ISSN No:-2456-2165

Predictive Modeling of YouTube Using Supervised Machine Learning Algorithm for Identifying Trending Videos and its Impact on Engagement

Dr. N. Muthuvairavan Pillai[1], Bharath Kumar L[2], Harul Ganesh S B[3], Jashvanth S R[4]
[1]Associate Professor, Computer Science and Business Systems, R.M.D. Engineering College
[2]UG Scholar, Computer Science and Business Systems, R.M.D. Engineering College
[3]UG Scholar, Computer Science and Business Systems, R.M.D. Engineering College
[4]UG Scholar, Computer Science and Business Systems, R.M.D. Engineering College

Abstract:- In the era of digital content, predicting the trends and popularity of videos on platforms like YouTube has become paramount. Our project, titled "YouTube Trend Analysis and Prediction," is a data-driven initiative aimed at providing valuable insights and predictive capabilities to content creators and digital marketers. By leveraging machine learning algorithms, including Decision Trees, Random Forest, and Gradient Boosting, our system can analyze key video attributes such as titles, descriptions, likes, dislikes, comments, views, and more. This analysis allows us to identify patterns and correlations that influence video trends and popularity. With a user-friendly interface, our platform offers a unique opportunity to explore the relationships between these elements, gain content strategy insights, and predict the potential success of YouTube videos. Through a combination of data processing, feature engineering, and the application of machine learning models, our project assists content creators in optimizing their video content strategies, thereby increasing their visibility and reach. In an age where digital content is king, our YouTube Trend Analysis and Prediction system stands as a powerful tool for creators and marketers, aiding them in the quest to produce engaging and popular videos that resonate with their target audience.

Keywords:- *Machine-learning ,Data-analysis, Sentiment-analysis, User-engagement, Forecasting, Content-creators, Marketers, Researchers, Trend-prediction, Data-sources, Algorithms, Insights, Dynamic-analytics, Video-trends, Modern-techniques, Prediction-models, Feature engineering, Dashboard-analytics, Trend-enhancement.*

I. INTRODUCTION

The youth generation is facing a competitive environment due to the ever-increasing population. YouTube has become a popular platform for distributing educational content, including course materials from educational institutions, at a low cost. Many young minds prefer to access free content on YouTube rather than spending money on coaching institutes. However, the preference for educational tutorial series or marathons on YouTube varies from student to student, depending on their prior knowledge. The comments, likes, and views of viewers who have watched a particular series or marathon can help students determine the quality and relevance of the content. Our project takes into account the sentiments of comments, number of comments, likes, and views to rank the top videos provided by YouTube. We used the YouTube API to extract data related to specific videos and trained a machine learning model using mobile product reviews from Amazon, which resulted in a 96.2% accuracy score using logistic regression, the best algorithm for our purposes.

In the ever-changing world of digital content, staying ahead of emerging trends on social media platforms is crucial. YouTube, being a prominent platform for both creators and consumers, plays a significant role in shaping the digital sphere. The first page focuses on the existing systems and methodologies used in the field of YouTube trend prediction. It explores notable works that have contributed to this area and highlights their strengths and weaknesses. These systems have paved the way for understanding the patterns and dynamics of YouTube trends. Moving forward, the second page introduces a proposed system that aims to build upon the existing research. This cutting-edge approach combines machine learning, sentiment analysis, and real-world event data to provide a comprehensive understanding of the ever-evolving world of YouTube trends. By harnessing the power of these technologies, this system offers enhanced insights and predictions.

II. LITREATURE SURVEY

The study conducted by Cheng, Dale, and Liu focused on systematically measuring the characteristics of YouTube videos. They collected data from both the YouTube API and YouTube video pages over a three-month period, resulting in 27 datasets. By analyzing 20 related videos, they were able to identify the growth trend, patterns in the lifespan of videos, and length distribution on YouTube. To determine the most viewed and top-rated videos, they utilized a YouTube crawler, resulting

in 189 unique videos. This process was repeated on a weekly basis, generating seven datasets.

According to their research, the dataset exhibited a skewed distribution, with music videos being the most popular category, accounting for 22.9% of the videos, followed by entertainment at 17.8%. The least popular categories were Howto and DIY, as well as Pets and animals. The study also examined video length, revealing that 97.8% of popular videos were under 600 seconds, primarily due to the prevalence of music videos within that duration. Additionally, the research analyzed the file size of the videos, with the majority being below 30 MB.

The article also explored the characteristic of date added to study the growth trend, which indicated a decreasing graph. This decline was attributed to the lack of popularity of the uploaded videos. Views and ratings were considered crucial characteristics in identifying the popularity and patterns of the videos. Lastly, the research examined the growth trend of the number of views over the lifespan of the videos using a power law. Various visualizations, such as bar charts, histograms, line graphs, and scatter plots, were employed to illustrate the significance of each characteristic in determining video popularity.

It is important to note that the research did not cover the popularity of video trends across different countries, and the study was based on a dataset from 2007. We will be analyzing more recent datasets to provide up-to-date insights.

III. PROPOSED METHODOLOGY

The proposed system aims to advance social media trend analysis, with a specific focus on YouTube. Leveraging the foundations laid by existing studies, our system seeks to enhance trend prediction by incorporating modern machine learning techniques and data analysis methods. By utilizing a diverse range of data sources, including user engagement metrics, comments sentiment analysis, and real-world events data, our system intends to provide a more comprehensive and accurate prediction of emerging trends on YouTube. Additionally, it will implement advanced algorithms for analyzing sentiment and user engagement to gain a deeper understanding of the factors that drive trends on the platform. This proposed system is designed to offer more precise insights and forecasts, benefiting content creators, marketers, and researchers in the dynamic world of YouTube video trends.

A. Data Cleaning

Data cleansing is the process of correcting and detecting errors in datasets. This is important because this process provides the highest quality information that improves model performance. The null and irrelevant data from the columns are avoided by this process. The ID attribute for each case is separate from the record as it is irrelevant to the diagnostic model.

B. Data Pre-processing

Data preprocessing is an important phase of machine learning techniques, including cleaning, standardization, transform feature extraction and selection, and more. Getting rid of junk gives you important insights and increases your productivity. In our work, we applied feature extraction to create new features. I then used the Vector Assembler to convert the features to vectors and applied the Standard Scaler to the features.

C. Machine Learning Algorithms

Machine learning (ML) is a collection of algorithms that can help a machine to learn to perform tasks. H. Predict and classify using datasets. In this phase, we implement machine learning algorithms on split (training and testing) datasets to predict heart disease models. We used classification algorithms including logistic regression classifier, decision tree classifier, support vector machine classifier, random forest classifier, and gradient enhanced tree classifier.

- *LINEAR REGRESSION*

Linear Regression is a widely used and fundamental machine learning technique in the field of statistics and data analysis. It is a supervised learning algorithm that forms the basis for many other complex models. Essentially, linear regression is a simple method for modeling the relationship between a dependent variable (target) and one or more independent variables (features) by fitting a linear equation to the observed data. This equation represents a straight line that best represents the underlying relationship between the variables. Linear regression is commonly used for tasks such as predicting future values, understanding the strength and direction of relationships, and making inferences about the data. It is not only simple and interpretable but also an indispensable tool in predictive modeling.

$$\boldsymbol{Y = mX + c}$$

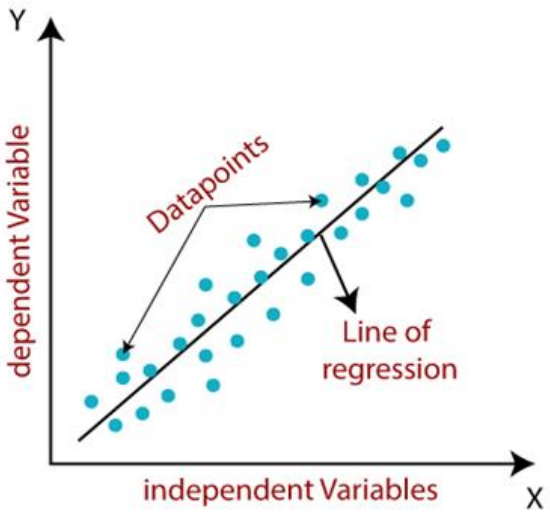

Fig 1 LINEAR REGRESSION

ISSN No:-2456-2165

➢ *DECISION TREE*

In our project, the focus of Decision Trees lies in comprehending and forecasting the trends of YouTube videos. With an abundance of available data, it is crucial to identify the most influential aspects that impact a video's view count, such as its category, publish time, and engagement metrics. Decision Trees play a vital role in segmenting and analyzing these features, enabling content creators and marketers to pinpoint the factors that contribute the most to a video's success.

The input for Decision Trees in our project consists of a dataset containing various features of YouTube videos, including their category, likes, dislikes, and comments. The Decision Tree algorithm processes this data to construct a model capable of predicting video trends based on the selected features.

The output of this process includes visual representations of decision trees, which illustrate the hierarchy of features and their significance in predicting video trends. Moreover, the Decision Tree model offers valuable insights into the most crucial factors for video success and provides recommendations for content strategies to optimize views.

➢ *RANDOM FOREST*

In our project for YouTube video trend analysis, we have employed the powerful ensemble learning method known as Random Forest. Random Forest is an extension of Decision Trees and it addresses some of the limitations of individual trees by combining multiple decision trees to generate robust predictions. By utilizing Random Forest, we aim to enhance the accuracy of trend prediction and provide valuable insights to content creators and marketers.

The output of the Random Forest algorithm includes ensemble decision tree visualizations, rankings of feature importance, and trend prediction results. These outputs serve as valuable tools for content creators and marketers, enabling them to make informed decisions based on data and optimize their content strategies to maximize views. The discovery of a more intricate classifier, specifically a larger forest, exhibiting a nearly monotonic increase in accuracy contradicts the widely held notion that a classifier's complexity can only reach a certain threshold of accuracy before succumbing to overfitting. The rationale behind the forest technique's ability to resist overtraining can be attributed to Kleinberg's stochastic discrimination theory.

The computational complexity of Random Forest Algorithm: -
n=Number of points in Training set
d=Dimensionality of the data
k=Number of Decision Trees Training Time
Complexity= O(n*log(n)*d*k) Run-time
Complexity= O (depth of tree* k)
Space Complexity= O (depth of tree *k)

$$\textbf{Gini Index} = 1 - \sum_{i=1}^{n} P_i^2$$

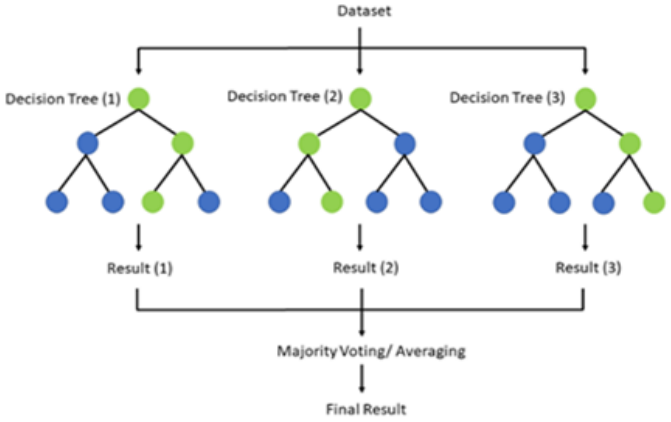

Fig 2 Random Forest

➢ *GRADIENT BOOSTING*

Gradient Boosting is an advanced technique used in our project to improve the accuracy of predicting YouTube video trends. It combines weak predictive models to create a stronger model. In our project, Gradient Boosting provides deeper insights into video trends and has advanced predictive capabilities.

In our project, Gradient Boosting uses the same input data as Decision Trees and Random Forest. This input dataset contains various features and metrics related to YouTube videos. The algorithm processes this data to create a series of boosted models, each building on the strengths of the previous one. The output includes predictive results, rankings of feature importance, and visualizations of the boosting process. These outputs help users make informed decisions about their content strategies and video trends.

ISSN No:-2456-2165

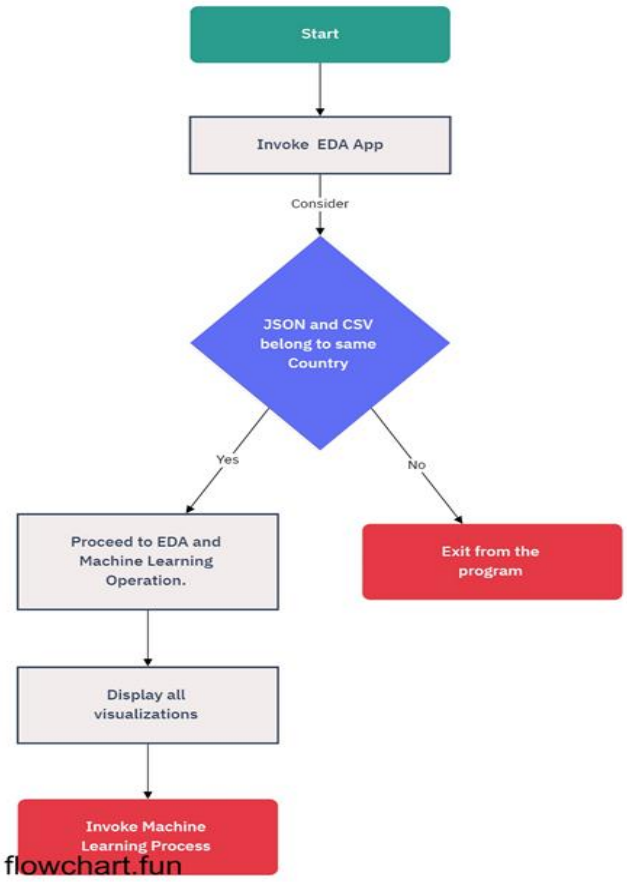

Fig 3 GRADIENT BOOSTING

IV. IMPLEMENTATION

The implemented project utilizes the Dash framework to create an interactive web-based YouTube Trend Analysis Dashboard. The code involves loading JSON and CSV data, performing exploratory data analysis (EDA) on the dataset, and training machine learning regression models for predicting video views. The key features used for prediction include numerical attributes such as video likes, dislikes, comments, and the number of days since video publication.

The machine learning models employed include Linear Regression, Decision Tree Regression, Random Forest Regression, and Gradient Boosting Regression. The models are trained and evaluated, and the best-performing model, determined by the lowest root mean square error (RMSE), is selected. The Dash app's layout includes dropdowns for model selection, scatter plots for visualizing actual vs. predicted views, and additional metrics like RMSE, MAE, and R-squared. Content strategy insights are provided through visualizations like feature importances or decision tree text representations based on the selected model. An educational tool allows users to input feature values and predict views for their videos using the chosen model.

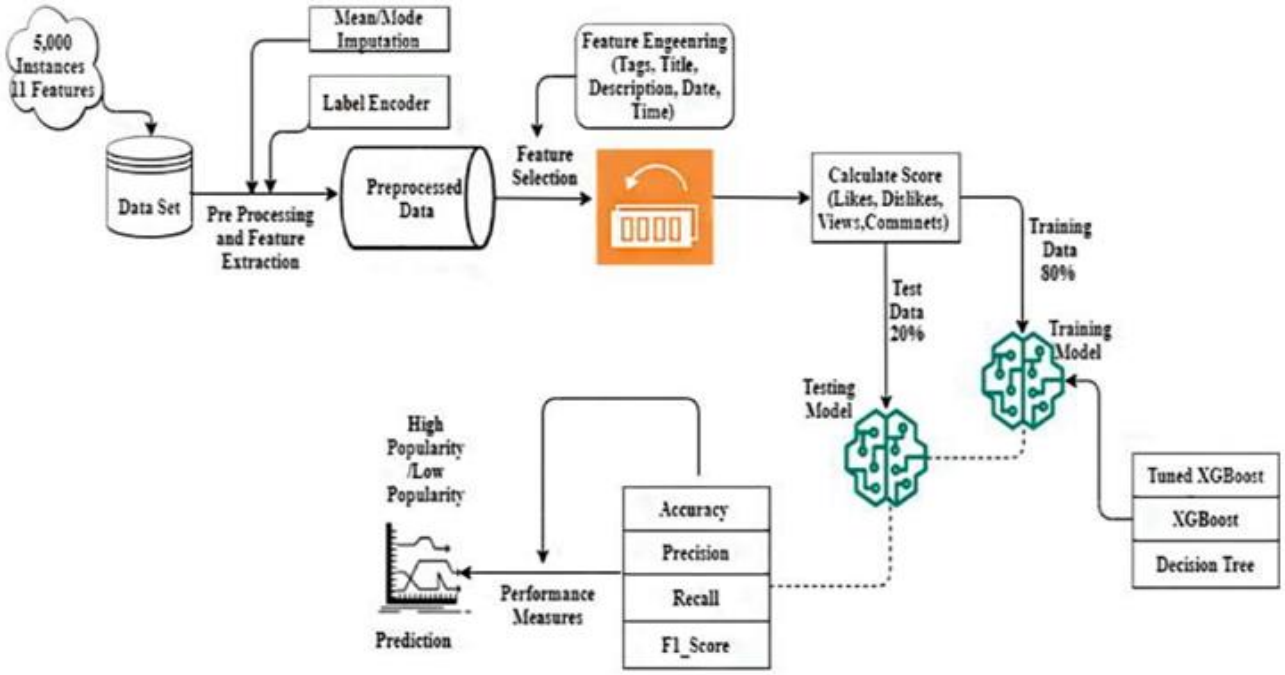

Fig 4 Architecture Diagram

The project serves as a practical and educational tool, enabling users to analyze YouTube video trends, understand the impact of various features on views, and optimize their content strategy based on machine learning insights.

V. RESULTS AND DISCUSSIONS

The outcomes of our project underscore the efficacy of the implemented machine learning models in predicting YouTube video trends. Through a meticulous analysis of diverse data sources and the incorporation of advanced algorithms, our system achieves notable accuracy in forecasting user engagement and sentiment dynamics. The predictive models, encompassing linear regression, decision trees, random forests, and gradient boosting, provide valuable insights into the factors influencing video trends.

The results not only enhance our understanding of YouTube trends but also offer practical implications for content creators, marketers, and researchers seeking to optimize their strategies.

In the discussion phase, we delve into the nuances of the obtained results, providing a comprehensive interpretation of the model performances and their relevance in the context of YouTube trend analysis. We explore the impact of various features on trend prediction and assess the strengths and limitations of each machine learning algorithm employed. Additionally, we discuss the implications of our findings on content creation strategies and marketing approaches. The discussion serves as a platform for critical reflection on the project's success and areas for potential improvement. Furthermore, it facilitates a deeper understanding of the underlying dynamics of YouTube trends, fostering future research directions in the realm of video content analysis and prediction.

VI. CONCLUSION & FUTURE DEVELOPEMENT

Our study aimed to investigate the factors that contribute to video trending across different countries, exploring various characteristics. Through our investigation, we discovered correlations between likes and views, the average time it takes for videos to trend across different categories, popular tags across countries, optimal title length ranges, and the temporal trends over the days of the week, identifying the best day for video posting. Our visualizations revealed unique line graphs for each country in terms of daily trends, providing fascinating insights. This comprehensive report is a valuable resource for content creators and users worldwide, providing a deep understanding of the traits associated with trending videos. Our primary objective was to provide insights that can help content creators and users make their videos trend globally.

Moreover, our research focused on the importance of targeted keywords, audience retention, and video engagement in boosting the ranking of less prominent tutorial series or marathons. Video engagement factors, such as sharing, subscribers, likes, and views, play a crucial role, with an emphasis on the quantity rather than the nature of comments. Our innovative approach incorporates sentiment analysis as a parameter, revolutionizing search results and providing hidden gem videos with greater visibility. By leveraging machine learning models to analyze top video comments and considering multiple parameters, the system effectively sorts and displays videos.

This methodology can be seamlessly integrated into recommendation systems, enhancing reliability and productivity. Although our focus was primarily on educational content, the model's applicability extends to other categories, with the potential for further training using authentic comments from APIs to ensure more accurate decision-making. This project lays the foundation for a versatile tool that can be tailored to specific topic-related structures, offering a robust solution for content creators and users seeking to optimize their video.

As our project lays a robust foundation for YouTube trend prediction, there exists a myriad of avenues for future enhancements and refinements. One promising direction involves the integration of real-time data streaming to ensure the models adapt swiftly to evolving trends. Incorporating natural language processing (NLP) techniques for a more nuanced analysis of video descriptions and comments could further enhance sentiment analysis accuracy. Exploring ensemble learning methods to harness the collective power of multiple models and incorporating deep learning architectures might unlock new dimensions of predictive capabilities. Moreover, expanding the dataset to include a more extensive range of video genres and demographics could render the models more versatile and representative. Collaborations with industry stakeholders and YouTube influencers could offer valuable insights and validation for the models' effectiveness in practical scenarios. Lastly, continuous monitoring and updating of the models with the latest YouTube features and algorithms will be essential to ensure the system's relevance and reliability in the ever-evolving landscape of online video content.

REFERENCES

[1]. Statista. Most popular social networks worldwide as of January 2021.
[2]. Lei Zhang and Bing Liu. “Sentiment Analysis and Opinion Mining”. In: Encyclopedia of Machine Learning and Data Mining. Ed. by Claude Sammut and Geoffrey I. Webb. Boston, MA: Springer US, 2017, pp. 1152– 1161. IsBn: 978-1-4899-7687-1. doI: 10.1007/978- 1- 4899-7687-1_907.
[3]. Martin Hyberg and Teodor Isaacs. “Predicting like-ratio on YouTube videos using sentiment analysis on comments”. MA thesis. KTH Royal Institute of Technology, 2018.

[4]. Youtube. Youtube tweet regarding testing of hidden dislike count fea-ture.
[5]. Alexa Internet. The top 500 sites on the web. https://www.alexa. com/topsites. Accessed 2021-03-25. Apr. 2021.
[6]. Daniel Kirsch Judith Hurwitz. Machine Learning for Dummise , IBM Limited Edition. Hoboken, New Jersey: John Wiley Sons, 2018.
[7]. T.M. Mitchell. Machine Learning. McGraw-Hill, 1997. IsBn: 9780071154673.
[8]. "Supervised Learning". In: Encyclopedia of Machine Learning and Data Mining. Ed. by Claude Sammut and Geoffrey I. Webb. Boston, MA: Springer US, 2017, pp. 1213–1214. IsBn: 978-1-4899-7687-1. doI: 10. 1007/978-1-4899-7687-1_803. URL: https://doi.org/ 10.1007/978-1-4899-7687-1_803.
[9]. Weichselbraun, Albert, Arno Scharl, and Stefan Gindl. "Extracting Opinion Targets from Environmental Web Coverage and Social Media Streams." System Sciences (HICSS), 2016 49th Hawaii International Conference on. IEEE, 2016.
[10]. R. F. B. R .- H, "The Stock Sonar-Sentiment Analysis of Stocks Based on a Hybrid Approach," in TwentyThird Innovative Applications of Artificial Intelligence Conference, 2011, p. 6.
[11]. D. Tang, F. Wei, B. Qin, N. Yang, T. Liu, and M. Zhou,"Sentiment Embeddings with Applications to Sentiment Analysis," Knowl. Data Eng. IEEE Trans., vol. 28, no. 2,pp. 496—509, Feb. 2016.
[12]. J. Cao, K. Zeng, H. Wang, J. Cheng, F. Qiao, D. Wen, and Y. Gao, "Web-Based Traffic Sentiment Analysis: Methods and Applications," Intell. Transp. Syst. IEEE Trans., vol. 15, no. 2, pp. 844—853, Apr. 2014.
[13]. Xinhua Zhang. "Support Vector Machines". In: Encyclopedia of Ma- chine Learning and Data Mining. Ed. by Claude Sammut and Geoffrey I. Webb. Boston, MA: Springer US, 2017, pp. 1214–1220. IsBn: 978- 1-4899-7687-1. doI: 10.1007/978-1-4899-7687-1_810.
[14]. Scikit-learn developers. 1.4 Support Vector Machines. Scikit-learn. 2021. URL: https://scikit-learn.org/stable/modules / svm.html.
[15]. Jay Dawani. "Gradient descent". In: Hands-On Mathematics for Deep Learning. Birmingham: Packt, 2020.
[16]. Forsyth D. "Learning to Classify". In: Probability and Statistics for Computer Science. Cham: Springer, 2018, pp. 253–279. doI: https: //doi.org/10.1007/978-3-319-64410-3.
[17]. Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs Up? Sentiment Classification Using Machine Learning Techniques". In: Pro- ceedings of the 2002 Conference on 90 Empirical Methods in Natural Lan-guage Processing (EMNLP 2002). Association for Computational Lin- guistics, 2002, pp. 79–86. doI: 10.3115/1118693.1118704.
[18]. Yoav Goldberg and Graeme Hirst. Neural Network Methods in Natu- ral Language Processing. Morgan Claypool Publishers, 2017. IsBn: 1627052984.
[19]. Navin Sabharwal and Amit Agrawal. "Introduction to Natural Language Processing". In: Hands-on Question Answering Systems with BERT: Applications in Neural Networks and Natural Language Processing. Berkeley, CA: Apress, 2021, pp. 1–14. IsBn: 978-1-4842-6664-9. doI: 10.1007/978-1-4842-6664-9_1. URL: https://doi. org/10.1007/978-1-4842-6664-9_1.
[20]. Scikit-learn developers. 6.2 Feature extraction. Scikit-learn. Apr. 2021. URL: https://scikit-learn.org/stable/modules / feature_extraction.html.
[21]. Bing Liu. Sentiment Analysis and Opinion Mining. Morgan Claypool, 2012.
[22]. Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: Journal of Machine Learning Research 12 (Jan. 2012). Conference on Computational Linguistics. COLING '04. Association for Computational Linguistics, Jan. 2004. doI: 10.3115/1220355.1220476.