Н. Б. Шаховська, Х. Ю. Гірак [5] запропонували унікальну шкалу слів, які є емоційно забарвленими, описана шкала охоплює ранжування слів, також здійснюється визначення коефіцієнта важливості за допомогою методики Фішберна. Кожна окрема методика відрізняється коефіцієнтами, нормами, використанням математичних шкал, які побудовано на логарифмічній основі, однак, варто підкреслити, що завданням методик є визначення порядку слів, фраз без глибинного аналізу їхньої тональності, емоційного забарвлення і відношення між ними. За результатами проведеного дослідження авторами здійснено розробку платформи для розрахункової оцінки, яка, в свою чергу є інтегрованою, яка дасть змогу визначити думку користувача, автора тощо.

Проте, незважаючи на масштабність наукових досліджень проведення порівняльного аналізу існуючих методів аналізу тональності тексту є актуальним питанням та потребує детального опрацювання.

Мета роботи. Метою даної роботи є проведення порівняльного аналізу існуючих методів аналізу тональності тексту.

Виклад основного матеріалу й обґрунтування отриманих результатів. На сьогодні, виділяють два основні класи алгоритмів машинного навчання - це навчання з учителем (supervised learning) і навчання без вчителя (unsupervised learning). Варно зазначити, що крім цих класів виділяють також алгоритми навчання з підкріпленням (reinforcement learning) і рекомендаційні системи (recommender systems). У рамках дослідження у межах даної роботи було обрано методи які відносяться до підходу supervised learning. Алгоритм навчання з вчителем приймає марковані дані і створює модель, яка виконує передбачення, надаючи нові дані. Це можуть бути як завдання класифікації, так і завдання регресії.

На основі цих даних відбуваються процеси навчання моделі (побудова такого алгоритму, який для будь-якого об'єкту класифікації може дати найбільш точну відповідь) та її тестування.

Основою аналізу обрано датасет Large Movie Review Dataset [6]. Даний датасет включає в себе 50 000 відгуків про фільми. Окреслений набір відгуків про фільми використовується для бінарного сентимент-аналізу та містить 25000 відгуків для навчання, а також відповідну кількість для тестування. Варто зазначити, що ця вибірка є збалансованою: половина даних є позитивними відгуками, а інша – негативними.

В умовах реалізації дослідження здійснено впровадження певних класифікаторів, до складу яких варто віднести: Long short-time memory (LSTM), Naïve Bayes Classifier, Convolutional Neural Network (CNN).

Naïve Bayes Classifier (наївний байесовский класифікатор (НБК)) – один з найбільш примітивних класифікаторів, заснованих на теоремі Байеса з умовою виконання суворої незалежності імовірнісних компонент. Дане припущення розглядає кожне слово в тексті окремо і незалежно від інших. Внаслідок цього, процес навчання НБК складний і нетривіальний в силу витрати великої кількості ресурсів для отримання мінімальної бази, придатної для подальшої класифікації тексту.

З математичного аспекту, наївний байесовский класифікатор – це певна структурна модель машинного навчання, яка базується на теоремі Байєса:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)},$$

де P(A) – ймовірність гіпотези A, P(B|A) – ймовірність того, що подія B відбудеться при істинності гіпотези A, P(A|B) – ймовірність гіпотези A при умові, що подія B відбудеться, P(B) – повна ймовірність того, що подія B настане [7].

Long short–time memory (LSTM) є різновидом рекурентних нейронних мереж (РНМ), яка здатна вивчати довготривалі залежності [8].

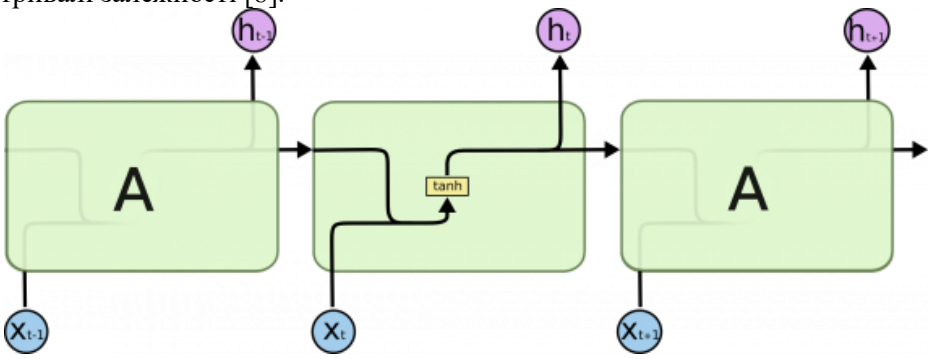

Рис.1 Архітектура LSTM з одношаровим повторюваним модулем

Загальні конструкції рекурентних нейронних мереж мають форму відмінного від аналогів ланцюжка модулів нейронної мережі, що повторюються з певною динамічною моделлю. У стандартних РНС цей повторюваний модуль має просту структуру, наприклад, один шар повторюваного модуля. Архітектура LSTM з одношаровим повторюваним модулем наведена на рисунку 1.

Ключовим поняттям LSTM є стан осередку: горизонтальна лінія, що проходить через верхню частину діаграми. Стан осередку нагадує конвеєрну стрічку. Яка проходить через весь ланцюжок, піддаючись незначним лінійним перетворенням. У LSTM кількість інформації в стані осередку є не стабільною та постійно змінюється, в залежності від потреб. Для цього використовуються певні структури, які ретельно настроюються, окреслені структури мають назву гейт. Гейт – це «ворота», які дають можливість пройти інформації, або навпаки не пропускають останню. Гейти складаються з сигмовидного шару нейронної мережі і операції поточного множення.

На виході сигмовидного шару наведено низку чисел, формат яких походить від нуля до одиниці, даний формат дозволяє визначити, яка кількість відсотків кожної окремої заявленої одиниці інформації пропустити далі. Значення «0» означає «не пропустити нічого», значення «1» – «пропустити все».

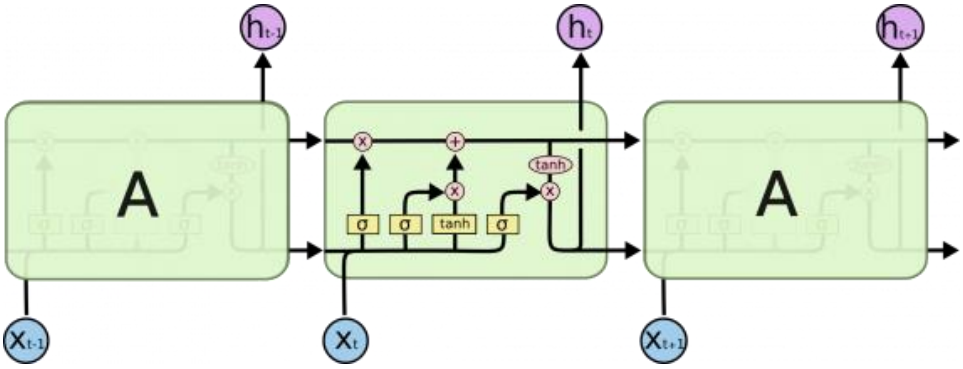

Рис.2 Архітектура LSTM, що складається з чотирьох шарів, які взаємодіють один з одним

Архітектура LSTM, що складається з чотирьох шарів, які взаємодіють один з одним наведена на рисунку 2, на наведеній діаграмі кожна лінія є вектором. Рожеве коло означає поточну операцію, наприклад, підсумовування векторів. Під жовтими осередками розуміються шари нейронної мережі. Поєднання ліній є об'єднанням векторів, а знак розгалуження – копіювання вектора з подальшим зберіганням в різних місцях.

Convolutional Neural Networks (CNN) – архітектура штучних нейронних мереж, яка базується на чергуванні шарів згортки та шарів-агрегаторів [9].

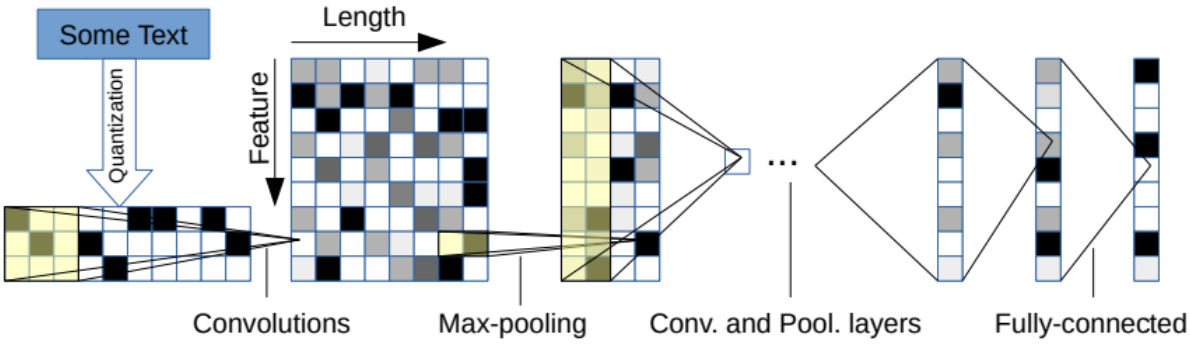

Рис. 3 Архітектура згорткової нейронної мережі

Згорткові нейронні мережі [2], спочатку не призначалися для роботи з текстом, вони використовувалися в «комп'ютерному зорі» і розпізнаванні образів. Згорткова нейронна мережа – це особливий вид нейронних мереж прямого поширення (рис. 3). Під прямим поширенням розуміється те, що поширення сигналів по нейронах йде по порядку, від першого шару до останнього. Прихованих шарів в мережі може бути досить багато, все залежить від кількості даних і складності завдання.

Основною особливістю таких мереж є наявність шарів, що чергуються по типу «згортка – субдискретизація», яких може бути безліч. Операція згортки (рис. 4) має на увазі, що кожен фрагмент входу по-елементно множиться на невелику матрицю ваг (ядро), а результат підсумовується. Ця сума є елементом виходу, який називається картою ознак. Зважена сума входів пропускається через функцію активації.

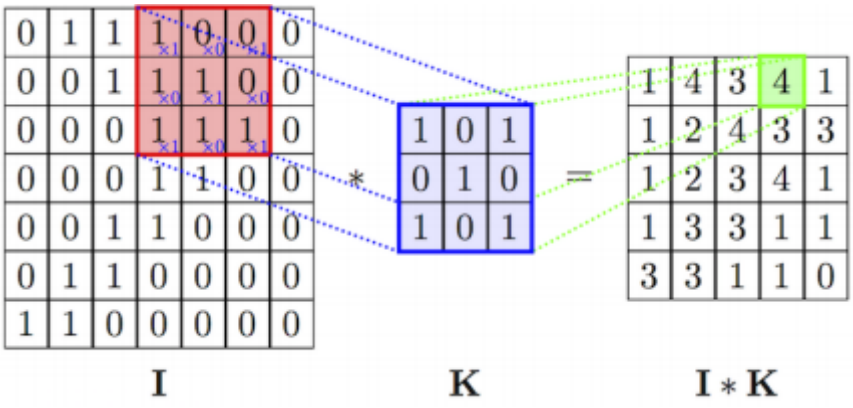

Рис. 4 Операція згортки

Для того щоб здійснити реалізацію таких класифікаторів як Long short-time memory, Naïve Bayes Classifier, Convolutional Neural Network необхідно обов'язково залучити низку бібліотек, які використовуються для машинного навчання : Keras та TensorFlow.

Keras - відкрита нейромережева бібліотека, написана на мові **Python** [10].

TensorFlow - це програмна бібліотека з відкритим вихідним кодом для високопродуктивних чисельних розрахунків, розроблена компанією Google [11].

Отже, для порівняння було реалізовано наступні класифікатори:

1. Naïve Bayes Classifier
2. LSTM + Keras framework
3. 1D CNN + Keras framework
4. 3D CNN + TensorFlow + Bert Tokenizer

Для того, щоб здійснити порівняння описаних класифікаторів та алгоритмів, доцільним є використати такі параметри як: час тренування, час тестування та точність класифікації. У рамках даної наукової роботи за точність будемо приймати відсоткове співвідношення кількості правильних передбачень класифікатора до розміру вибірки. В таблиці 1 представлено отримані результати:

Таблиця 1. Результати роботи

Класифікатор	Час тренування, сек	Точність тренування, %	Час тестування, с	Точність тестування, %

Naïve Bayes Classifier	~ 1 сек	98,0 %	~ 1 сек	88,06 %
LSTM + Keras	2 829 сек	95,03 %	97 сек	75,32 %
1D CNN + Keras	461 сек	99,0 %	21 сек	88,57 %
3D CNN + Tensorflow + Bert Tokenizer	2 304 сек	99,27 %	4 сек	89,72 %

На основі проведеного дослідження та у відповідності до отриманих результатів, які наведені у таблиці 1 варто зазначити, що найбільш негативні результати за всіма параметрами показала Long short–time memory мережа. Даний факт є можливість пояснити тим, що набір даних, який був обраний є доволі малий, аби повною мрою продемонструвати переваги рекурентних нейронних мереж.

У відповідності до параметру, який підкреслив час тренування та тестування найбільш привабливий результат показав класифікатор Байєса, однак варто зазначити, що окрім часового параметру він незначно поступається в точності класифікаторам, які є побудованими на основі згорткових нейронних мереж. Проте, варто зробити наголос, що максимальним недоліком баєсівського класифікатора є те, що якщо в тестовому наборі наявне слово, яке, в свою чергу, не знайдене у базі тренувальних даних, цей факт негативно вплине на ефективність класифікатора.

Під час проведення, порівняння класифікаторів на згорткових нейронних мережах, можна зробити висновок, що збільшення кількості шарів згортокової нейронної мережі та використання Bert Tokenizer збільшує час на навчання мережі, проте, це максимально зменшує час, який відводиться на тестування та максимально підвищує точність класифікації, що є беззаперечним позитивним фактом.

Висновки та перспективи подальшого дослідження. В ході виконання даної роботи було реалізовано декілька класифікаторів для проведення сентимент–аналізу та порівняно їх за ефективністю та часом навчання та тестування. Здійснено порівняння якості роботи за чотирма класифікаторами: Naïve Bayes Classifier; LSTM + Keras framework; 1D CNN + Keras framework; 3D CNN + TensorFlow + Bert Tokenizer. Доведено, що класифікатор Naïve Bayes Classifier демонструє результати краще, ніж інші методи. Так як вирішення задачі сентимент аналізу тексту пов'язано з обробкою великої кількості текстів, застосування згорткових нейронних мереж буде ефективніше з точки зору обчислювальної складності, ніж LSTM. Важливо відзначити, що в даній роботі розкрита проста архітектура зготкової нейронної мережі. Поліпшення параметрів моделі та по символьній передпідготовки даних [12], а також збільшення розміру початкового корпусу текстів дозволить значно підвищити характеристики класифікатора.

Список бібліографічного опису.

1. Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.
2. Кобець Н.М., Ковалюк Т.В. Задачі сентимент-аналізу як інструмент моніторингу міркувань користувачів соціальних мереж.//Матеріали наукової конференції студентів, магістрантів та аспірантів 23 – 24 квітня 2018 року «Інформатика та обчислювальна техніка» – ІОТ-2018, Київ, 2018. С.151-153.
3. Мазниченко Н. І. Автоматичний аналіз текстів електронних документів. Сучасні тенденції розвитку української науки : матеріали Всеукр. наук. конф., 6–7 трав. 2017 р.. Переяслав-Хмельницький, 2017. Вип. 2. С. 136–139.
4. Шингалов Д. В. Методи автоматичного аналізу тональності контенту у соціальних мережах для виявлення інформаційно-психологічних впливів / Д. В. Шингалов, Є. В. Мелешко, Р. М. Минайленко, В. А. Резніченко // Збірник наукових праць Кіровоградського національного технічного університету. Техніка в сільськогосподарському виробництві, галузеве машинобудування, автоматизація. 2017. Вип. 30. С. 196-202. Режим доступу: http://nbuv.gov.ua/UJRN/znpkntu_2017_30_29.
5. Шаховська Н. Б. Шкалювання емоційно забарвлених слів для використання у методах класифікації тональності / Н. Б. Шаховська, Х. Ю. Гірак // Вісник Національного університету "Львівська політехніка". Інформаційні системи та мережі. 2017. № 872. С. 195-203. Режим доступу: http://nbuv.gov.ua/UJRN/VNULPICM_2017_872_23.
6. Large Movie Review Dataset [Електронний ресурс]. - URL: https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews
7. Stuart J. Russell and Peter Norvig. Artificial Intelligence: A Modern Approach, Prentice Hall, 2003.
8. Understanding RNN and LSTM [Електронний ресурс]. – URL: https://towardsdatascience.com/understanding-rnn-and-lstm-f7cdf6dfc14e
9. Применение свёрточных нейронных сетей для задач NLP [Електронний ресурс] – URL: https://habr.com/ru/company/ods/blog/353060/
10. TensorFlow [Електронний ресурс] – URL: https://en.wikipedia.org/wiki/TensorFlow
11. Keras [Електронний ресурс] – URL: https://en.wikipedia.org/wiki/Keras
12. Understanding of Convolutional Neural Network (CNN) — Deep Learning [Електронний ресурс] – URL: https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148

13. Sentiment Analysis: Concept, Analysis and Applications [Електронний ресурс] – URL: https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17
14. Dalibor Bužić. Sentiment Analysis on Text Documents. URL: https://www.researchgate.net/publication/336716648_Sentiment_Analysis_of_Text_Documents

References

1. Bing Liu. (2012) Sentiment Analysis and Opinion Mining, *Morgan & Claypool Publishers*, May 2012.
2. Kobec N.M., Kovalyuk T.V. (2018) Zadachi sentiment-analizu yak instrument monitoringu mirkuvan koristuvachiv socialnih merezh.*Materiali naukovoyi konferenciyi studentiv, magistrantiv ta aspirantiv 23 – 24 kvitnya 2018 roku «Informatika ta obchislyuvalna tehnika» – IOT-2018*, Kiyiv. S.151-153.
3. Maznichenko N. I. (2017) Avtomatichnij analiz tekstiv elektronnih dokumentiv. *Suchasni tendenciyi rozvitku ukrayinskoyi nauki : materiali Vseukr. nauk. konf., 6–7 trav. 2017 r.*. Pereyaslav-Hmelnickij. Vip. 2. S. 136–139.
4. Shingalov D. V. (2017) Metodi avtomatichnogo analizu tonalnosti kontentu u socialnih merezhah dlya viyavlennya informacijno-psihologichnih vpliviv / D. V. Shingalov, Ye. V. Meleshko, R. M. Minajlenko, V. A. Reznichenko // *Zbirnik naukovih prac Kirovogradskogo nacionalnogo tehnichnogo universitetu. Tehnika v silskogospodarskomu virobnictvi, galuzeve mashinobuduvannya, avtomatizaciya.* Vip. 30. S. 196-202. Rezhim dostupu: http://nbuv.gov.ua/UJRN/znpkntu_2017_30_29.
5. Shahovska N. B. (2017) Shkalyuvannya emocijno zabarvlenih sliv dlya vikoristannya u metodah klasifikaciyi tonalnosti / N. B. Shahovska, H. Yu. Girak // *Visnik Nacionalnogo universitetu "Lvivska politehnika". Informacijni sistemi ta merezhi*. № 872. S. 195-203. Rezhim dostupu: http://nbuv.gov.ua/UJRN/VNULPICM_2017_872_23.
6. Large Movie Review Dataset [Elektronnij resurs]. - URL: https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews
7. Stuart J. (2003) Russell and Peter Norvig. Artificial Intelligence: *A Modern Approach, Prentice* Hall.
8. Understanding RNN and LSTM [Elektronnij resurs]. – URL: https://towardsdatascience.com/understanding-rnn-and-lstm-f7cdf6dfc14e
9. Primenenie svyortochnyh nejronnyh setej dlya zadach NLP [Elektronnij resurs] – URL: https://habr.com/ru/company/ods/blog/353060/
10. TensorFlow [Elektronnij resurs] – URL: https://en.wikipedia.org/wiki/TensorFlow
11. Keras [Elektronnij resurs] – URL: https://en.wikipedia.org/wiki/Keras
12. Understanding of Convolutional Neural Network (CNN) — Deep Learning [Elektronnij resurs] – URL: https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148
13. Sentiment Analysis: Concept, Analysis and Applications [Elektronnij resurs] – URL: https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17
14. Dalibor Bužić. Sentiment Analysis on Text Documents. URL: https://www.researchgate.net/publication/336716648_Sentiment_Analysis_of_Text_Documents