I. J. Education and Management Engineering, **2025, 1, 1-9**
Published Online on February 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijeme.2025.01.01

Analysis of Human Behavior and Interests Based on Text Data

Irada Alakbarova*
Ministry of Science and Education of Azerbaijan, Institute of Information Technologies, Azerbaijan
E-mail: airada.09@gmail.com
ORCID iD: https://orcid.org/0000-0002-9876-3035
*Corresponding Author

Received: 12 July, 2024; Revised: 21 August, 2024; Accepted: 01 November, 2024; Published: 08 February, 2025

Abstract: Information technology has revolutionized data collection and analysis, offering unprecedented opportunities to study human behavior. Various information registers, the internet of things, and electronic demographic platforms that collect and analyze user data from various online sources provide a unique opportunity to predict human behavior using machine learning methods. This study applies machine learning to analyze textual data derived from diverse sources: demographic data, scientific articles, employee documents, and social media content. The primary goal is to identify a person's area of interest and predict their behavior. We propose using Support Vector Machines (SVM) as a robust and versatile machine learning algorithm for text data analysis. SVM's ability to handle diverse data types makes it well-suited for analyzing complex human behavior patterns. By classifying documents into relevant topics, SVM can help assess how employee behavior aligns with organizational goals and performance metrics. This research aims to contribute to human behavior analysis by demonstrating the effectiveness of machine learning techniques, particularly SVM, in extracting meaningful insights from textual data.

Index Terms: Human Behavior; Machine Learning; TF-IDF; SVM; Demographic Data

1. Introduction

The digital revolution has changed every aspect of our lives, from the way we communicate to the way we work, learn, and shop. Along with these advances, the study of human behavior in the digital age has become increasingly important as we strive to understand how people interact with and are influenced by the vast array of digital technologies. The development of digital technologies, the global spread of the Internet, and the introduction of social networks into human life and their consequences provide new data and sources of information for the study of human behavior and demographic problems.

The digital revolution has radically changed research practices in all fields. Now everyone can collect data in different forms and even use mobile phone applications to monitor their health, and social relationships and study people's behavior. To ensure the security of personal data and the effective use of this data in analytics, the use of new sources, centralized collection, storage, and processing of personal data must be carried out through an electronic demographic (e-demographic) system. The formation of the information society and the use of information technologies in demographic research have created a new field of science – digital demography or electronic demography (e-demography) [1]. E-demography can be considered a new field of science that studies the demographic situation in the country and society, offering new technologies for the effective management of civil society.

The use of intelligent algorithms in e-demography ensures the creation of an intelligent e-demography system, based on achieving flexible, competitive, effective results that meet the requirements of Industry 4.0. The diversity of the source, and the fact that the data has different formats and structures, is constantly updated, and represents big data, complicates the issue of processing this data. At the same time, the security of citizens' data must be ensured. This is only possible if the capabilities of e-demography are available and a unified population register is created.

The creation of a unified population register requires a comprehensive approach to ensure the security and confidentiality of personal data. The key to successfully solving this problem is a combination of legal, technical, and organizational measures, as well as a commitment to ethical principles. When analyzing personal data, it is necessary to collect and store only those data that are necessary to achieve the set goals. Taking all necessary measures to protect the

confidentiality of data, ensuring transparency in the processes of collecting, processing, and storing data, as well as bearing responsibility for possible negative consequences are the main tasks of analysts working with personal data.

The main data for the analysis of human behavior are search queries in web browsers, social network data, data collected in government registers, electronic services data, feedback data on the relationship between government and citizens, telephone conversations, credit card transactions, information systems data insurance, medicine, education, etc.

The principles of behavior analysis using ML algorithms are used in education, medicine, banking and insurance operations, in the study of social relations, determining the psychological state of an individual, organizing services for children and people with disabilities, as well as in the field of healthcare and in the areas of increasing production efficiency [2, 3]. ML is changing the landscape of human behavior analysis.

These algorithms offer powerful tools for analyzing large data sets, identifying hidden patterns, and predicting future human behavior [2]. By analyzing vast data sets, these algorithms can reveal hidden data and relationships, allowing applications in areas such as marketing, healthcare, human-computer interaction, and more. Text, graphic, video, and audio data collected in electronic state registers, various information systems, clouds, and social networks constitute big data, which creates certain problems in their collection, structuring, and processing. An effective solution to the big data problem is only possible with the help of intelligent algorithms. Analysis of human behavior using intelligent algorithms can give an idea of a person's personal qualities, interests, and even emotional states, and predict his future actions in society.

The article is devoted to the analysis and prediction of human behavior using SVM on the digital demography platform. For this purpose, international experience in the field of human behavior is studied. A brief overview of various approaches for big data mining is given to determine human behavior and interests and a methodology for predicting human behavior is described. A comparative analysis of ML algorithms used to identify interests and analyze human behavior is carried out.

2. Related Work

There are different approaches to mining human behavior. These approaches include methods ranging from conventional statistical calculations to classification and optimization models. In work [4], using decision tree classifiers, the authors try to predict the behavior of students in the process of continuing their studies in graduate school (master's degree). The proposed approach is an implementation of the J48 algorithm analysis tool on data collected through surveys of students in different majors and on the demographic data of these students.

One of the proposed models for studying egocentric activity and human behavior in everyday life is the model of detection and assessment of behavioral activity [5]. In the first stage of the approach, the state of the object (human) is determined using a table of state-activity relationships. In the second stage, the image is obtained using a convolutional neural network (CNN), and the data obtained by determining the entropy of the image is classified into topics. In the third step, the motion data set is reclassified using an SVM. Abnormal behavioral activity is determined by classifying the overall results obtained at the end of the study.

The approach proposed in [6] takes into account the characteristics of the external background (view) and movement. The architecture of deep neural networks (DNN) is presented in the form of two networks: an appearance network and an object motion network. The joint work of these two networks makes it possible to determine the behavior of an object taking into account background movement.

In [7] addresses the problem of detecting hidden social networks in Wikipedia by detecting conflicting articles using probability theory and clustering unstructured text-type data. To solve the problem, the fuzzy c-means algorithm was used. Fuzzy clustering allows, under certain conditions, to divide a given set into fuzzy sets. Using this algorithm, it is possible to collect semantically relevant texts into clusters.

In [8], decision trees (TD) and C4.5 algorithms are used to determine the behavior of university students. The level of adaptation of students to online classes was determined using the Random Forest (RF) algorithm.

Another approach proposes to group people (objects) according to certain characteristics in order to determine the characteristics of their behavior [9] and anomalies in human behavior. To solve the problem, a tokens model was built. For this purpose, ML algorithms were used: DT, SVM and DNN. The authors are trying to determine the most effective algorithm for identifying anomalies in human behavior.

Taking into account the influence of human psychology on behavior, in [10] they try to study the behavior of an individual based on the analysis of texts written by him. The purpose of the study is to determine the psychological mood, understand and predict individual behavior by conducting intellectual analysis of text data. Data sources used were ISI Web of Science, Engineering Village Compendex, ProQuest Dissertations, and Google Scholar databases. Sentimental analysis of unstructured texts on various topics is carried out by searching these databases using keywords, the psychological state of the author is studied by classifying texts into positive and negative classes. The authors state that NLP, clustering and classification algorithms are mainly used in the analysis of text data.In [10], changes in people's behavior based on demographic characteristics were analyzed by selecting tokens using Naive Bayes (NB) and Support Vector Machine (SVM) classifiers.

In [11], an informative feature is determined for each audio recording and the effectiveness of feature classification is assessed. The assessment is carried out according to parameters including sound frequency, speech rate (number of

words expressed per unit of time), pitch contour, sound power rating, etc. Special software has been developed for voice stress analysis (VSA). Deep learning algorithms and graph theory were used to build the system.

In [12] the authors state that the excess of text data in various information systems forces researchers to analyze human behavior based on unstructured data. The main purpose of their research is to discuss the basic methods of predicting and identifying human behavior. The authors came to the conclusion that articles on the analysis of human behavior can be symmetrically divided into two classes. The first class contains articles in which the collection of unstructured text data for the analysis of human behavior is obtained from three sources: a list of surveys provided by the respondents themselves, (2) social networks and an assessment of the relationship between these two data sets. The assessment is carried out by applying correlation analysis. The second class focused on approaches based on labeled unstructured text data and on unlabeled unstructured text data. The study provides a complete understanding of the work of intelligent algorithms designed to analyze unstructured text data to predict human behavior.

In [13] the authors argue that thanks to the technological revolution, researchers can link the daily use of words to analyze human behavior. This article examines several text analysis methods and describes how the Linguistic Inquiry and Word Count (LIWC) program works. LIWC is a program that counts and classifies words into psychologically meaningful categories. LIWC is capable of detecting hidden knowledge across a wide variety of experimental conditions, including emotionality, social relationships, attentional focus, thinking styles, and human interests.

As we can see, there are many different approaches to analyzing behavior. However, due to the expansion of cyber-physical systems and the emergence of “big data”, there is a need to continue this research.

Documents and personal data that reflect human behavior can be obtained from various sources. Such sources include state registers, tracking devices, help systems, mobile devices, social networks, the Internet of Things, etc. Modern technologies make it possible to connect all possible state registers in the shortest possible time and collect data into one database.

3. Popular Algorithms for Analyzing Human Behavior

Analysts can choose the most appropriate tool to solve problems by understanding the strengths and weaknesses of different algorithms. It is important to consider not only the algorithm's accuracy, but also interpretability, scalability, and suitability for the specific data and problem being solved. This section discusses several well-known ML algorithms used to analyze human behavior and its interests based on text data:

1. K-Nearest Neighbors (KNN): KNN is widely used for classification and regression problems. The simplicity and efficiency of KNN make it a popular choice for both beginners and experienced practitioners [5]. KNN uses a distance metric (Euclidean distance, Manhattan distance, etc.) to determine nearest neighbors. KNN can process both discrete and continuous data, making it suitable for analyzing a wider range of human behavior data, including numerical characteristics such as purchase amounts or time spent on a website, behavioral anomaly detection, image recognition, spam filtering, and medical diagnostics. Its nonparametric nature allows it to adapt to complex data distributions without making any prior assumptions. KNN is relatively robust to noise and outliers in the data because it relies on aggregate information from multiple neighbors and can achieve high classification accuracy, especially when the data is well represented in the training set. But, as the number of features (dimensions) in the data increases, the distance metric becomes less meaningful, potentially leading to poor classification performance.

2. Support Vector Machine (SVM): SVM algorithms excel at finding hyperplanes that best separate data points belonging to different classes [7]. They are robust to outliers and efficient for high-dimensional data. SVMs can analyze a wide variety of data types, making them a versatile tool. SVM algorithms are particularly good at processing high-dimensional data, which refers to data with a large number of features. Because they focus on identifying the most important hyperplane separating data points, regardless of dimension. This makes them suitable for tasks such as text classification and image feature analysis. SVM is successfully applied to both linearly separable and nonlinearly separable data. SVM usually demonstrates higher classification accuracy on complex data, especially in the presence of nonlinear dependencies. The choice of kernel allows SVM to be adapted to different types of data and tasks. SVMs may be less susceptible to overfitting than deep neural networks and require fewer computational resources. However, SVM algorithms can be computationally expensive for large data sets and provide limited interpretability.

3. Decision Trees (TD): These tree structures classify data points based on a series of sequential decisions [6, 7]. They are interpretable, process categorical and numerical data well, and can identify important features that influence behavior. TD algorithms start with a root node representing the entire data set. Subsequent nodes represent splits on the independent variables, resulting in leaf nodes that assign data points to the corresponding classes.

4. Naive Bayes (NB): It is a simple and efficient classification algorithm based on Bayes' theorem [8]. It assumes conditional independence of features and works by calculating the probability of an object belonging to each class based on its features. Naive Bayes is well suited for analyzing categorical data often found in human behavioral datasets, such as demographics, online activity patterns, and sensor data. Naive Bayes is efficient for learning and prediction, allowing you to quickly analyze large text datasets. Its simplicity and efficiency make it a popular choice for exploratory analysis and classification tasks.

5. Convolutional Neural Networks (CNN): These powerful algorithms excel at big data analysis [10]. They can be used to analyze human behavior based on visual and text data. For example, solving problems with facial recognition, user facial expressions during video calls, or identifying areas of interest in a virtual environment. However, CNN algorithms require significant amounts of labeled data for training and can be computationally expensive.

6. Logistic Regression (LR): This fundamental and popular statistical technique is excellent for classification problems [11]. It can analyze user data to predict the likelihood that a certain behavior will occur (such as purchasing a product, subscribing to a service, etc.). LR estimates the probability of a binary outcome based on a set of independent variables (for example, demographic data, online behavior data, etc.). LR does not parse text or images directly. However, these types of data can be converted into numerical characteristics (for example, word frequency for text or image pixel intensity for images).

4. Methodology for Analysis of Human Behavior

The choice of the most suitable ML algorithm for analyzing human behavior depends on the following factors:

• *Data type:* text, image, video files, sensor data, etc.
• *Objective of the task:* classification, regression, clustering, prediction or anomaly detection determines the appropriate category of the algorithm.
• *Data size and complexity:* It is important to consider computational efficiency and scalability for large data sets.
• *Computational resources:* It is important to consider the hardware requirements of the methods.

To analyze human behavior, we use documents and personal data of employees of the Institute of Information Technologies of the Republic of Azerbaijan. We take documents reflecting employee behavior from three sources:

1) *Demographic data.* Public and private institutions collect demographic data that can be used to predict social processes, analyze citizen behavior, and effectively manage enterprises and resources. For our experiment, we can obtain employee demographic data from the ict.az website, where it is publicly available. It should be noted that this site does not contain all the data about employees. For the experiment, we need data: year of birth, place of study, and specialty. With the consent of employees, we will use the personal data of these employees to analyze people's behavior.
2) *Scientific articles by employees.* Scientific articles by employees stored in the institute's electronic library are in the public domain. Analysis of these articles can provide information about the quality and productivity of employees. In addition, ML algorithms gain knowledge about the psychological state, interests, and moods of employees.
3) *Data on social networks*. Tracking user interactions on social networks, such as time spent in certain areas of the site, topic and length of comments written by him, etc. can reveal preferences and patterns. All major open social networks such as Twitter, Facebook, etc. provide their data to the online community through their Application Programming Interfaces (APIs). API plays a key role in the collection, delivery to users and analysis of personal data [14]. Using the API, we can identify your friends' friends, including their profile pages, the photos and comments they enter, and even their phone numbers. Like any other software product, a modern API has its software development lifecycle, including design, testing, creation, and versioning. The purpose of mining employee text data is to determine how important these employees are to the organization. Text data analysis using ML (text mining) consists of four sequential stages:

• text pre-processing and data indexing;
• vector representation of text data;
• building and training a classifier using ML methods;
• assessment of the quality of classification.

5. Pre-processing and Data Indexing

Pre-processing of the text includes tokenization, and removal of functional words (semantically neutral words such as conjunctions, prepositions, articles, etc.). Next, morphological analysis is carried out (marking by parts of speech and lemmatization are carried out). This allows you to significantly reduce the size of space.

The main stages of preprocessing:

1. Data collection: All available data sources (databases, files, API, etc.) are identified. Data is exported into a format convenient for analysis (e.g. CSV, Excel).
2. Data cleaning: The collected information is subjected to tokenization, de-duplication, lemmatization, and stop word removal. Tokenization means breaking the text into individual words (tokens). This simplifies the text analysis process. Since each word becomes a separate unit that can be analyzed. Lemmatization is the process of reducing a word to its lemma, that is, its dictionary form. The same word can have several lemmas depending on the context. Stop word removal is a basic stage of text data preprocessing, which involves

excluding the most frequently occurring words from the text that do not carry a significant semantic load (for example, "and", "in", "on").

3. Integration into a single format: Data is brought to a common scale and format (e.g. normalization, standardization): all texts are brought to a single representation, which is necessary for the operation of ML algorithms.

Document indexing is the construction of a numerical text model that translates the text into a representation convenient for further processing.

6. Vector Representation of Text Data

Most text mining tasks, especially text classification and clustering, require a vector representation for each document.To analyze text data, we need to convert the text into a numeric representation that the algorithm can understand [7, 15]. Common approaches include:

- *Bag of Words (BoW):* Represents each document as a word count vector, ignoring word order [16].
- *N-grams:* Considers sequences of words (n-grams) to account for local context and word order.
- *Term Frequency - Inverse Document Frequency (TF-IDF):* The TF-İDF algorithm allows you to evaluate the importance of words in a document, determine the weight of words, etc. The TF-IDF algorithm is mainly used for large text data sets. It takes into account both the frequency of occurrence of a word in a specific document (term frequency) and its overall frequency across all documents in the corpus (inverse document frequency) [17].

The vector model is the most widely used model for representing texts. Let's pretend that:
$D = \{d_1, d_2, ..., d_i\}$ – sets of documents and a vector model of each document d_i can be expressed as follows:

$$d_i = [\omega_{i1}, \omega_{i2}, ..., \omega_{ij}] \quad i = 1, ..., n.$$

where, ω_{ij} – is the weight of word t_j in document d_i.

The weight of a word in the text t_j is determined as follows:

$$\omega_{ij} = \mathrm{TF}_{ij} \times \mathrm{IDF}_j, \; i = 1, ..., n, \; j = 1, ..., m, \tag{1}$$

where, TF_{ij} – frequency of use of the word t_j in the text

$$\mathrm{TF}_{ij} = \frac{m_{ij}}{m_i}; \; i = 1, ..., n, \; j = 1, ..., m \tag{2}$$

where, m_i – total number of documents, m_{ij} – number of terms in the text. The significance of terms is measured as follows:

$$\mathrm{IDF}_j = \log(n / n_j) \tag{3}$$

where, n_j is the document number with word t_j, and n is the total number of words. After this description, it becomes clear that metrics must be given to determine the proximity between two vectors. The weight w_{ij} is calculated by the TF-IDF scheme:

$$\omega_{ij} = \frac{m_{ij}}{m_i} \times \log \frac{n}{n_j}; \; i = 1, ..., n \tag{4}$$

There are different metrics for calculating the proximity between two vectors: cosine metric, Euclidean distance, etc. A metric is defined to determine the proximity between two vectors. We use the cosine metric to determine the semantic similarity of words.

Based on this metric, the proximity of vectors $\mathrm{d}_i = [\omega_{i1}, \omega_{i2}, .. \omega_{ij}]$ and $\mathrm{d}_l = [\omega_{l1}, \omega_{l2}, .. \omega_{lj}]$ is calculated as follows:

$$sim(\mathrm{d}_i, \mathrm{d}_l) = \frac{\sum_{j=1}^{m} \omega_{ij} \omega_{lj}}{\sqrt{\sum_{j=1}^{m} \omega_{ij}^2 \cdot \sum_{j=1}^{m} \omega_{lj}^2}}, \quad i,l = 1,2...,n \tag{5}$$

where ω_{ij} and ω_{lj} are the j -th elements of the frequency of occurrence of the compared vectors.

After text processing is completed, you can proceed directly to the application of machine learning algorithms. In other words, after calculating the proximity between the vectors, the classification and prediction phase begins using a support vector machine (SVM) algorithm. SVM algorithms excel at finding hyperplanes that best separate data points belonging to different classes [18]. They are robust to outliers and efficient for high-dimensional data. SVMs provide efficient, more accurate, and faster learning of text classifiers from examples.

7. Classification of Documents Using SVM

The SVM algorithm is designed in such a way that it searches for points closest to the hyperplane. These points are called support vectors. Next, the algorithm calculates the distance between the support vectors and the hyperplane. This distance is called the margin [19].

SVM is trained on datasets where text samples are labeled with specific behavioral categories (e.g., positive sentiment, aggressive language, term). By analyzing these labeled examples, SVM learns to identify features and patterns within the text that distinguish different behavioral categories. Once trained, SVM can classify new, unseen text data points. Based on the learned features, the SVM predicts the most likely behavioral category for a new text sample. SVM creates an optimal hyperplane that separates data points in a high-dimensional space S. Classifying documents by topic using SVM involves dividing the semantic space by a hyperplane. In the input space, a hyperplane separates documents related to different topics. We can represent any pair of parallel hyperplanes as follows:

$$\omega^t x_i + \omega_0 = \pm 1 \tag{6}$$

where, input data is marked as $x_i \in R^d$. ω – is the weight vector of the hyperplane, t – is the training dataset.

The mathematical description of SVM is expressed by defining the hyperplane:

$$\frac{1}{2}\left|\omega^2\right| + \mathrm{C}\sum_{\mathrm{i}=1}^{N} \zeta_i \tag{7}$$

where C – is a preset tuning parameter and evaluates the fit errors in the classification function. ζ_i – classification error and it can be written as a vector: $\xi_i = (\xi_1, .. \xi_N)^t$.

As we have already said, the primary task of SVM maximizes the margin between two classes and also minimizes the training error ζ_i. For this, we first solve the optimization problem:

$$\min_{\omega,b,\zeta} imize\ f(\omega,b,\zeta) \equiv \frac{1}{2}\omega^t\omega + C(\sum_{i=1}^{N}\zeta_i) \tag{8}$$

$$\text{Subject to: } \mathrm{y}_\mathrm{i}\ (\omega^{\mathrm{t}}\phi(x_i) + b) \geq 1 - \zeta_i,$$
$$\zeta_i \geq 0\ ,\ 1 \leq i \leq N,$$

where $\varphi(x_i)$ – are nonlinear input data in feature space: $x = (x_1, x_2, ..., x_i)\ \alpha\ \ \phi(x) = (\phi(x_1), \phi(x_2), ..., \phi(x_i))$

b – the bias in the feature space S. Is a shift factor that allows changing the decision boundary, $\omega^t\varphi(x_i) + b$ – boundary loss function defining the hyperplane. $\sum_{i=1}^{N}\zeta_i$ – is the sum of deviations of incorrectly classified objects and considered empirical. Introducing Lagrange multipliers *a*, *b* quadratic programming is defined as follows:

$$L(\omega,b,\zeta;\alpha,b)=\frac{1}{2}\omega^t\omega+C\sum_{i=1}^{N}\zeta_i-\sum_{i=1}^{N}\alpha_i\left[y_i(\omega^t x_i-b)+\zeta_i-1\right]-\sum_{i=1}^{N}\mu_i\zeta_i \tag{9}$$
$$0\le\alpha\le C,\ i=1,..N\ ,\quad \alpha=(\alpha_1,...,\alpha_N)$$

The active binding constraints determine the parameters of the SVM hyperplane. They correspond to a support vector, which are non-zero Lagrange multipliers. The performance of the trained model is estimated on a separate set of control data. Data intended for analysis are divided into training and testing data. An SVM classifier is then generated that matches the training data and classification is performed with regularization parameter C=1. SVM learns from labeled data sets where past behavior is categorized into specific outcomes. This allows SVM to identify patterns and relationships between features (demographics, social media activity, and science articles).

We use Python and sci-kit-learn libraries for classification. The library consists of 98 tokens, including employee demographics, social media data (time spent online and the content of comments they leave), topics of scientific articles and the scientific databases in which they are published, and other categories. Here, the "Demographic Data" parameters display information from the electronic demographics system – about the age, education, and places of work of employees. The experiments were carried out using the "Document Subject" parameter. The database defines categories of areas of interest of employees, and the "useful employee" category includes employees whose scientific articles are included in authoritative scientific databases (WoS, Scopus, Google Scholar, CrossRef), giving preference to the topic of information technology. The Helpful Employee categories are defined as follows: Very Helpful Employee, Helpful Employee, Not So Helpful Employee, Unhelpful Employee.

The classifier SVM has the functions of initialization "init()", training "fit()", prediction "predict()", finding the loss function "hinge_loss()" and finding the general loss function of the classical algorithm with a soft margin "soft_margin_loss()". For each epoch of the training set (X_train, Y_train) we will take one element from the sample and calculate the gap between this element and the position of the hyperplane at a given time. Depending on the size of the gap, we will change the weights of the algorithm using the gradient of the loss function ζ_i.

As can be seen, the SVM classifier determines the area of interest of employees by dividing the data into classes through a hyperplane. SVM also offers a powerful framework for predicting human behavior. Given the relevance of the topic, predicting human behavior will be considered in further research.

8. Comparative Analysis of Machine Learning Algorithms

Accuracy, Precision, Recall, and F1 score metrics provide a robust estimate of SVM performance and ensure that the model is not overfitted. To confirm the effectiveness of cross-validation in preventing overfitting, we conducted experiments on several benchmark datasets with different class distributions. The formula F1 score used in these calculations is given below:

$$F1\,score=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{10}$$

The experiment on the comparative analysis of ML algorithms was carried out in the Python software package. The algorithms were evaluated using the metrics Accuracy, Precision, Recall, F1-score. The results of the analysis are shown in Table 1.

Table 1. Results of comparative analysis of ML algorithms

no	Algorithms	Accuracy	Precision	Recall	F1-score
1.	**K-NN**	0.949	0.965	0.934	0.948
2.	**SVM**	0.953	0.988	0.911	0.949
3.	**DT**	0.931	0.952	0.895	0.931
4.	**NB**	0.837	0.804	0.938	0.857
5.	**CNN**	0.942	0.937	0.934	0.937
6.	**LR**	0.957	0.984	0.928	0.955

Based on the results of a comparative analysis of ML algorithms (Table 1), we can say that SVM, LR and KNN demonstrate the highest overall accuracy: 0,953, 0,957 и 0,949. The NB classifier performed the lowest with a precision of 0.804, but has the highest recall rate (0.938), meaning it correctly identifies a large proportion of real positive cases. LR and SVM achieve the highest F1 scores (0.955 and 0.949), providing a balance of precision and recall for comprehensive performance measurement. A comparative analysis of MO algorithms has proven that the most suitable algorithms for analyzing human behavior are SVM, CNN, and LR with an accuracy of 0.984.

By effectively leveraging user data, now known as big data, and carefully creating intelligent models accordingly, valuable insights can be gained for a variety of applications. However, limitations such as linearity assumptions and overfitting must be taken into account.

9. Conclusion

It is impossible to study and predict human behavior using traditional statistical methods, where information technology is rapidly developing and data about people is increasing every minute. ML algorithms have become powerful tools for analyzing vast amounts of data collected through sensors, social media, and other digital sources. By understanding the strengths and weaknesses of ML algorithms, you can choose the most suitable tool for your specific needs. It is important to consider not only the algorithm's accuracy, but also interpretability, scalability, and suitability for the specific data and problem being solved.

Based on learned patterns, SVM can classify new, unseen data points and allow you to understand and predict human behavior. While SVM are powerful, their internal workings can be complex, making it challenging to understand how they arrive at specific behavioral classifications. But despite these challenges, the study proved that SVM can revolutionize our understanding of human behavior by analyzing huge amounts of text data.

A four-step machine learning approach including preprocessing, vectorization, classifier training, and quality assessment yielded good results in text data analysis. In summary, this study demonstrates the potential of machine learning to effectively analyze and predict human behavior using a variety of data sources. SVMs often achieve high classification accuracy, especially with well-structured datasets and effective feature representation. The quality and representativeness of the training data are crucial for SVM performance. Biased or incomplete data can lead to inaccurate classifications.

References

[1] Claudio Bosco, Grubanov-Boskovic Sara, Iacus M. Stefano, Minora Umberto, Sermi Francesco, Spyridon Spyratos. “Data Innovation in Demography, Migration and Human Mobility”, European Union, Luxembourg, 144 p., 2022. DOI: 10.2760/027157, JRC127369

[2] Reza Zafarani, Huan Liu. “Behavior Analysis in Social Media”, IEEE Intelligent Systems, Vol.29, No.4, pp.1–4, 2014. https://reza.zafarani.net/publications/files/BehaviorAnalysis.pdf

[3] Barbara Calabrese, Mario Cannataro, Nicola Ielpo. “Using Social Networks Data for Behavior and Sentiment Analysis”, Proceedings of the 8th International Conference on Internet and Distributed Computing Systems, Vol.9258, pp.285–293, September 2015. DOI: 10.1007/978-3-319-23237-9_25

[4] Vasile P. Bresfelean. “Analysis and predictions on students’ behavior using decision trees in Weka environment”, Proceedings of the ITI 29th International Conference on Information Technology Interfaces, Cavtat, Croatia, pp.1–6, 25–28 June 2007. DOI: 10.1109/ITI.2007.4283743

[5] Dharmar Selvithi. “FPGA Based Human Fatigue and Drowsiness Detection System Using Deep Neural Network for Vehicle Drivers in Road Accident Avoidance System”, Learning and Analytics in Intelligent Systems, Vol.6, pp.69–91. 2020. DOI:

[6] 10.1007/978-3-030-35139-7_4

[7] Byeongho Heo, Kimin Yun, Jin Y. Choi. “Appearance and motion based deep learning architecture for moving object detection in moving camera”. Proceedings of the IEEE International Conference on Image Processing, Beijing, China, pp.1827–1831, 17-20 September, 2017. DOI: 10.1109/ICIP.2017.8296597

[8] Rasim Alguliyev, Ramiz Aliguliyev, Irada Alakbarova. “Extraction of hidden social networks from wiki-environment involved in information conflict”, International Journal of Intelligent Systems and Applications, Vol.8, No.2, pp.20–27, 2016, DOI: 10.5815/ijisa.2016.02.03

[9] Mahmudul H. Suzan, Nishat A. Samrin, Ali A. Biswas, Aktaruzzaman Pramanik. “Students’ Adaptability Level Prediction in Online Education using Machine Learning Approaches”. Proceedings of the 12th International Conference on Computing Communication and Networking Technologies, Kharagpur, India, pp.1–5, 6–8 July, 2021, DOI: 10.1109/ICCCNT51525.2021.9579741

[10] Xiaofeng Ma, YanYang. Zhurong. Zhou. “Using machine learning algorithm topredict student pass rates in online education”, Proceedings of the 3rd International Conference on Multimedia Systems and Signal Processing, New York, pp.156–161, 28 April 2018. DOI: 10.1145/3220162.322018

[11] Trstenjak Bruno, Dzenana Đonko. “Determining the impact of demographic features in predicting student success in Croatia”, Proceedings of the 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, Croatia, pp. 1222–1227, 26–30 May 2014, DOI: 10.1109/MIPRO.2014.6859754

[12] Kelly R. Damphousse. “Voice Stress Analysis: Only 15 Percent of Lies About Drug Use Detected in Field Test”, National İnstitute of Justice Journal, No.259, pp.8–12, 2008.

[13] Mohammad R. Davahli, Waldemar Karwowski, Elgar Gutierrez-Franco, Krzysztof Flok, Grzegorz Wrobel, RedhaTalar, Tareq Ahram. “Identification and Prediction of Human Behavior through Mining of Unstructured Textual Data”, Symmetry, Vol.12, No.1902, pp.1–23, 2020, DOI: 10.3390/sym12111902

[14] Yla R. Tausczik, James Pennebaker. “The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods”, Journal of Language and Social Psychology, Vol.29, No.1, pp.24–54. 2010. DOI: 10.1177/0261927X09351676

[15] Fernando N. der Vlist, Anne Helmond, Marcus Burkhardt, Tatjana Seitz. “API Governance: The Case of Facebook’s Evolution”. Social Media + Societ, pp.1–24, 2022. DOI: 10.1177/20563051221086228

[16] Zihang Lin, Yuwei Zhang, Qingyuan. Gong, Yang. Chen, Atte Oksanen, Aaron Ding. “Structural Hole Theory in Social Network Analysis: A Review”, IEEE Transactions on Computational Social Systems, Vol.9, No.3, pp.724–739, 2022, DOI: 10.1109/TCSS.2021.3070321
[17] Md. Abdur Rahman, Abu Nayem, Mahfida Amjad, Md. Saeed Siddik. "How do Machine Learning Algorithms Effectively Classify Toxic Comments? An Empirical Analysis", International Journal of Intelligent Systems and Applications (IJISA), Vol.15, No.4, pp.1–14, 2023. DOI:10.5815/ijisa.2023.04.01
[18] Lin Xiang. “Application of an Improved TF-IDF Method in Literary Text Classification”, Hindawi: Advances in Multimedia, Vol.2022, pp.1–10, 2022, DOI: 10.1155/2022/9285324
[19] Rahul Khanna, Marie Awad. “Efficient Learning Machines: Theories, Consepts, and Applications for Engineers and System Designers”. New York: Apress. 263 p. 2015.
[20] Raquel Rodr ́guez-Pérez, Jürgen Bajorath. “Evolution of Support Vector Machine and Regression Modeling in Chemoinformatics and Drug Discovery”, Journal of Computer-Aided Molecular Design, Vol.36, No.5, pp.1-8, 2022. DOI: 10.1007/s10822-022-00442-9.

Authors’ Profiles

Irada Alakbarova in 1984 she graduated from the Faculty of Automation of production processes, Azerbaijan Institute of Oil and Chemistry named after M. Azizbayov. In the same year, she was accepted for employment at the Institute of Information Technology. In currently holds the post of Sector Chief of the Institute of Information Technology. In 2018, the defense of the dissertation on the “Development of methods and algorithms for analysis of information war technologies in a wiki environment” and she received his Ph.D. (2018). In currently conducts research in the field of Social Network Analysis, Text Analysis, Clustering, Social Credit Analysis, and Big Data Analytics. She is the author of 50 articles and three books.

How to cite this paper: Irada Alakbarova, "Analysis of Human Behavior and Interests Based on Text Data", International Journal of Education and Management Engineering (IJEME), Vol.15, No.1, pp. 1-9, 2025. DOI:10.5815/ijeme.2025.01.01