ЕКОНОМІКА

DOI: https://doi.org/10.32782/2521-666X/2024-87-1
УДК 339.543:328.185:004.8:519.2

Боженко В.В.
кандидат економічних наук,
Сумський державний університет
Гончарук І.Г.
аспірант,
Сумський державний університет

Bozhenko Victoria, Honcharuk Ivan
Sumy State University

АНАЛІЗ ТОНАЛЬНОСТІ ТЕКСТІВ НА ОСНОВІ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ МОНІТОРИНГУ СУСПІЛЬНИХ НАСТРОЇВ ЩОДО КОРУПЦІЇ[1]

SENTIMENTAL ANALYSIS OF TEXTS BASED ON MACHINE LEARNING METHODS TO MONITOR PUBLIC ATTITUDES TOWARDS CORRUPTION

Стаття присвячена аналізу тональності текстів публікацій в засобах масової інформації, присвячених питанням корупції, а також проведення аналізу контенту щодо корупції в світі та Україні. Для дослідження тональності тексту публікацій використано методи машинного навчання, а саме аналізатор VADER. Для аналізу обрано англомовні статті, які опубліковані в газеті The Guardian. Періодом дослідження обрано 2021–2024 роки. Проведений аналіз продемонстрував, що домінантною емоційною тональністю статей про корупцію у світі в газеті The Guardian була негативна. Водночас емоційне забарвлення новинних текстів щодо питань корупції в Україні є більш негативним порівняно зі світовими новинами з аналогічної проблематики. Результати дослідження мають практичне значення та можуть бути використані громадськими організаціями, державними установами при оцінюванні ефективності політики щодо запобігання та протидії корупції в країні.

Ключові слова: *корупція, сентимент аналіз, машинне навчання, засоби масової інформації.*

Monitoring public attitudes towards corruption is critical for governments, organisations and researchers seeking to understand the causes of its spread and ways to prevent and counteract these destructive processes in society. The article is devoted to analysing the corruption-related tone of media publications and making an analysis of the content on corruption in the world and in Ukraine. The relevance of solving this scientific problem is due to the fact that the use of machine learning methods allows processing large amounts of data from various sources and obtaining a detailed and dynamic understanding of public attitudes towards corruption. The article analyses the areas of application of sentiment analysis at different levels of economic relations. Machine learning methods, namely the VADER analyser, were used to study the tone of the text of publications about corruption. The articles published in The Guardian were selected for analysis. The study period was 2021–2024. For the analysis of the text tone, Orange Data Mining was chosen. The study of the issue of tone of texts in the media in the article is carried out in the following logical sequence: collection of text data on corruption, data pre-processing (cleaning and preparing the text for classification), transformation of textual data into numerical representations for machine learning models, and sentiment classification based on machine learning methods. The study found that during the first six months of the year with about 910 publications on corruption issues were published annually. One third of the corruption stories in The Guardian in 2022 mentioned Ukraine. The analysis showed that the dominant emotional tone of articles about corruption in the world was negative. At the same time, the emotional colouring of news texts on corruption in Ukraine is more negative compared to world news on similar issues. The results of the study are of practical importance and can be used by non-governmental organizations, government institutions and international organisations in assessing the effectiveness of policies to prevent and combat corruption in the country.

Keywords: *corruption, machine learning, media, sentiment analysis.*

[1] Роботу виконано в межах науково-дослідної теми, що фінансується за рахунок коштів державного бюджету (номер державної реєстрації 0124U000544).

Постановка проблеми. Повсюдне поширення корупції має далекосяжні наслідки, впливаючи на соціально-економічну та політичну структуру суспільства в усьому світі. Моніторинг суспільних настроїв щодо корупції є критично важливим для урядів, організацій та дослідників, які прагнуть зрозуміти причини її поширення та способи запобігання та протидії даним деструктивним процесам у суспільстві. За даними міжнародної організації Transparency International понад 80 % населення світу проживає в країнах, де індекс сприйняття корупції нижче середнього світового показника, а саме 43 ум. од. [2]. Попри довготривалу боротьбу урядів країн та міжнародних організацій з корупцією, вона досі створює загрози для національної безпеки цих країн та стримує темпи економічного зростання та соціального добробуту населення.

З появою цифрових засобів комунікації та соціальних мереж з'явилася значна кількість текстових даних, що відображають суспільні настрої, і це дає унікальну можливість для більш ефективного та всебічного аналізу. Методи машинного навчання пропонують потужний інструментарій для аналізу цих текстових даних. Використовуючи такі методи, як обробка природної мови, аналіз настроїв і глибоке навчання, можна автоматизувати оцінку тональності текстів, надаючи в реальному часі інформацію про ставлення громадськості до корупції. Ці методи можуть обробляти великі обсяги даних з різних джерел, зокрема постів у соціальних мережах, новинних статей, блогів та онлайн-форумів, що дає змогу отримати детальне та динамічне розуміння суспільних настроїв.

Аналіз останніх досліджень і публікацій. Методи машинного навчання передбачають навчання алгоритмів на маркованих наборах даних для розпізнавання настрою в тексті. Найпоширенішими алгоритмами для аналізу текстових даних є наївний Байєс [1; 16; 19], метод опорних векторів [5] і нейронні мережі [7; 21; 23]. Для підвищення точності аналізу текстових даних на практиці використовують гібридні методи, які поєднують підходи на основі лексики та машинного навчання [10; 18; 20].

Аналіз тональності тексту має практичне застосування на різних рівнях управління економічними відносинами. Дослідження публіцистичних та інформаційних повідомлень в Інтернеті, а також телевізійних матеріалів можуть використовуватися для оцінювання реакції громадськості на економічну політику держави, або доповнювати традиційні економічні показники для того, щоб ідентифікувати ранні сигнали про переломні моменти в економіці, такі як рецесія чи підйом. Зокрема, Lukauskas M. та ін. (2022) при прогнозуванні макроекономічних показників проаналізували настрої в засобах масової інформації [9]. У роботі Saito T. (2022) проаналізовано економічний звіт президента Сполучених Штатів для оцінювання економічного стану [17]. За результатами дослідження виявлено, що саме рівень безробіття, а не політичні фактори, такі як Конгрес і вибори, є визначальним фактором впливу на економічну політику в країні. Аналізуючи оголошення про вакансії, відгуки працівників та дискусії в соціальних мережах, аналіз настроїв дозволяє ідентифікувати ключові тенденції на ринку праці, задоволеність працівників та нові вимоги до навичок. Devi G. та ін. (2023) для оцінювання рівня задоволеності умовами праці співробітниками використання техніку Flair Pytorch, систему програмування природною мовою типу вбудованого програмування та унікальну стратегію для Twitter SA з акцентом на емоджі [15]. За результатами аналізу контенту соціальної мережі Twitter за допомогою класифікатора наївного Байєсу у роботі Qaiser S. та ін. (2020) встановлено, що 65 % людей мають негативні настрої щодо впливу технологій на зайнятість і технологічний прогрес; тобто у суспільстві формується думка, що люди повинні здобувати нові навички, щоб мінімізувати вплив структурного безробіття [14]. Крім цього, на основі аналізу соціальних мереж, новин та інших відкритих даних про компанію (веб-сайт, фінансові звіти, відгуки споживачів) з використанням інструментів машинного навчання можна визначити настрої інвесторів відповідної компанії та спрогнозувати її ринкову вартість у майбутньому, проаналізувати рівень лояльності споживачів до бренду та розробити клієнторієнтовану політику компанії тощо. У роботах Murugavalli S. та інші [11] та Nichifor та інші [13] оцінено рівень лояльності клієнтів до продукту з використанням методів сентимент-аналізу. У наукових статтях Nemes L. та ін. [12], Wu W. та ін. [22] використано методи машинного аналізу текстових даних для прогнозування рівня волатильності цін на товари та послуги.

Для розробки ефективних антикорупційних стратегій важливим є точне визначення стимулів до корупції. Традиційні методи оцінювання корупції є досить ресурсномісткі та вимагають тривалого часу для проведення дослідження, починаючи від збору даних та закінчуючи презентацією результатів оцінювання корупції. Машинне навчання, аналітика великих даних та технологія блокчейн є передовим напрямком досліджень у цій сфері. Постійна міждисциплінарна співпраця та методологічний прогрес матимуть вирішальне значення для подолання складної та еволюціонуючої природи корупції в усьому світі.

Мета статті полягає в аналізі тональності текстів публікацій, пов'язаних з корупцією, за допомогою методів машинного навчання, а також проведенні порівняльного аналізу контенту щодо корупції в світі та Україні.

Виклад основного матеріалу. Аналіз тональності текстів або аналіз думок – це метод, який використовується в обробці природної мови для визначення емоційного тону, що стоїть за текстом. Він передбачає аналіз текстових даних для виявлення та класифікації думок, висловлених у тексті, як позитивних, негативних або нейтральних. Аналіз настроїв використовує машинне навчання, аналіз текстів і комп'ютерну лінгвістику для обробки даних з різних джерел. Аналіз тональності тексту публікацій, присвячених питанням корупції, включає кілька етапів вилучення та аналізу суб'єктивної інформації з текстових даних (рис. 1). Для аналізу тональності тексту було обрано Orange Data Mining [3; 4], який було розроблено групою науковців з Університету Любляни.

Основними етапами для аналізу тональності тексті, присвячених питанням корупції, є:

1. Збір текстових даних

Для аналізу обрані англомовні статті, які опубліковані в газеті The Guardian. У роботі проаналізовано настрої, висловлені в газеті The Guardian щодо питань корупції. Протягом першого півріччя 2021–2024 років було ідентифіковано 3651 статей та інших інформаційних повідомлень, в яких згадувалося питання корупції (рис. 2). Щорічно протягом першого півріччя у середньому публікувалося близько 910 публікацій щодо питань корупції на вищезазначеному інформаційному ресурсі. З початком війни росії проти України зріс фокус уваги до нашої країни, у тому числі в питаннях протидії та запобігання корупції. Так, у 2022 році в третині інформаційних повідомлень щодо корупції в газеті The Guardian згадується Україна. У наступні два роки (2023 р., 2024 р.) дещо зменшилася кількість даних повідомлень в інформаційному просторі, де обговорюються питання корупції та України, проте залишається досить високим (близько 25 % від загального обсягу).

2. Попередня обробка тектових даних

У межах даного етапу було проведено токенізацію (розбиття тексту на окремі слова або лексеми), створено перелік стоп-слів та їх видалення з аналізу, які не відображають настрою автора (the, and, by, of та інші), а також видалення шуму (нерелевантних символів, знаків пунктуації та цифр).

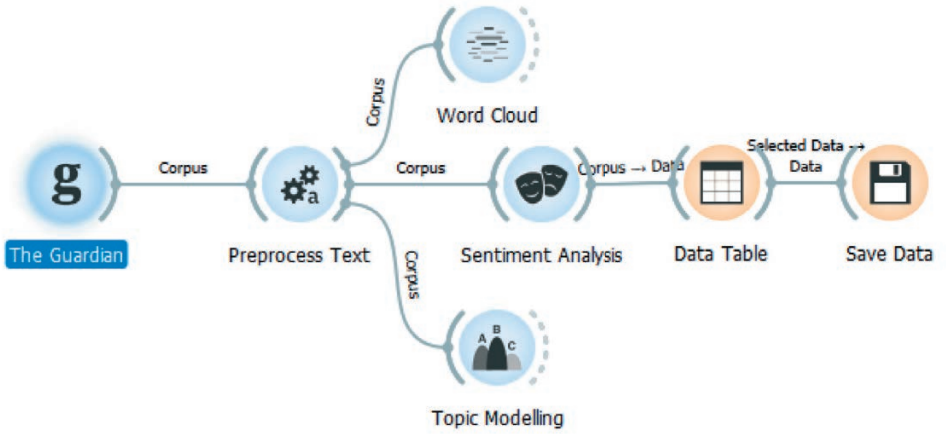

Рис. 1. Блок-схема аналізу тональності тексту публікацій у сфері корупції

Джерело: Orange Data Mining

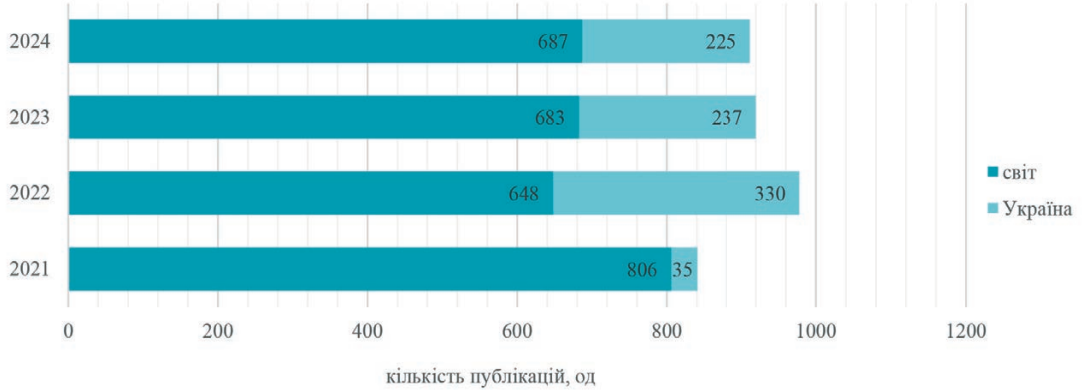

Рис. 2. Кількість публікацій з тематики корупція, опублікованих в газеті The Guardian протягом 2021–2024 рр.

Джерело: власна розробка авторів

3. Перетворення текстових даних у числові представлення для моделей машинного навчання, що передбачає екстракцію ознак або векторизацію тексту. У межах даного дослідження використано модель «мішок слів».

4. Класифікація настроїв на основі методів машинного навчання

Для дослідження тональності тексту використано аналізатор VADER (Valence Aware Dictionary for sEntiment Reasoning) [6], який визначає полярність настрою (позитивний, негативний або нейтральний) певного фрагмента тексту. VADER використовує комбінацію лексики настроїв, тобто набір лексичних ознак, які зазвичай маркуються відповідно до їхньої семантичної спрямованості. Класифікація тональності текстів за допомогою аналізатору VADER відбувається у розрізі трьох груп: позитивні (pos), негативні (neg) та нейтральні емоційні оцінки (neu), на основі яких визначається узагальнена оцінка (compound). Узагальнена оцінка може набувати значення в діапазоні від -1 до +1. Чим вища узагальнена оцінка, то більш позитивний текст, і навпаки. Зазвичай використовують такі порогові значення для віднесення тексту до однієї з трьох категорій: якщо узагальнена оцінка більше або дорівнює 0,05, то текст позитивний; якщо узагальнена оцінка менше або дорівнює -0,05; в інших випадках оцінка інтерпретується як нейтральна. На основі використання аналізатору VADER визначено узагальнену оцінку тональності тексту публікацій щодо корупції протягом 2021–2024 рр. (рис. 3).

Проведений аналіз продемонстрував, що протягом 2021–2024 р. домінантною емоційною тональністю статей про корупцію у світі в газеті The Guardian була негативна. Зокрема, у 2021 році узагальнена оцінка тональності тексту публікацій, присвячених питанням корупції, становила – 0,16 ум.од., тоді як у 2023 році -0,19 ум.од. Виключенням є 2022 рік, коли відбулося у світі незначне превалювання інформаційних повідомлень з позитивною тональністю. Іншими словами автори цих статей висвітлюють позитивні емоційні тони щодо питань корупції (наприклад, успішні практики протидії та запобігання корупції, удосконалення процесуальних процесів щодо викривлення та покарання винних у корупції та пов'язаних з нею злочинів тощо).

Емоційне забарвлення новинних текстів щодо питань корупції в Україні є більш негативним порівняно зі світовими новинами з аналогічної проблематики. Тенденція загальної негативізації новин щодо корупції та України в газеті The Guardian пояснюється низкою чинників: протистоянням політичних сил в державах Заходу щодо підтримки України, виділення значної за обсягом фінансової допомоги для України та посилення вимог до транспарентності публічних установ, реальні факти викриття масштабних корупційних правопорушень в Україні тощо. Це дозволяє говорити, що в англомовних засобах масової інформації питання протидії та запобігання корупції в Україні має негативне забарвлення, тобто автори даних інформаційних повідомлень висвітлюють негативне ставлення до подолання даного деструктивного процесу.

Висновки. Підводячи підсумок, сентимент-аналіз є ефективним новітнім інструментом для моніторингу суспільних настроїв та інформаційного забезпечення антикорупційних стратегій. У межах даного дослідження використано методи машинного навчання для запровадження більш ефективного, дієвого і комплексного підходу до моніторингу та аналізу громадської думки, що дозволяє зацікавленим сторонам швидше і точніше реагувати на складну і мінливу природу корупції. Проведений сентимент-аналіз засвідчив про негативізацію новин щодо корупції як на світовому, так і на національному рівнях.

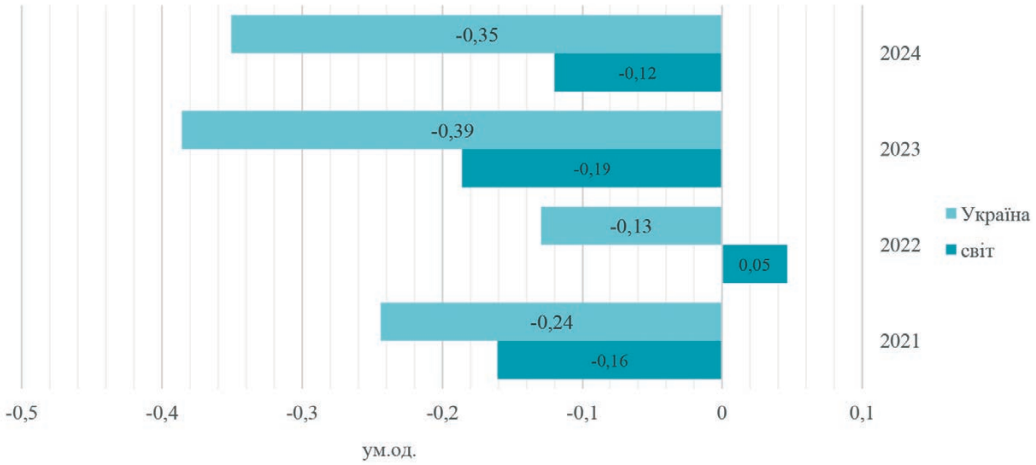

Рис. 3. Узагальнена оцінка тональності тексту публікацій, присвячених питанням корупції, в газеті The Guardian

Джерело: власна розробка авторів

References:

1. Abbas M., Ali K., Jamali A., Ali Memon K., & Aleem Jamali A. (2019) Multinomial Naive Bayes Classification Model for Sentiment Analysis. *IJCSNS International Journal of Computer Science and Network Security*, no. 19(3), pp. 62–67. Available at: https://www.researchgate.net/publication/334451164

2. Corruption Perception Index. (2023). Transparency International. Available at: https://images.transparencycdn.org/images/CPI-2023-Report.pdf

3. Demšar J., & Zupan B. (2013) Orange: Data mining fruitful and fun – A historical perspective. *Informatica (Slovenia)*, no. 37(1), pp. 55–60.

4. Demšar J., Curk T., Erjavec A., Gorup Č., Hočevar T., Milutinovič M., … Zupan B. (2013) Orange: Data mining toolbox in python. *Journal of Machine Learning Research*, no. 14, pp. 2349–2353.

5. Han K. X., Chien W., Chiu C. C., & Cheng Y. T. (2020) Application of support vector machine (SVM) in the sentiment analysis of twitter dataset. *Applied Sciences (Switzerland)*, no. 10(3). DOI: https://doi.org/10.3390/app10031125

6. Hutto C. J. & Gilbert E. E. (June 2014) VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI.

7. Jianqiang Z., Xiaolin G., & Xuejun Z. (2018) Deep Convolution Neural Networks for Twitter Sentiment Analysis. *IEEE Access*, no. 6, pp. 23253–23260. DOI: https://doi.org/10.1109/ACCESS.2017.2776930

8. Jura M., Spetz J., & Liou D. M. (2022) Assessing the Job Satisfaction of Registered Nurses Using Sentiment Analysis and Clustering Analysis. *Medical Care Research and Review*, no. 79(4), pp. 585–593. DOI: https://doi.org/10.1177/10775587211035292

9. Lukauskas M., Pilinkienė V., Bruneckienė J., Stundžienė A., Grybauskas A., & Ruzgas T. (2022) Economic Activity Forecasting Based on the Sentiment Analysis of News. *Mathematics*, no. 10(19). DOI: https://doi.org/10.3390/math10193461

10. Mahmood A. T., Kamaruddin S. S., Naser R. K., & Nadzir M. M. (2020) A combination of lexicon and machine learning approaches for sentiment analysis on facebook. *Journal of System and Management Sciences*, no. 10(3), pp. 140–150. DOI: https://doi.org/10.33168/JSMS.2020.0310

11. Murugavalli S., Bagirathan U., Saiprassanth R., & Arvindkumar S. (2017) Feedback analysis using Sentiment Analysis for E-commerce. *International Journal of Latest Engineering Research and Applications*, no. 02(03), pp. 2455–7137.

12. Nemes L., & Kiss A. (2021) Prediction of stock values changes using sentiment analysis of stock news headlines. *Journal of Information and Telecommunication*, no. 5(3), pp. 375–394. DOI: https://doi.org/10.1080/24751839.2021.1874252

13. Nichifor E., Brătucu G., Chițu I. B., Lupșa-Tătaru D. A., Chișinău E. M., Todor R. D., … Bălășescu S. (2023) Utilising Artificial Intelligence to Turn Reviews into Business Enhancements through Sentiment Analysis. *Electronics (Switzerland)*, no. 12(21). DOI: https://doi.org/10.3390/electronics12214538

14. Qaiser S., Yusoff N., Ahmad F. K., & Ali R. (2020) Sentiment analysis of impact of technology on employment from text on twitter. *International Journal of Interactive Mobile Technologies*, no. 14(7), pp. 88–103. DOI: https://doi.org/10.3991/IJIM.V14I07.10600

15. Devi G. D., & Kamalakannan S. (2023) Sentimental Analysis (SA) of Employee Job Satisfaction from Twitter Message Using Flair Pytorch (FP) Method. *Lecture Notes on Data Engineering and Communications Technologies,* no. 131, pp. 367–380. DOI: https://doi.org/10.1007/978-981-19-1844-5_28

16. Ressan M. B., & Hassan R. F. (2022) Naïve-Bayes family for sentiment analysis during COVID-19 pandemic and classification tweets. *Indonesian Journal of Electrical Engineering and Computer Science*, no. 28(1), pp. 375–383. DOI: https://doi.org/10.11591/ijeecs.v28.i1.pp375-383

17. Saito T. (2022) Sentiment Analysis of the Economic Report of the President in the United States:What Determines Economic Recognition? *Kodo Keiryogaku (The Japanese Journal of Behaviormetrics)*, no. 49(2), pp. 197–205. DOI: https://doi.org/10.2333/jbhmk.49.197

18. Sham N. M., & Mohamed A. (2022) Climate Change Sentiment Analysis Using Lexicon, Machine Learning and Hybrid Approaches. *Sustainability (Switzerland)*, no. 14(8). DOI: https://doi.org/10.3390/su14084723

19. Song J., Kim K. T., Lee B., Kim S., & Youn H. Y. (2017) A novel classification approach based on Naïve Bayes for Twitter sentiment analysis. *KSII Transactions on Internet and Information Systems*, no. 11(6), pp. 2996–3011. DOI: https://doi.org/10.3837/tiis.2017.06.011

20. Srivastava R., Bharti P. K., & Verma P. (2022) Comparative Analysis of Lexicon and Machine Learning Approach for Sentiment Analysis. *International Journal of Advanced Computer Science and Applications*, no. 13(3), pp. 71–77. DOI: https://doi.org/10.14569/IJACSA.2022.0130312

21. Sugumaran P., & Uma A. B. B. K. (2022) Real-time twitter data analytics of mental illness in COVID-19: sentiment analysis using deep neural network. *Indonesian Journal of Electrical Engineering and Computer Science*, no. 26(1), pp. 560–567. DOI: https://doi.org/10.11591/ijeecs.v26.i1.pp560-567

22. Wu W., Xu M., Su R., & Ullah K. (2024) Modeling crude oil volatility using economic sentiment analysis and opinion mining of investors via deep learning and machine learning models. *Energy*, no. 289. DOI: https://doi.org/10.1016/j.energy.2023.130017

23. Xiang R., Li J., Wan M., Gu J., Lu Q., Li W., & Huang C. R. (2021) Affective awareness in neural sentiment analysis. *Knowledge-Based Systems*, no. 226. DOI: https://doi.org/10.1016/j.knosys.2021.107137

24. Zheng Y., Long Y., & Fan H. (2022) Identifying Labor Market Competitors with Machine Learning Based on Maimai Platform. *Applied Artificial Intelligence*, no. 36(1). DOI: https://doi.org/10.1080/08839514.2022.2064047