INTERNATIONAL JOURNAL OF eBUSINESS AND eGOVERNMENT STUDIES
Vol: 15 **No:** 1 **Year:** 2023 **ISSN:** 2146-0744 (Online) (**pp.** 324-341) **Doi:** 10.34109/ijebeg.2023150115
Received: 19.12.2022 | **Accepted:** 25.04.2023 | **Published Online:** 10.05.2023

-RESEARCH ARTICLE-

A NEW MODEL FOR CREATING DIGITAL AVATARS OF APPLICANTS BASED ON SOCIAL MEDIA DATA

Dmitriy Rodionov

Graduate School of Industrial Economics, Institute of Industrial Management, Economics and Trade, Peter the Great St. Petersburg Polytechnic University, Saint Petersburg, Russia
Email: drodionov@spbstu.ru

Polina Pashinina

Graduate School of Industrial Economics, Institute of Industrial Management, Economics and Trade, Peter the Great St.Petersburg Polytechnic University, Saint Petersburg, Russia
Email: pashinina_pa@spbstu.ru

Irina Smirnova

Graduate School of Industrial Economics, Institute of Industrial Management, Economics and Trade, Peter the Great St.Petersburg Polytechnic University, Saint Petersburg, Russia
Email: ir_almazova@mail.ru

Evgenii Konnikov

Graduate School of Industrial Economics, Institute of Industrial Management, Economics and Trade, Peter the Great St.Petersburg Polytechnic University, Saint Petersburg, Russia
Email: konnikov.evgeniy@gmail.com

─Abstract─

A thorough digital picture of our world has been created as a result of the information environment's rapid growth, which has completely changed how we view and interact with reality. This revolutionary advancement has created opportunities for precise

Citation (APA): Rodionov, D., Pashinina, P., Smirnova, I., Konnikov, E. (2023). A New Model For Creating Digital Avatars Of Applicants Based On Social Media Data. *International Journal of eBusiness and eGovernment Studies, 15*(1), 324-341. doi: 10.34111/ijebeg.2023150115

Vol: 15 **No:** 1 **Year:** 2023 **ISSN:** 2146-0744 (Online) (**pp.** 324-341) **Doi:** 10.34109/ijebeg.2023150115

quantitative evaluation of the online social milieu. In this study, we sought to learn more about the online information background of college students who actively participate in extracurricular and academic activities regarded as very valuable by their individual institutions. To accomplish our goal, we created a sophisticated algorithm that can evaluate how closely an applicant's information background vector resembles that of the current student group. We want to uncover possible university candidate accounts on the VK social network by using this cutting-edge analytic technology, thereby improving the effectiveness of the admissions process. In addition to being a priceless tool for evaluating the digital world, our mathematical approach to processing internet data shows potential for reaching other socially relevant goals. This approach unlocks novel opportunities for creating tools that can effectively evaluate the digital landscape, offering substantial benefits across diverse fields. By leveraging the power of mathematical evaluation, we can make informed decisions, gain insights, and drive progress in this increasingly interconnected world. In addition, this mathematical evaluation could also lead to informed decision-making, drive progress, and facilitate advancements in a wide range of socially significant endeavors within the rapidly evolving information environment.

Keywords: digital avatar, virtual profile, digital environment, social media, text analysis, LDA.

1. INTRODUCTION

In the current dynamic environment, fast advancements in technology and the growing influence of social media platforms have transformed the way recruitment processes are conducted. Traditional resume-based evaluations often fall short in capturing the multidimensional aspects of an applicant's personality and character. In response to this limitation, a new paradigm is emerging that leverages social media data to create digital avatars of applicants, providing a holistic and nuanced representation of their online presence (Kaufman & Horton, 2014). In this dire environment, university recruitment companies usually focus on a broad demographic cohort, typically defined by the school year. At the same time, not every graduate is predisposed to participate in university studies actively. For educational institutions, the rationalization of marketing strategies for attracting applicants to improve the informational quality of the educational offer and increase the costs of recruiting such students whose educational efficiency is consistent with the reality of the university remains relevant (Jing et al., 2023). Instead of spreading generalized information blocks, the university can target potential applicants with information about educational programs, admission requirements, and study benefits. Such optimization is beneficial to the educational institution, allowing it to attract active partners who are ready to work productively. Ultimately, this is the leading indicator of the university's scientific development, the quality of education, and the degree of achievement of the directed social orientation (Back et al., 2010; Ullah et al., 2023).

Vol: 15 **No:** 1 **Year:** 2023 **ISSN:** 2146-0744 (Online) (**pp.** 324-341) **Doi:** 10.34109/ijebeg.2023150115

In this case, criteria need to be identified to categorize people as potential applicants based on an obvious solution to this issue is to analyze current students who meet the expectations of an active and highly potential student to identify common characteristics that form the desired informational background the university (Bonnema & Van der Waldt, 2008). Such analysis of students and later potential applicants should be automated with many participants in mind. The most mechanized, audience-oriented, and least resource-demanding approach, in this case, can be an analysis of some ready-made textual information.

Therefore, the aim is to develop a mathematical tool for assessing the potential applicant's suitability in terms of interests and personal characteristics obtained by analyzing public data, the informational environment of an active group of Polytechnic University students, which would be a summary of the need for targeted recruitment of this person to participate in the admission campaign. Therefore, this paper presents a novel model for generating digital avatars by integrating various data sources from social media platforms. These avatars serve as visual and interactive representations of an applicant's digital identity, encompassing their interests, social connections, online behavior, and communication style (Chambers, 2013). This strategy seeks to improve the hiring process by providing a more thorough and accurate assessment of an applicant's fit inside a business by utilizing the large quantity of information readily available on social media.

Both theoretically and practically, it is important to create a new paradigm for digital candidate avatars based on social media data (Zeng et al., 2010). Theoretically, through leveraging the developing disciplines of digital identity and human-computer interaction, this technique increases our understanding of how social media data may be utilized to generate virtual representations of individuals. By analyzing the nuances of social media accounts, including content created by users, preferences, as well as interactions, this model contributes to the theoretical foundations of individualized avatars and gives light on the difficulties of translating real-world identities into virtual environments. Practically speaking, this objective has wide-ranging effects across multiple disciplines. Digital avatars can provide a more thorough review of applicants for jobs and recruiting, giving businesses access to information beyond traditional resumes. Additionally, in gaming and virtual reality applications, such avatars enhance immersion and customization, providing users with a more authentic and tailored experience. Moreover, digital avatars can facilitate personalized marketing strategies, enabling businesses to better understand their target audience and deliver highly targeted advertisements (Miao et al., 2022). Overall, this objective merge theoretical advancements with practical applications, paving the way for innovative uses of social media data in avatar creation and significantly impacting diverse industries.

Vol: 15 **No:** 1 **Year:** 2023 **ISSN:** 2146-0744 (Online) (**pp.** 324-341) **Doi:** 10.34109/ijebeg.2023150115

The Previous discussion, established growing interest in leveraging social media data to create digital avatars for applicant assessment. By combining NLP, machine learning, and ethical considerations, researchers aim to develop an objective and reliable method to assess applicants' suitability and cultural fit. While promising results have been obtained, challenges such as ethical concerns, privacy protection, and bias mitigation still require careful consideration. As the field advances, it is crucial to maintain transparency and adhere to ethical standards to ensure fair and responsible implementation of this innovative model in recruitment processes. Therefore, this research has mainly concern on new model for creating digital avatars of applicants based on social media data. This model significance lies in its potential to streamline the applicant screening process, providing admissions teams with objective and quantifiable insights into candidates' personalities and online behavior. By identifying applicants whose digital avatars closely align with the ideal vectors, universities can potentially identify individuals who demonstrate positive social qualities, active engagement, and a high level of emotional intelligence. This can lead to more informed and efficient decision-making in the university admissions process. Based on previous discussion, this model significance could be ignored. The research was divided into five chapters, introduction, literature review, research methodology, data analysis and results, discussion and conclusion.

2. LITERATURE REVIEW

The digital environment increasingly and accurately reflects the social aspect of life (Ahmad et al., 2019). According to the Digital 2023 Global Overview Report (Kemp, 2023) the number of internet users at the beginning of 2023 was 5.16 billion, of which 4.76 billion used social networks. It was also calculated that the average internet user spends 6 hours 43 minutes in the digital environment daily – thus, (allowing 8 hours per day for sleep) internet interaction amounts to an average of 40% of a person's life (Twenge, Martin, & Spitzberg, 2019).

Accordingly, internet activity attracts more and more research interest every year. Most often, such activity in the adolescent-youth environment is studied. This is related, firstly, to the fact that youth form the most significant part of the internet audience (Leijse, Koning, & van den Eijnden, 2023). Secondly, the influence of the digital revolution on the psycho-emotional state of young people causes the most significant concern (Ali, 2022; binti Ab Aziz & Balraj, 2022; Levenson et al., 2016) as it coincides with the rapid physical, social, behavioral, and cognitive development typical of this group. In general, the digital trace left on social networks is a highly informative indicator of social experience, psychological characteristics, and personal qualities of a person (Sultan, Scholz, & van den Bos, 2023).

Vol: 15 **No:** 1 **Year:** 2023 **ISSN:** 2146-0744 (Online) (**pp.** 324-341) **Doi:** 10.34109/ijebeg.2023150115

As online reality improves the ways to process digital streams of information and interpret the results, the complexity and multifaceted nature of the virtual imprint of the social environment is understood and the methods for analyzing it are becoming more sophisticated, increasing the scientific effectiveness. From the differentiation of the studied control groups based on the user/non-user principle (Katz & Rice, 2002; Kuznetsov, Gorovoy, & Rodionov, 2021) and mathematical estimation of quantitative parameters of overall social activity on the internet (Steinfield, Ellison, & Lampe, 2008) researchers come to the necessity of dividing users by type, the purpose of use (Cummings, Butler, & Kraut, 2002; Van den Eijnden et al., 2008).

However, a number of works confirm that this may be insufficient (Khan et al., 2020) of high importance is the basic level of a person's psycho-emotional state that is, not only the goal with which he or she uses the social network but also how they relate to the social network, to the act of using and to their own goal in the end. Thus, for example, longitudinal (cohort) and cross-sectional analysis of the same data within the context of the internet will give different results (Kraut & Burke, 2015) cohort is preferable in terms of the accuracy of conclusions - cross-sectional does not take into account the initial predisposition to using the Internet. Longitudinal research methods also significantly improve mathematical analysis – despite the digitization of communications, they remain connected with social factors – they must therefore be processed about the peculiarities of social environment analysis (Mikami et al., 2010). In addition, the factor of active/passive behavior on social networks should be taken into account - a person is more of a consumer of information or a creator recently, the trend of expanding the use of social networks has catalyzed the transition of users from positions of passive information consumption (Web 1.0) to active creation of an information flow (Web 2.0) (O'reilly, 2007; Ozimek, Brailovskaia, & Bierhoff, 2023).

Research experience in virtual reality has shown that the most practical and effective are targeted studies of a clearly defined demographic cohort, with the possibility of a temporal lag in the cause-effect interaction (Jing et al., 2023). Therefore, generating a mathematical representation of the informational background of a strictly defined cohort of research users of a social network is significant in modern conditions (Prell, 2011). The interpretation of the data presented in a mathematical model can be used in social studies. Here, identifying an individual's personal qualities is of primary interest a connection has been established between the behavioral model of using a social network and the user's personality traits and personal qualities (Bastani, Namavari, & Shaffer, 2019). Moreover, the psycho-emotional direction of the information vector is usually understood intuitively when acquainted with a social account. It has been shown that strangers viewing an internet account can predict the level of the owner's extraversion and openness (Ellison, Steinfield, & Lampe, 2007).

Vol: 15 **No:** 1 **Year:** 2023 **ISSN:** 2146-0744 (Online) (**pp.** 324-341) **Doi:** 10.34109/ijebeg.2023150115

However, such empirical personal assessments do not allow for the processing of a large amount of information needed to create a digital model of behavior and a statistically reliable mathematical analysis of the information environment. Accordingly, the collection of analytical data should be automated mechanically and aimed at obtaining a tolerable level of trust for the amount of information processed. Social networks allow users to be clustered in various ways by subscriptions and interests, frequency and content of published information, age, gender, worldview, political and religious views, tastes, and more. One of the most common methods of analyzing internet activity is a study technically supported by traditional survey systems (Błachnio, Przepiorka, & Pantic, 2016). The principle of structuring information on the internet allows for the use of more modern technological methods of interaction with the information flow. The presented public textual information in the internet account profile can be processed, weighed, and evaluated by various topic modeling tools. There are various methods and procedures for text analysis, the main task of which is to extract significant information. However, not every methodology produces satisfactory results for the researcher regarding the level of trust (Steinfield, Ellison, & Lampe, 2008). The standard approach is content analysis – in which the output about some characteristics of the text array is formed based on its content via the description of syntactic and semantic communication features. Words that appear together are assigned to a common theme (Sultan, Scholz, & van den Bos, 2023). The content analysis itself is also varied in methods. They distinguish between thematic focused on studying the emergence of topics or terms, semantic related to the study of the interrelationship of topics within sentences and network aimed at extracting a network of interrelated topics – methods (Roberts, 2020). At the same time, they do not exclude each other and can be combined to improve the quality of the analysis.

Previous discussion shown that digital landscape has undergone a profound transformation with the advent of social media platforms, allowing individuals to construct and showcase their online identities. This proliferation of user generated content has opened up new possibilities for harnessing social media data to create digital avatars, which hold immense theoretical and practical significance. From a theoretical perspective, the emergence of this novel model intersects with the realms of internet studies and digital identity, shedding light on the intricate dynamics between online personas and real-world individuals. Exploring the theoretical underpinnings of digital avatars based on social media data not only expands our understanding of the complex interplay between technology and self-representation but also paves the way for innovative applications in numerous practical domains. By harnessing the vast amounts of user-generated content available on social media platforms, this new model offers practical implications for areas such as recruitment, gaming, virtual reality, and targeted marketing strategies.

Vol: 15 **No:** 1 **Year:** 2023 **ISSN:** 2146-0744 (Online) (**pp.** 324-341) **Doi:** 10.34109/ijebeg.2023150115

Seeking importance previously, when choosing one content analysis method, it is essential to consider the desired outcome. If the goal is to classify content automatically, a text clustering procedure must be implemented which is more efficiently done using network theory. When using a relation model based on word matches and simultaneously classifying two dimensions – topics and terms the results are more realistic and can be more accurately interpreted empirically (Celardo & Everett, 2020). One of the most effective techniques for semantic analysis is the application of Latent Dirichlet Allocation (LDA). The main advantages of LDA for evaluating a text corpus are, first, that this method does not require pre-processing of the data (e.g., annotation of every text block) – a semantic annotation and analysis of the source texts replace this. Secondly, the approach extracts hidden topics while assigning a probability model to each topic. Thirdly, the LDA method allows identifying several topics within one document, which is more in tune with the reality of the social field than accepting the condition that only one topic can be addressed in a single document – which is an inescapable part of the work with the LSA (Latent Semantic Analysis) tool, which is an effective method of reducing the dimensionality of term-document matrices using singular value decomposition and while it provides a noticeable compression of large datasets while preserving the structural basis, due to its one-topic focus, it may be ineffective in reflecting reality (Brier & Hopp, 2011).

The discussion in literature review is being highlighted the growing interest in leveraging social media data to create digital avatars for applicant assessment. By combining NLP, machine learning, and ethical considerations, researchers aim to develop an objective and reliable method to assess applicants' suitability and cultural fit. While promising results have been obtained, challenges such as ethical concerns, privacy protection, and bias mitigation still require careful consideration. As the field advances, it is crucial to maintain transparency and adhere to ethical standards to ensure fair and responsible implementation of this innovative model in recruitment processes. Therefore, this research has mainly concern on new model for creating digital avatars of applicants based on social media data.

3. MATERIALS AND METHODS

The LDA approach to topic modeling involves treating each document as a set of topics in a particular proportion. And each topic – as a set of keywords again, in a certain proportion. Once the algorithm is given the number of topics, all it does is map the distribution of topics in documents and the distribution of keywords within topics. A topic is nothing more than a set of dominant keywords. By simply looking at the keywords, one can determine what the topic is about and mathematically, this can be described as the probability distributions on words representing topics and probability distributions on topics representing documents. The selection of words and topics follows a distribution, and the entire data-generating process combines generative

Vol: 15 **No:** 1 **Year:** 2023 **ISSN:** 2146-0744 (Online) (**pp.** 324-341) **Doi:** 10.34109/ijebeg.2023150115

processes. The generative process itself describes the generation of each word in each document – in LDA following an algorithm (Srinivasa-Desikan, 2018).

– in every topic:
– vector selection ϕ_t by distribution $\phi_t \sim Dir(\beta)$
– in every document:
– vector selection $\theta_t \sim Dir(\alpha)$ – reflecting the proportion of the presence of the topic in the document
– for each word of document:
– topic selection z_w by distribution $z_w \sim \theta_d$
– word selection $w \sim p(w \mid z_w, \phi)$ with probability given by β

The distribution is modeled mathematically using the Latent Dirichlet Allocation (LDA), generated from a Dirichlet integral (in particular, the normalizing constant is drawn from the Dirichlet integral), which is a distribution on a (n – 1) dimensional vector (Yang & Li, 2015).

$p(x_1, \dots, x_{n-1} \mid a_1, \dots, a_n) = \frac{1}{B(\alpha)} \prod_{i=1}^{n} x_i^{\alpha_i - 1}$, where
$x_n = 1 - x_1 - \cdots - x_{n-1}, x_i > 0.$

Thus, we get a distribution over distributions – a vector that is decomposed into non-negative elements, whose total sum to 1, i.e. a discrete probability distribution (Yang & Li, 2015). The set of probability distributions can be decomposed in the following manner (Yang & Li, 2015).

$$p(w, z, \theta, \phi \mid \alpha, \beta) = \left[\prod_{t=1}^{T} p(\phi_t \mid \beta)\right]\left[\prod_{d \in D} (p(\theta_d \mid \alpha) \prod_{w \in d} p(z_w \mid \theta_d) p(w \mid \phi_{z_w}))\right]$$

Additionally, the main advantage of the LDA method is that the final distribution obtained is a conjugate prior for the distribution with a controllable probability. These qualities significantly simplify the calculation of posterior probabilities when using Bayes' theorem.

On input, the LDA mathematical model receives a text array of data. On output, it provides a set of existing themes in the text array with a probabilistic distribution of significant tokens. In this way, we get a descriptive system of the existing text space, with the identification of possible directions of digitized thought - correspondingly, each thought receives its own vector in the identified n-dimensional space (by n-number of topics). Tokens here are word-markers that have a semantic weight in certain model topics of the text array - that is, each token-word is attached to a theme and, encountered in thought, adds a certain weight to bring it closer to the ideal vector of its theme.

Vol: 15 **No:** 1 **Year:** 2023 **ISSN:** 2146-0744 (Online) (**pp.** 324-341) **Doi:** 10.34109/ijebeg.2023150115

Since the mathematical characteristics of topics are embedded directly into tokens, the text to be evaluated must be prepared for mechanical analysis in such a way that words that do not carry any semantic topic load do not have the opportunity - by virtue of the functions they perform - to end up in the range of significant ones, having a tendency to high frequency of occurrence. In other words, before the operation of the LDA model, the array of data must be brought into a certain order. Noise should be cleaned up - stop words, functionally necessary for making the text readable to humans. These are general words: prepositions, particles, adverbs, conjunctions, introductory words, pronouns, interjections, various symbols and punctuation marks; and dependent words, which have meaning only in combination with the main component (for example, name and patronymic - dependent words when used together with the main component - surname).

Stemming is also considered good practices in various evaluation methods - reducing a word to its semantic basis (cleansing from prefixes and suffixes) - where the semantic basis is not always equivalent to the morphological basis; and lemmatization - bringing the form of a word to the normal form, lemma (nominative case for nouns in Russian). These practices help to reduce the duplication of tokens in the array and, as a consequence, defocus the significance of tokens.

Consequently, the mathematics of analysis sets certain requirements for the text itself, which goes into processing - it should be a large text array reflecting, in a methodologically significant way, the object of study.

Within the framework of our research - when digitizing the informational background generated by the selected student group - such a meaningful array of data suitable for analysis can be obtained from the public information of student's internet accounts. This solution has a number of advantages: first, public internet information is created by an individual who takes responsibility to some degree, minimizing the interfering factor of spontaneity and randomness of the emergence of certain relationships when assessing the significance of the results. Secondly, the information collected in this manner is already logically structured at the initial stage, which significantly simplifies the tasks of clustering and cohorting objects of study. Thirdly, this information represents a sufficiently large set, covering the requirements of LDA analysis and creating the necessary level of trust in the results. Thus, the choice stops on the use of public data of student accounts in social networks.

Of the hundreds of social networks for our research, VKontakte was chosen. Besides being the most popular social network in Russia with the highest audience reach, flow of messages and number of authors, VK has a convenient special interface - VK API (Application Programming Interface) - allowing you to get information from the vk.com

Vol: 15 **No:** 1 **Year:** 2023 **ISSN:** 2146-0744 (Online) (**pp.** 324-341) **Doi:** 10.34109/ijebeg.2023150115

database using HTTP requests to a special server. That is, the process of parsing - automated information collection - is quite lenient and open.

After a full definition of the mathematical and technical components of the research, it is necessary to determine, from a social point of view, the control group of students who's public VK accounts will be used as ideal indicators. Following the logic of studying the significance of academic and extracurricular activities undertaken by students, the following classification characteristics were appointed:

- 90% or more marks of "excellent" for the academic year;
- attendance of more than 10 events organized by the university;
- representing the university at extracurricular events;
- 3 or more achievements in the fields of academic, research, social, cultural and sports activities.

That is, the research should cover the various aspects of student life. The developed method of conducting the research is presented in Figure 1.

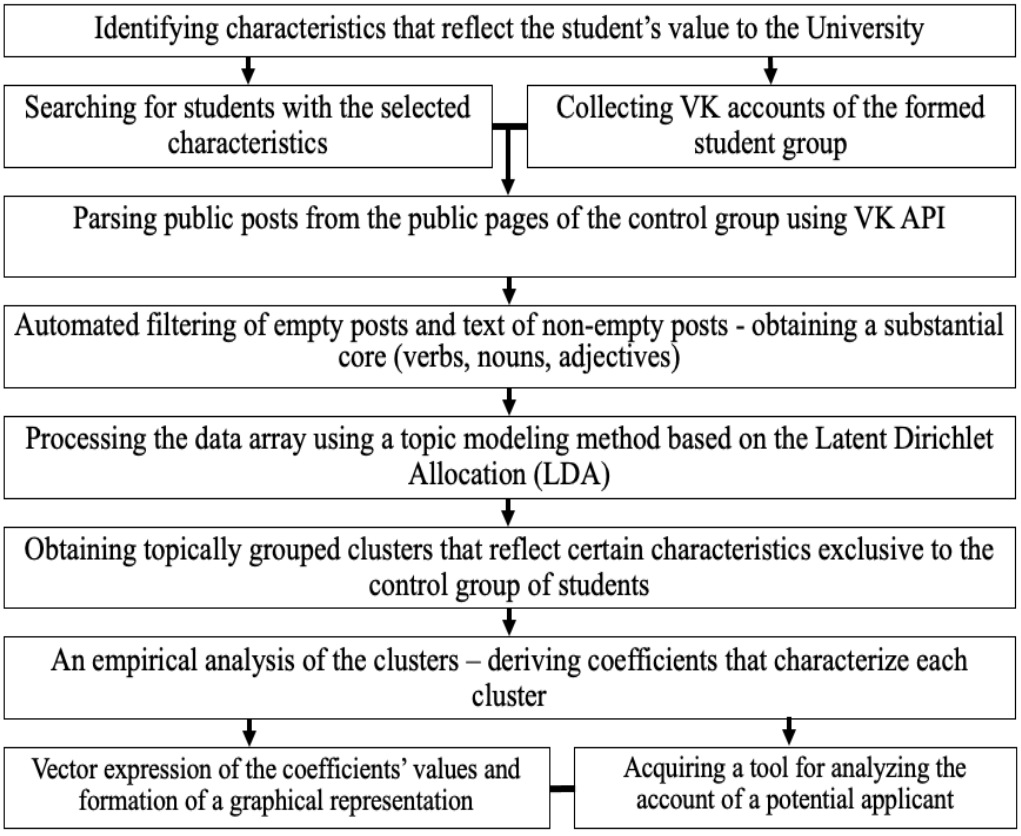

Figure 1 –Algorithm for conducting the study

Vol: 15 **No:** 1 **Year:** 2023 **ISSN:** 2146-0744 (Online) (**pp.** 324-341) **Doi:** 10.34109/ijebeg.2023150115

4. RESULTS AND DISCUSSION

According to the developed algorithm for the study, four thematic clusters were obtained from the output of the LDA analysis, clearly expressed in the processed text array, and the most significant tokens for each cluster. Empirically, each theme was given a name. The clusters obtained, their description and significant tokens are presented in Table 1.

Table 1 – Coefficients characterizing a certain cluster and the degree of presence of tokens from the student's profile in it

Coefficient	Tokens	Description
Emotional reaction coefficient	Ахаха (Ha ha), круто (cool), камон (let's do), чувства (feelings), лол (lol), блин (cowabunga), добро (good)	The coefficient reflects the ability and propensity of a person to express emotions. With a high informativeness of the public part of the user account, this coefficient is rarely lowered below average values. Its presence affects the degree of expression of the other coefficients, as it is an indicator of the person's willingness to share some information through posts in social networks.
Positively-colored reaction coefficient	Классные (great), спасибо (thank), огромное, удачно, какие, благодарю (I thank you very much), большое	The coefficient digitizes the emotional tint of the information flow created by the internet user. High values indicate the presence of a positive social interaction experience and the ability to maintain good will and openness in interpersonal relationships.
Sociality and activity coefficient	Едем (let's go), ребят (guys), можем (we can), должны (we must), завтра (tomorrow), сегодня (today), города (cities), днём (day)	The coefficient expresses the user's tendency towards event-filled activities and the reaction towards the social color of them. It is mainly represented by place and time tokens, as well as verbs in the plural - combining the factor of multilayeredness and the share of collectivization into a general system of activity manifestation.
Invasiveness of own emotions coefficient	Поздравляю (congratulations), хочу (I want), любимой (my beloved,), настроение (mood), лучшая (the best), самая (the most), жить (live), красиво (beautifully)	The coefficient reveals the author's skill in expressing approval towards the environment and its mastery of supportive expression functions.

Vol: 15 **No:** 1 **Year:** 2023 **ISSN:** 2146-0744 (Online) (**pp.** 324-341) **Doi:** 10.34109/ijebeg.2023150115

Each theme, with the distribution of tokens included in it, is presented in the form of an ideal thematic vector. Together, these vectors form a four-dimensional informational space, which mathematically reflects the set of positive personal qualities of Polytechnic University students, in terms of the effectiveness of scientific development and the achievement of social goals.

DISCUSSION

Thus, the share of the presence of a theme for each cluster is represented through the analysis of the public data of the control group of students accounts by ideal vectors. To form the corresponding vectors characterizing the informational content of the public VK accounts of the users to be verified for the ratio with the informational background of highly productive students of the Polytechnic University, the formula of the topic presence in the studied text is derived:

$\mathrm{K}_i = \frac{(\sum t_i \cap \mathrm{T})}{\mathrm{T}}$, where

K_i – share of thematic cluster in the news array (i – from 1 to 4, by cluster number);
t_i – a token belonging to the set forming the evaluated cluster;
T – the set of all tokens that form the news array.

The four vectors of correspondence of the informational background of a particular theme form a digital avatar of the applicant, which is then subjected to mathematical significance assessments - in the end, those users are of interest whose value of any coefficient individually exceeds the threshold value corresponding to the bright expression of the coefficient, or the value of the sum of coefficients exceeds a threshold value established for the sum.

As a demonstration of the consistency of the digital avatar of the applicant with the informational background of the university, a graphic interpretation of the vector presence of each thematic cluster in the posts to the Internet user's wall (Figure 2) is proposed.

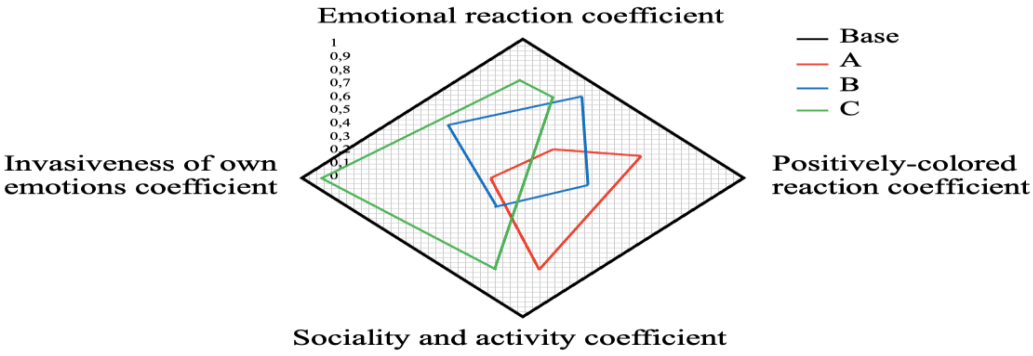

Figure 2: Graphical representation of possible ways to estimate the informational background of internet accounts

Vol: 15 **No:** 1 **Year:** 2023 **ISSN:** 2146-0744 (Online) (**pp.** 324-341) **Doi:** 10.34109/ijebeg.2023150115

The methodology for assessing an internet user to determine the rationality of attracting them to university education is presented in Figure 3.

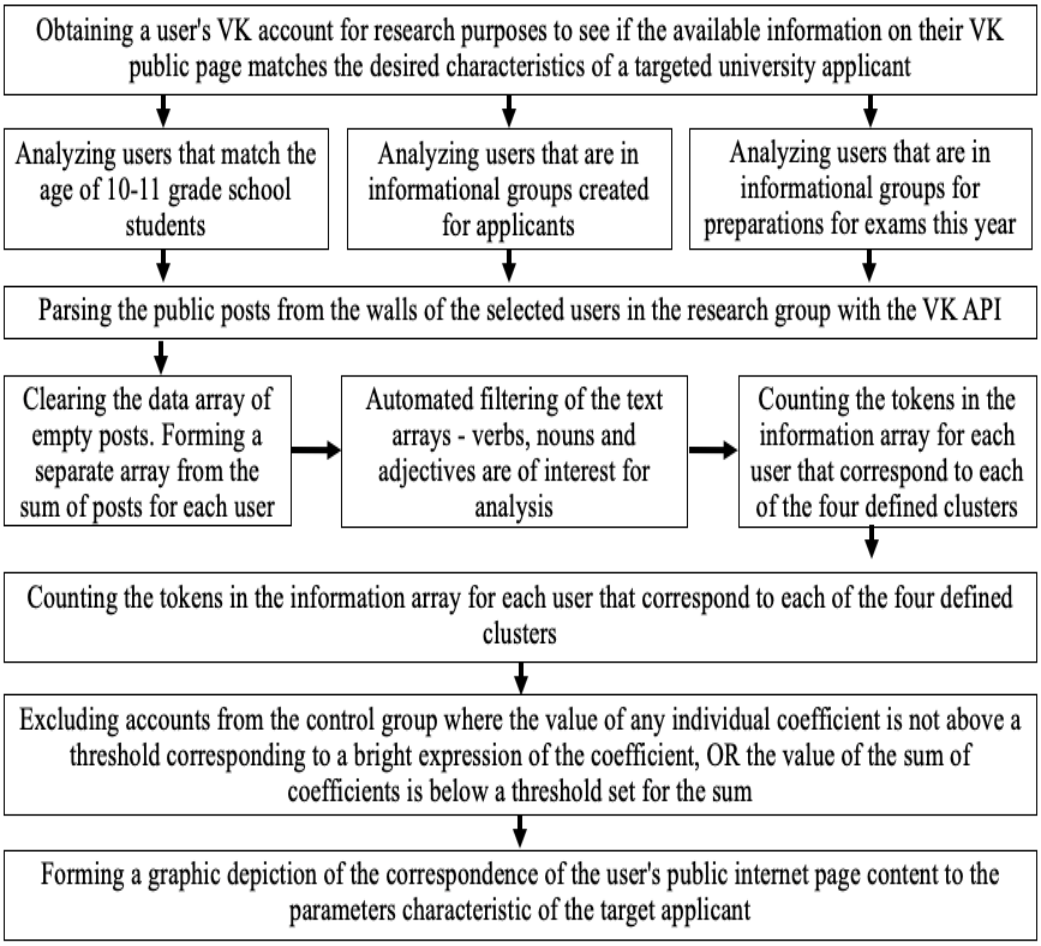

Figure 3 – an algorithm for processing the applicant's account

5. CONCLUSION AND RECOMMENDATIONS

Through text analysis methods, a four-dimensional informational space of popular topics in public internet posts of active students was formed using LDA. A mathematical tool for analyzing internet accounts of users in the VK social network, determining their digital avatar, a mathematical vector representation of the degree of expression of a number of topics in the texts of public posts, was created during the study. The main practical result is a methodology for assessing the correspondence of the obtained digital avatar of the user to the university's informational space. This algorithm allows to identify people for the use of additional resources efficiently when attracting them to participate in the admission process. An unexplored recommendation for the application

Vol: 15 **No:** 1 **Year:** 2023 **ISSN:** 2146-0744 (Online) (**pp.** 324-341) **Doi:** 10.34109/ijebeg.2023150115

of this method is the need for periodic updating of the states of the basic informational space which, as the mathematization of the social environment, has the typical for it trends of variability. It is also recommended that new model for creating digital avatars of applicants based on social media data would greatly enhance the recruitment process. By analyzing an applicant's social media presence, including their interests, activities, and online behavior, the model can generate accurate avatars that reflect their personality and suitability for a position. This approach would provide recruiters with valuable insights and aid in making more informed hiring decisions.

6. IMPLICATIONS AND FUTURE DIRECTIONS

The research has practical and theoretical implications, theoretically, study is important for the formation of a universal method of processing public internet data in general, which creates a tool for assessing the digital environment, aimed at achieving socially significant goals. Moreover, this study contributes to the field of personality assessment by exploring the potential of social media data as a valuable source of information for understanding individuals' traits, interests, and behaviors. The range of data utilized for personality evaluation is widened by this approach beyond conventional self-reported measurements. Second, it improves knowledge of the creation and representation of digital identities. The approach sheds light on how people present themselves online and the effects of their digital footprints on how people perceive their eligibility for particular jobs by evaluating social media data. on the other hand, it aided in the investigation of algorithmic decision-making and its influence on hiring procedures. The approach addresses concerns of bias, privacy, and openness while highlighting the possible advantages and disadvantages of adopting automated avatar generation. Last but not least, this model creates new research opportunities for the study of how social media, technology, and employment are intertwined, leading to a better understanding of the dynamics that are changing between online presence and actual professional outcomes.

This study has certain applications in addition to a theoretical approach. For instance, institutions might utilize this study to enhance their efforts at retention and recruiting. Universities can more efficiently target their outreach efforts by identifying students who are motivated by their school and who are likely to succeed. This study might also be used to monitor student development and pinpoint areas where they want further assistance. This information could then be used to provide students with the resources they need to succeed. Moreover, this model could enhance the objectivity of the hiring process by relying on data-driven algorithms rather than subjective judgments. This helps reduce unconscious biases and promotes fairness and equal opportunities. On the other hand, use of digital avatars allows for more thorough evaluations, capturing additional information beyond traditional application materials. This comprehensive assessment can assist recruiters in identifying potential red flags or positive attributes

that may not be apparent otherwise. Lastly, model enables time and cost efficiency by automating the avatar creation process, streamlining the initial screening stage, and reducing the need for manual evaluation of individual social media profiles

With significance contributions, the research has several limitations. This research was conducted on a small sample of students, so the findings may not be generalizable to the broader population. Additionally, the research only looked at one type of social media data, so it is not clear if the findings would be the same for other types of data. Future research could build on this research by conducting a larger study with a more diverse sample of students. Additionally, future research could look at how this research could be used to improve student recruitment, retention, and success. Other limitation of the study was a new model for creating digital avatars of applicants based on social media data is the potential for bias and privacy concerns. Social media data may not provide a comprehensive or accurate representation of an individual's true character or capabilities, leading to biased assessments. Moreover, privacy issues arise as the use of personal data from social media raises ethical questions regarding consent and data protection. Future research could involve refining the model to address bias and privacy concerns, incorporating a wider range of data sources beyond social media, such as professional profiles or educational records, and implementing robust mechanisms for obtaining informed consent and ensuring data privacy. Additionally, research could focus on developing algorithms that better capture nuanced aspects of an individual's personality and skills, as well as exploring ways to enhance the transparency and interpretability of the avatar creation process.

Funding

The research was funded by the Ministry of Science and Higher Education of the Russian Federation under the strategic academic leadership program “Priority 2030” (Agreement 075-15-2023-380 dated 20.02.2023).

REFERENCES

Ahmad, M., Beddu, S., Itam, Z. b., & Alanimi, F. B. I. (2019). State of the art compendium of macro and micro energies. *Advances in Science and Technology. Research Journal, 13*(1), 88-109. doi: https://doi.org/10.12913/22998624/103425

Ali, M. (2022). A Study of Citizen Satisfaction with the Spirit of Innovation and Work Validity of Basic Government Personnel in Shandong Province, China. *Journal of Advances in Humanities Research, 1*(2), 1-16. doi: https://doi.org/10.56868/jadhur.v1i2.32

Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., et al. (2010). Facebook profiles reflect actual personality, not self-idealization. *Psychological science, 21*(3), 372-374. doi: https://doi.org/10.1177/0956797609360756

Bastani, K., Namavari, H., & Shaffer, J. (2019). Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. *Expert Systems with Applications, 127*, 256-271. doi: https://doi.org/10.1016/j.eswa.2019.03.001

binti Ab Aziz, N. S., & Balraj, B. M. (2022). Soft Skills for Employability from Academics Perspectives. *Journal of Advances in Humanities Research, 1*(3), 16-36. doi: https://doi.org/10.56868/jadhur.v1i3.36

Błachnio, A., Przepiorka, A., & Pantic, I. (2016). Association between Facebook addiction, self-esteem and life satisfaction: A cross-sectional study. *Computers in Human Behavior, 55*, 701-705. doi: https://doi.org/10.1016/j.chb.2015.10.026

Bonnema, J., & Van der Waldt, D. (2008). Information and source preferences of a student market in higher education. *International journal of educational management, 22*(4), 314-327. doi: https://doi.org/10.1108/09513540810875653

Brier, A., & Hopp, B. (2011). Computer assisted text analysis in the social sciences. *Quality & Quantity, 45*(1), 103-128. doi: https://doi.org/10.1007/s11135-010-9350-8

Celardo, L., & Everett, M. G. (2020). Network text analysis: A two-way classification approach. *International Journal of Information Management, 51*, 102009. doi: https://doi.org/10.1016/j.ijinfomgt.2019.09.005

Chambers, D. (2013). *Social media and personal relationships: Online intimacies and networked friendship*. Springer. doi: https://doi.org/10.1057/9781137314444

Cummings, J. N., Butler, B., & Kraut, R. (2002). The quality of online social relationships. *Communications of the ACM, 45*(7), 103-108. doi: https://doi.org/10.1145/514236.514242

Ellison, N. B., Steinfield, C., & Lampe, C. (2007). The benefits of Facebook "friends:" Social capital and college students' use of online social network sites. *Journal of computer-mediated communication, 12*(4), 1143-1168. doi: https://doi.org/10.1111/j.1083-6101.2007.00367.x

Jing, Z., Turi, J. A., Lu, S., & Rosak-Szyrocka, J. (2023). Sustainability through Factory-Based Learning in Higher Education. *Sustainability, 15*(6), 5376. doi: https://doi.org/10.3390/su15065376

Katz, J. E., & Rice, R. E. (2002). *Social consequences of Internet use: Access, involvement, and interaction*. MIT press. doi: https://doi.org/10.1108/14636690310495274

Kaufman, I., & Horton, C. (2014). *Digital marketing: Integrating strategy and tactics with values, a guidebook for executives, managers, and students*. Routledge. doi: https://doi.org/10.4324/9781315879451

Kemp, S. (2023). *Digital 2023: Global Overview Report*. Simon Kemp. Retrieved from https://datareportal.com/reports/digital-2023-global-overview-report

Khan, U., Cheng, Y., Shah, Z. A., & Ullah, S. (2020). Resistance in disguise and the re-construction of identity: a case of the Pashtuns in Pakistan. *Inter-Asia Cultural Studies, 21*(3), 374-391. doi: https://doi.org/10.1080/14649373.2020.1797121

Kraut, R., & Burke, M. (2015). Internet use and psychological well-being: Effects of activity and audience. *Communications of the ACM, 58*(12), 94-100. doi: https://doi.org/10.1145/2739043

Kuznetsov, M., Gorovoy, A., & Rodionov, D. (2021). Web Innovation Cycles and Timing Projections–Applying Economic Waves Theory to Internet Development Stages. In *International Scientific Conference on Innovations in Digital Economy* (pp. 3-21). Springer. doi: https://doi.org/10.1007/978-3-031-14985-6_1

Leijse, M. M., Koning, I. M., & van den Eijnden, R. J. (2023). The influence of parents and peers on adolescents' problematic social media use revealed. *Computers in Human Behavior, 143*, 107705. doi: https://doi.org/10.1016/j.chb.2023.107705

Levenson, J. C., Shensa, A., Sidani, J. E., Colditz, J. B., & Primack, B. A. (2016). The association between social media use and sleep disturbance among young adults. *Preventive medicine, 85*, 36-41. doi: https://doi.org/10.1016/j.ypmed.2016.01.001

Miao, F., Kozlenkova, I. V., Wang, H., Xie, T., & Palmatier, R. W. (2022). An emerging theory of avatar marketing. *Journal of Marketing, 86*(1), 67-90. doi: https://doi.org/10.1177/0022242921996646

Mikami, A. Y., Szwedo, D. E., Allen, J. P., Evans, M. A., & Hare, A. L. (2010). Adolescent peer relationships and behavior problems predict young adults' communication on social networking websites. *Developmental psychology, 46*(1), 46–56. doi: https://doi.org/10.1037/a0017420

O'reilly, T. (2007). What is Web 2.0: Design patterns and business models for the next generation of software. *Communications & strategies, 1*(65), 17-37. Retrieved from https://ssrn.com/abstract=1008839

Ozimek, P., Brailovskaia, J., & Bierhoff, H.-W. (2023). Active and passive behavior in social media: Validating the Social Media Activity Questionnaire (SMAQ). *Telematics and Informatics Reports, 10*, 100048. doi: https://doi.org/10.1016/j.teler.2023.100048

Prell, C. (2011). *Social network analysis: History, theory and methodology*. SAGE. Retrieved from https://www.torrossa.com/en/resources/an/4913142

Roberts, C. W. (2020). *Text analysis for the social sciences: methods for drawing statistical inferences from texts and transcripts*. Routledge. doi: https://doi.org/10.4324/9781003064060

Srinivasa-Desikan, B. (2018). *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing Ltd. Retrieved from https://www.perlego.com/book/771695

Steinfield, C., Ellison, N. B., & Lampe, C. (2008). Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. *Journal of applied developmental psychology, 29*(6), 434-445. doi: https://doi.org/10.1016/j.appdev.2008.07.002

Sultan, M., Scholz, C., & van den Bos, W. (2023). Leaving traces behind: Using social media digital trace data to study adolescent wellbeing. *Computers in Human Behavior Reports, 10*, 100281. doi: https://doi.org/10.1016/j.chbr.2023.100281

Twenge, J. M., Martin, G. N., & Spitzberg, B. H. (2019). Trends in US Adolescents' media use, 1976–2016: The rise of digital media, the decline of TV, and the (near) demise of print. *Psychology of Popular Media Culture, 8*(4), 329–345. doi: https://doi.org/10.1037/ppm0000203

Ullah, S., Khan, U., Begum, A., Han, H., & Mohamed, A. (2023). Indigenous knowledge, climate change and transformations of Gwadar fishing community. *International Journal of Climate Change Strategies and Management, ahead-of-print*(ahead-of-print), 1-20. doi: https://doi.org/10.1108/IJCCSM-06-2022-0069

Van den Eijnden, R. J., Meerkerk, G.-J., Vermulst, A. A., Spijkerman, R., & Engels, R. C. (2008). Online communication, compulsive Internet use, and psychosocial well-being among adolescents: a longitudinal study. *Developmental psychology, 44*(3), 655-665. doi: https://doi.org/10.1037/0012-1649.44.3.655

Yang, Q., & Li, W. (2015). The Lda Topic Model Extension Study. In *International Conference on Logistics Engineering, Management and Computer Science (LEMCS 2015)* (pp. 857-860). Atlantis Press. doi: https://doi.org/10.2991/lemcs-15.2015.169

Zeng, D., Chen, H., Lusch, R., & Li, S.-H. (2010). Social media analytics and intelligence. *IEEE Intelligent Systems, 25*(6), 13-16. doi: https://doi.org/10.1109/MIS.2010.151