

Saint Petersburg State University Graduate School of Management  
Master of Business Analytics and Big Data

**ANALYSIS OF SOCIAL MEDIA DATA TO OPTIMIZE THE MARKETING  
STRATEGY OF GSOM HIGHER EDUCATION PROGRAMS**

Master Thesis

Master's Thesis by the 2<sup>nd</sup> year  
students  
Master in Business Analytics and Big Data

D.A. Doroshkova  
M.O. Ryleeva  
A.D. Lisitsyna

Research Advisor:  
A.V. Gorovoy  
Senior Lecturer, Department of Information Technologies in Management

Saint Petersburg




2023

## ЗАЯВЛЕНИЕ О САМОСТОЯТЕЛЬНОМ ХАРАКТЕРЕ ВЫПОЛНЕНИЯ ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

Мы, Дорошкова Дарья Александровна, Лисицына Анна Дмитриевна и Рылеева Мария Олеговна, студентки второго курса магистратуры направления «Бизнес-аналитика и Большие Данные», заявляем, что в нашей магистерской диссертации на тему «Анализ данных социальных сетей для оптимизации маркетинговой стратегии высших образовательных программ ВШМ», представленной в службу обеспечения программ магистратуры для последующей передачи в государственную аттестационную комиссию для публичной защиты, не содержится элементов плагиата.

Все прямые заимствования из печатных и электронных источников, а также из защищенных ранее выпускных квалификационных работ, кандидатских и докторских диссертаций имеют соответствующие ссылки.

Нам известно содержание п. 9.7.1 Правил обучения по основным образовательным программам высшего и среднего профессионального образования в СПбГУ о том, что «ВКР выполняется индивидуально каждым студентом под руководством назначенного ему научного руководителя», и п. 51 Устава федерального государственного бюджетного образовательного учреждения высшего образования «Санкт-Петербургский государственный университет» о том, что «студент подлежит отчислению из Санкт-Петербургского университета за представление курсовой или выпускной квалификационной работы, выполненной другим лицом (лицами)».




 _____	Дорошкова Д.А.	(Подпись студента)
 _____	Лисицына А. Д.	(Подпись студента)
 _____	Рылеева М.О.	(Подпись студента)
_____	01.06.2023	(Дата)

STATEMENT ABOUT THE INDEPENDENT CHARACTER OF  
THE MASTER THESIS

We, Doroshkova Daria Aleksandrovna, Lisitsyna Anna Dmitrievna and Ryleeva Mariia Olegovna, (second) year master student, program «Business Analytics and Big Data», state that our master thesis on the topic «Analysis of Social Media Data to Optimize the Marketing Strategy of GSOM Higher Education Programs», which is presented to the Master Office to be submitted to the Official Defense Committee for the public defense, does not contain any elements of plagiarism.

All direct borrowings from printed and electronic sources, as well as from master theses, PhD and doctorate theses which were defended earlier, have appropriate references.

We are aware that according to paragraph 9.7.1. of Guidelines for instruction in major curriculum programs of higher and secondary professional education at St. Petersburg University «A master thesis must be completed by each of the degree candidates individually under the supervision of his or her advisor», and according to paragraph 51 of Charter of the Federal State Institution of Higher Education Saint-Petersburg State University «a student can be expelled from St. Petersburg University for submitting of the course or graduation qualification work developed by other person (persons)».

 _____	Дорошкова Д.А.	(Student's signature)
 _____	Лисицына А. Д.	(Student's signature)
 _____	Рылеева М.О.	(Student's signature)
_____	01.06.2023	(Date)

## Content

Introduction .....	3
Chapter 1. Use of social media for higher education promotion and problem statement .....	5
1.1. History of approaches to higher education promotion.....	5
1.2. Social media platforms' popularity .....	7
1.3. Problem description .....	8
1.4. Machine Learning approach in social media analytics .....	11
1.5. Research goals and expected results .....	14
Chapter 2. Exploratory data analysis and methodology for working with text data .....	16
2.1. Methodology of text processing .....	16
2.2. Data description and transformations .....	23
2.3. Dashboards for tracking main estimators .....	30
Chapter 3. Building Machine Learning models for VK data processing .....	34
3.1. Dataset preparation for model building .....	34
3.2. Modelling the dependencies between likes and post characteristics .....	36
3.3. Cluster analysis .....	46
Conclusion .....	52
Reference list .....	54

## Introduction

In recent years, social media has become an increasingly important tool for promoting higher education programs. As more and more students turn to social media to research colleges and universities, it has become essential for institutions to have a solid social media presence.

Social media has changed the higher education marketing landscape in an ambivalent way. On the one hand, it allows institutions to reach a wider audience and connect with prospective students in new ways and rather efficiently thanks to technologies of targeting advertisement. Universities can leverage social media data to create more focused content and engage with the right audience (Chaudhry, 2023). On the other hand, there are difficulties with getting the target audience to notice the content and engage with it. More than 30% of social media professionals struggle with it (Calus, 2023), and the educational sphere is undoubtedly not an exception. Colleges and universities are constantly creating content, but the question is how to actually reach the target audience. Usually, higher education establishments use social media to connect with prospective students by sharing information about academic programs, campus events, and student life and addressing concerns and questions. However, institutions sometimes struggle with creating engaging content that resonates with their target audience. If they manage to do so, it will help people feel like a part of the university culture and encourage them to share the posts further, therefore expanding the reach/invoking word-of-mouth marketing (Rogers).

Overall, social media can be a powerful tool for promoting higher education programs if institutions use it correctly. They should create informative, entertaining, visually appealing content that aligns with the institution's brand and builds a solid social media presence. Institutions can show their advantages to their target audience and attract more prospective students. Moreover, it is crucial for institutions to use social media strategically and to tailor their approach to the needs and interests of their potential students.

Our work consists of the following parts: Chapter 1 covers the history and trends of education marketing and describes the GSOM marketing strategy, particularly in the Russian social media network - VK. Moreover, the first chapter discusses the latest ML methods for evaluating social media marketing performance. Chapter 2 describes the methodology we chose for our research and VK API data used for data analysis. Also, the second chapter covers the dashboard design and its managerial use. Chapter 3 discusses the models for posts and followers' analysis and its main results. In Conclusion, we provide data-driven recommendations to the marketing department and discuss areas for further research.

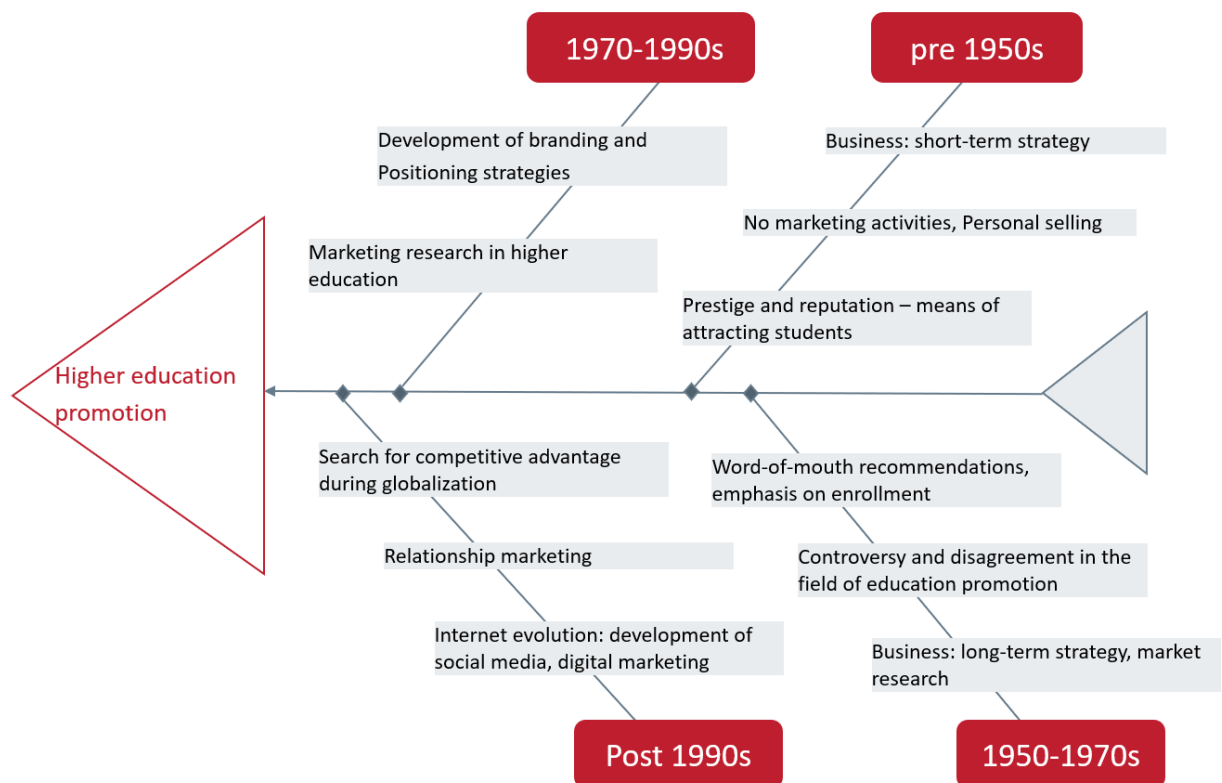
Generally, the work was distributed inside of the group equally. For Introduction and the first chapter, each participant wrote an equal volume of text, which was then discussed and revised together. The exploratory data analysis and data preparation was done by all the authors jointly. For the practical part the work was divided in the following way: A. D. Lisitsyna was responsible for the dashboard and all the work connected to it, M.O. Ryleeva built classification models and made k-means clusterization, and D.A. Doroshkova built regression models and DBScan based clusterization. Conclusions from the analysis were drawn and written together.

# Chapter 1. Use of social media for higher education promotion and problem statement

## 1.1. History of approaches to higher education promotion

The history of approaches to higher education promotion can be traced back to the early 20th century when institutions of higher learning began to recognise the importance of marketing their programs and services to potential students. In this section, we will examine several articles that provide insights into the evolution of higher education promotion and its current state.

Higher education promotion was always following the lead of marketing in business. Harrison-Walker (2010) mentioned in his article that the history of higher education promotion and marketing itself can be divided into four eras: the pre-1950s era, the 1950s-1970s era, the 1970s-1990s era, and the post-1990s era (Fig. 1).



**Figure 1** Fishbone diagram of the main historical eras of higher education promotion

Until the mid-50s, marketing methods were oriented mainly on short-term strategy, the centre of which was personal selling, advertising, and sales promotion. Higher education institutions had the same beliefs; admission offices and development teams focused on prestige and reputation as the primary means of attracting students. They were responsible for selling the institution to prospective students and lacked marketing activities.

In The 1950s-1970s, education and business marketing took different paths. While businesses began to adopt more long-term strategies to study the customer, his/her desire, and the conduct of the product, there was controversy and disagreement in the field of education. For example, articles mentioned that academics and professors were hesitant to define students as customers and argued that, at a young age, they were unable to assess what they really needed (Harrison-Walker, 2010) and that marketing activities aimed at students' needs would lower academic standards of educational quality (Constantinides, Stagno, 2011). Institutions were primarily focused on increasing enrollment rather than enhancing the quality of education. Still, institutions relied mainly on word-of-mouth recommendations.

In the 1970s-1990s era, higher education institutions began to adopt a more comprehensive approach to marketing, including the development of branding and positioning strategies. Institutions began to adopt more sophisticated marketing instruments, such as market research and branding, in order to compete for students in a more crowded marketplace. Finally, the post-1990s world has seen an Internet revolution and the rise of its components, such as social media. Online interactive applications, user-generated content (blogs, online communities), digital marketing and online education have transformed the way that higher education institutions promote their programs and services. At the same time, in the early 2000s, institutions recognised student retention as an increasingly important strategic theme and introduced a new concept of relationship marketing. Relationship marketing – a strategy aimed at improving and maintaining relationships with the stakeholders. Regarding higher education, the stakeholders are students, faculty members, professors, administration and other parties of universities or colleges. The idea is that stakeholders are ambassadors who define and develop the institution's brand name in the eyes of the general public (Jain et al., 2022). Some papers analyse such dependencies in the field of relationship marketing as the impact of courses held in university on student retention and the nature of relationships among students that affect long-term loyalty (Harrison-Walker, 2010). The combination of social media and relationship marketing can lead to improved communication, customer engagement and brand loyalty (Constantinides, Stagno, 2011). Another factor that has had an influence on higher education promotion in recent days is globalisation. It has increased the competition among universities all over the world in order to attract both talented students and professors (Hrusca, Kuca, 2020). In response to the trend, higher education institutions have started searching for a unique definition of what they offer to differentiate themselves and use social media presence as a potential competitive advantage (Chapleo, O'Sullivan, 2017).

Russia, along with other countries, has also passed these stages of development of marketing promotion in the field of education. However, because of the large geographic extent



and regional differentiation, universities in the country are still organised in a very hierarchical way that prevents them from using market-oriented and relationship marketing approaches (Bydanova et al., 2018).

Overall, higher education promotion has evolved significantly over the past century. The history of higher education promotion can be understood in terms of several key trends and developments. These include the increasing importance of branding and reputation, the rise of digital marketing and social media, and the growing emphasis on globalisation. Through the years, there has been a shift from traditional marketing strategies, such as print advertising, direct mail, and reputation-based approach, towards more comprehensive and sophisticated marketing strategies, which continue to improve.

## **1.2. Social media platforms' popularity**

As mentioned earlier, social media is being chosen as a marketing tool by higher education institutions more and more often because the younger generation actively uses it. In fact, students use several social media apps at once on a daily basis. The most popular are Twitter, Instagram, and Facebook. In Russia, the most popular social media is VK<sup>1</sup>.

It is also important to notice that VK is constantly developing and continues to grow in spite of the fact that it emerged nearly 20 years ago. For example, in March 2022, VK's monthly global audience grew by 2.4% to a record of 100.4 million. Half of this monthly audience is using VK every day. The most active segment of VK users are people of the age category 12-24 years old, which includes the age of potential students of colleges and universities. They spend almost an hour scrolling VK social media daily. What is more, VK improved the recommendation engine, which allowed it to increase the number of new subscriptions from news feeds to community pages by 25% (VK Press Office, 2022). All the listed factors allow us to conclude that VK is a prospective and progressive social media platform for promoting higher education programs.

Generally, social media platforms are a promising way to promote higher education programs also because the contact with the audience is personalised. Moreover, it is often cheaper to promote on social media than in other channels, which is of specific importance in unstable economic conditions (Zailskaite-Jakste and Kuvykaite, 2012). That is why many universities have started to use social media marketing in their promotion strategies. Before discussing the matter

---

<sup>1</sup> Similarweb, "Most Visited Social Media Networks Websites in Russia." URL: <https://www.similarweb.com/top-websites/russian-federation/computers-electronics-and-technology/social-networks-and-online-communities/#:~:text=vk.com%20ranked%20number%201,Media%20Networks%20websites%20in%20Russia.>

further, it is necessary to give a definition of social media marketing strategy. Social media marketing (SMM) is "the systematic planning, implementation, management, and control of all activities undertaken in social media that are aligned with overarching goals" (Zerres, 2020).

Another reason for using social media marketing more often is the fact that smartphones and apps, including social media apps, have become an integral part of everyday life (Kreshnik et al., 2022). So, it is reasonable to conclude that attracting people to the already usable environment (social media apps) is much easier than transferring them to websites. Moreover, social media apps are more catching than websites because they can deliver multimedia content, including not only text but also images, videos, and even gifs or memes. Furthermore, there is a strong connection among social networking users, enabling them to get feedback quickly (Hrusca, Kuca, 2020). To understand the promotion standards in the higher education sphere, some researchers selected the world's ten most influential universities and analysed their social media activity. It turned out that all universities were uploading posts every day, some even several times a day (Hrusca, Kuca, 2020). What is more, present and former students produced content and shared information regarding their accomplishments and life on campus to show the strong community and loyalty to potential students. However, while social media indeed influence students' choice of institution, traditional information channels such as campus visits, university websites, and brochures are still considered powerful ways of promotion (Constantinides, Stagno, 2011). One of the hypotheses for such a turn is that institutions do not use social media properly and present an unclear brand image.

### **1.3. Problem description**

#### **1.3.1 How to capture attention in the large flow of information**

Although technological advancements have revolutionised the way people access and use information, the fundamental problems arising from bounded rationality and limited attention still exist (Van Knippenberg et al., 2015). There are still some controversies about the existence of such a phenomenon as an information overload among scientists and researchers, but this difficulty is actually experienced by many people (Koltay, 2017). Information overload can be defined as "a state of affairs where an individual's efficiency in using the information in their work is hampered by the amount of relevant, and potentially useful, information available to them" (Bawden and Robinson, 2009). This phenomenon is influenced by personal factors and human's limited ability to process large amount of information which is constantly available in a modern world (Koltay, 2017). With the increased usage of the internet, the problem of information overload became more widespread. (Bawden and Robinson, 2020). However, it is important to note that the concept of information overload is not new. The problem of information overload was first recognised in the

late 1950s/early 1960s, especially in science and technology. It was explained by the increased number of publications available in the scientific world, thanks to the emergence of mechanised documentation and computerised information handling (Bawden and Robinson, 2009). Later, starting from the 1980s, it was claimed that communication technology was causing information overload (Koltay, 2017). Email and social media are currently considered the primary sources of overwhelming information. Moreover, the use of mobile devices, particularly smartphones, has only added to issues with overload (Bawden and Robinson, 2020).

Information overload can make it difficult for users to remember messages and brands they see (Cordero-Gutiérrez and Lahuerta-Otero, 2020). What is more, the abundance of information available through social media can result in users being overwhelmed with data, making it challenging to identify and consume the most relevant content. Mostly, users end up looking only through the top posts in their feed, i.e., the most recent ones. This can lead to a decrease in the quality of social media experience, user burnout, and, ultimately, user disengagement (Grineva and Grinev, 2012). Identifying disengagement factors is crucial for firms to maintain long-term relationships with their customers. Communities need to focus on increasing the engagement of those who join them on social media (Dutot and Mosconi, 2016). For measuring engagement, there is a special metric - engagement rate. It can be measured differently depending on the needs of a company. The most common method to define engagement rate for social media is to calculate the average number of interactions (likes, comments, and others) on the number of followers (Yost et al., 2021). The engagement rate is considered to be a crucial metric in the social media communication of universities (Marino and Lo Presti, 2016).

Information overload is not the only thing that causes struggles with engaging potential students in universities' content. The competition for students and resources among universities is increasing in social media (Harrison-Walker, 2010). That is why higher education institutions, the same as companies, need to develop a unique identity and image that reflects their values and mission (Maringe and Gibbs, 2009). For that purpose, institutions embrace marketing that is seen in this sphere to differentiate from competitors and attract students who are increasingly aware of the value of their education (Maringe and Gibbs, 2009).

### **1.3.2. Current marketing strategy**

GSOM is a Russian Business School established on the basis of SPBU. Russia's only holder of the "triple crown" of international accreditations - the independent associations AACSB, EQUIS and AMBA - is among the top 1% of business schools globally with similar recognition. GSOM SPbU prepares specialists in management on undergraduate, master's, postgraduate, MBA, Executive MBA, and corporate professional development programs. Currently, GSOM is

struggling with the engagement of potential undergraduate and master's students, i.e., enrollees, in the respective communities in VK.

VK is a popular social media site in the Russian Federation, with over 500 million registered users. Also, VK has been one of the few permitted social media promotion channels left in Russia since 2022.

The current marketing strategy of GSOM bachelor and master programs in VK can be divided into two tracks: PR and advertisement.

For the first track, VK is used as one of the main communication tools that help to interact with a broad audience as well as with applicants. The marketing department tries to build collaborations with other communities and increase organic reach. For example, it receives support from the VK department that promotes educational content. In return, the GSOM marketing department provides VK with opportunities to promote their employer brand in GSOM and develop projects that benefit VK. Through this partnership, educational broadcasts are shown to a wide audience of interest. The current goal of the PR track is to bring the engagement rate to 1%.

For the second track, advertisement, GSOM's social media marketing department is focused solely on VK. This platform is used to collect registrations for Doors Open Days and other presentations. The most targeted products, in this case, are undergraduate programmes; advertising at the graduate level also works, but to a lesser extent. It is important to mention that the most converting tool is VK lead forms. The growth of followers was happening mostly organically because of budget constraints. Now that these constraints are no longer a big issue, the marketing department plans to run targeting ads aimed at attracting followers to applicant groups to connect auto-funnels, subscriptions, and other tools.

The target audience is revised regularly on an annual basis - at the end of the enrollment process. However, it is worthy of note that it was done based on specific examples and small surveys that might not allow us to see the full picture.

### **1.3.3. Drawbacks of the current marketing strategy and required adjustments**

While building a marketing strategy, it is crucial to understand the needs and wants of different stakeholders, including students, staff, alums, and the wider community. This requires market research, segmentation, and targeting strategies to ensure effective marketing efforts (Maringe and Gibbs, 2009).

Currently, the GSOM marketing department lacks in-depth market research, competitor analysis, and segmentation. GSOM is still known only to a certain circle of people, unlike big state

universities. With a better understanding of the audience, their behaviour and interests, and the competitors' strategies, it will be possible to revise the current marketing strategy and improve it in order to increase visibility and engagement rate.

To identify factors that increase engagement rate in VK groups and to clarify the target audience, there is a need for comprehensive data analysis. The existing approach to analytics requires qualitative improvement, which could be enabled by more advanced techniques and tools, such as machine learning. More details on how and for what purposes machine learning can be used in social media analysis will be described in the next paragraph.

#### **1.4. Machine Learning approach in social media analytics**

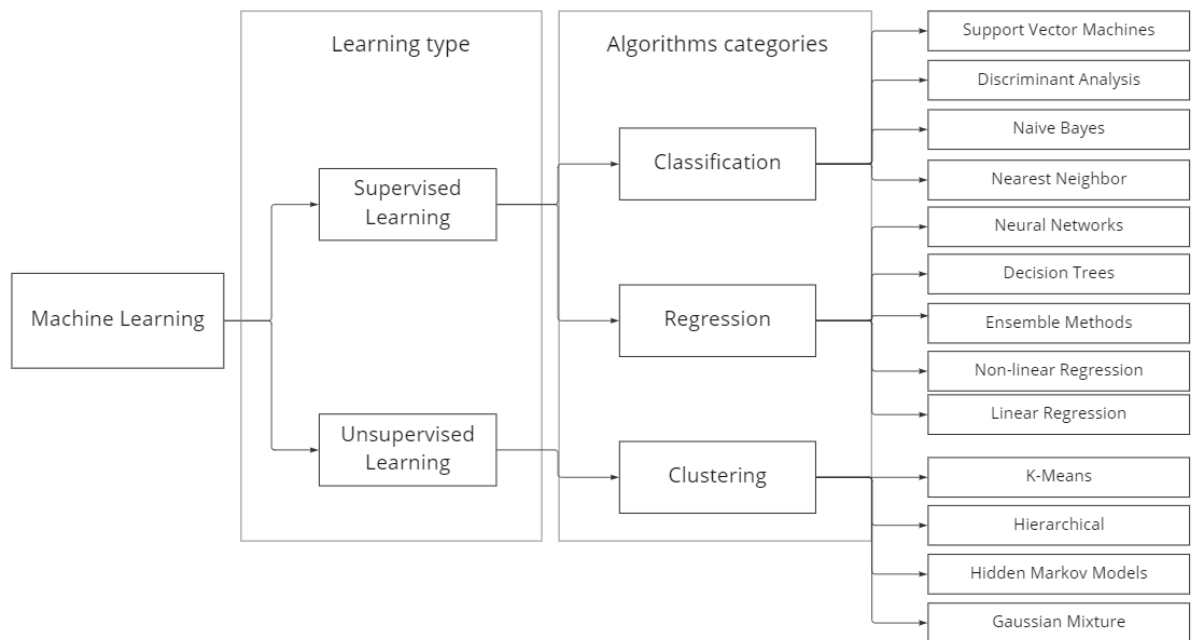
Due to the high significance of the results that can be obtained from the analysis of social networks, it is necessary to choose the most modern and effective methods of analysis. Machine learning algorithms are becoming increasingly popular in the field of social media marketing analysis, providing a number of advantages to companies seeking to optimise their social media strategies. Using machine learning algorithms, social media marketers can get information about customer preferences, behaviours and moods, as well as predict trends and predict customer behaviour.

In their study, Belfin et al. (2020) identified the importance of machine learning algorithms in social media marketing analysis, noting that these algorithms can help companies better understand their target audience and adapt their marketing strategies accordingly. The authors also noted that machine learning algorithms could help identify new trends, provide insight into customer behaviour and help predict customer behaviour in the future. Similarly, Basarslan et al. (2020) explored the use of machine learning algorithms for social media marketing analysis, emphasising the importance of natural language processing and sentiment analysis to understand customer sentiment and opinions. The authors noted that machine learning algorithms can help identify key topics and issues related to a company or product, allowing marketers to respond to customer feedback quickly and improve brand reputation.

Machine learning algorithms can extract valuable information from social media data that can help companies and researchers better understand user behaviour and trends. In order to better understand what is behind machine learning algorithms, as well as the correspondence between algorithms and marketing tasks, the necessary theory will be discussed further. Machine learning systems can be classified according to the amount and type of control they receive during training. There are four main categories: supervised and unsupervised learning, semi-supervised learning and reinforcement learning (A. Geron, 2019). Most machine learning tasks are solved using

supervised and unsupervised learning. Therefore, only methods related to these two types of training will be considered further.

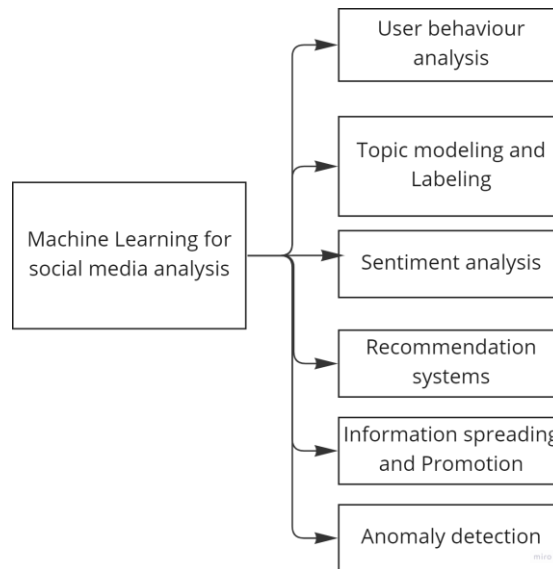
In supervised learning, the training data adjusted to the algorithm includes the desired solutions, called labels (A. Géron, 2019). The goal of the supervised learning algorithm is to use a data set to create a model that takes a feature vector  $x$  as input and outputs information that allows for output a label for this feature vector (A. Burkov, 2019). On the contrary, unsupervised learning uses unmarked data. A basic diagram of the structure of machine learning algorithms is shown on Fig. 2.



**Figure 2** Machine Learning Overview

Source: Batrinca et al., 2014

Most of these machine learning algorithms are widely used to analyse social networks. Figure 3 shows what machine learning is most commonly used in social media.



**Figure 3** Machine Learning for social media

By Hai Tao (Hai Tao et al., 2022) was proposed an approach using naive Bayes and maximum entropy algorithms to classify tweets as positive, negative or unbiased in order to determine the perception of business by customers. The study confirms the accuracy of machine learning algorithms for classifying and predicting data. The proposed models make it possible to reduce the number of project failures by 80% due to data mining in social networks.

The use of machine learning for sentiment analysis is discussed in the article by Mejova Y. (2009). Sentiment analysis is about recognizing the tone of text, which is applicable in business to analyse reviews and audience acceptance of products and content. The article discusses various tasks related to sentiment analysis, including the determination of moods or opinions, the classification of polarity and the definition of the target audience of opinions. The first task involves classifying the text as objective or subjective by studying adjectives and adverbs in sentences. The second is to classify the opinion as positive or negative. The third stage involves defining the topic of opinion and extracting related characteristics. Evidently, sentiment analysis is a useful and strong tool for analysing textual data, which has a wide range of applications.

In a study by Buckley et al. (2014) a new system architecture for improving customer relationships is suggested and loyalty in various industries based on natural language processing and machine learning to analyse the context of social media posts and product information in order to compare social networks with products and rank potential advertising. The use of the platform is demonstrated in the travel and tourism industry using Twitter as a social media platform. The system is aimed at targeting individual posts on social networks to improve marketing campaigns. As a result of the application of machine learning algorithms to the analysis of data from social networks, the percentage of user responses reached 15%.

Overall, these studies demonstrate the importance of machine learning algorithms for social media marketing analysis. By analysing large amounts of data from social networks, these algorithms can help companies gain valuable information about customer behaviour and moods, as well as predict future trends and behaviour.

## **1.5. Research goals and expected results**

Based on the previously reviewed literature and drawn conclusions, we formulated the following:

- Managerial problem - How to increase engagement rate in VK for GSOM bachelor and master enrollees.
- Research problem - To find out the followers' preferences and what content engages them the most.
- Research questions:
  - What are the behavioural patterns of the target audience in VK?
  - What are the preferences of the target audience in VK?
  - Which socio-demographic groups represent the overall subscribers of the GSOM groups?
- Hypotheses:
  - There is a difference in engagement rate between posts that contain video and posts that contain photos;
  - There is a dependency between specific words used in the post and followers' activity;
  - The day of the week and time of post-publishing influence engagement.
  - Preferences of different socio-demographic groups differ in VK

## **Summary of Chapter 1**

In this chapter, we discussed the history of marketing in education. We also looked at trends in social media use. Both history and modern trends coincide in the way that our life is highly digitalised. Institutions seek to find competitive advantage and state the brand and reputation using any tool possible. We then focused on the GSOM marketing strategy and the description of the main marketing promotion tool - VK. Based on the available information on the current state of marketing promotion in GSOM, we identified the shortcomings and analysed the latest ML methods of evaluating social media marketing performance. Findings in Chapter 1 assured us that



with the help of ML algorithms, we would be able to gain some valuable insights into VK groups subscribers' behaviour and sentiment of posts, as well as shape the GSOM's new marketing strategy.

Unfortunately, we did not manage to find an exhaustive amount of research devoted to marketing in social media in Russia, especially in the area of higher education programs promotion. Only foreign social media data was analysed before, and very few researchers investigated the issue of higher education marketing in social networks. Moreover, the Machine Learning technique was rarely used for these types of research, and we believe that this method can bring out new insights and help to understand the specificity of marketing strategy in Russian social media networks.

## **Chapter 2. Exploratory data analysis and methodology for working with text data**

### **2.1. Methodology of text processing**

#### **2.1.1. Natural Language Processing (NLP) techniques for text preprocessing**

Most of the data in social networks is photos and text: user posts, community posts, and comments. For the purpose of a deeper study of the audience, their interests and patterns, the text can be an extremely useful source. As a rule, machine learning models are designed to work with numbers, but there is a section that allows for applying ML algorithms to text data. The machine learning section studying texts and speech is called Natural Language Processing. NLP includes linguistics and artificial intelligence, which jointly recognise and process text data, as well as oral speech. An important part of natural language processing is data preprocessing, that is conducted before building models.

Preprocessing can consist of a different set of stages carried out in different order since it is based on the format of the source text, how much noise there is in it in the form of non-printable characters, symbols, abbreviations, as well as what goal they are trying to achieve in the study and what model they plan to use. Thus, the choice of the text data processing process should be determined by the source data, the volume of text, the problem to be solved and the required results (Araşlanov et al., 2020). The primary purpose of preprocessing, as well as its high significance, is that reducing noise in input data and reducing dimensionality will lead to a much higher performance of machine learning models. Also, the selection of features is often included in the preprocessing since it reduces the number of features, leaving the most significant for a particular model, increasing the interpretability of the results and, at the same time, reducing the training time of the model (improving computational time), and this is often critical in the framework of machine learning (Araşlanov et al., 2020). However, first of all, text data must be cleared. For this purpose, the most widespread steps are as follows: removal of digits and special characters, conversion to lowercase, removal of stop words, lemmatisation, stemming, POS (part of speech), and tokenization of tags (Hasan et al., 2019).

Stopword removal involves the elimination of frequently repeated words that do not carry a semantic load: conjunctions, articles, prepositions, and adjectives bearing no particular statistical relevance (Mooney, Nahm, 2003). Such words will only overload the future machine learning model, increasing the training time and reducing performance.

Summation and lemmatisation are methods of normalising word forms (Singh, Gupta, 2016). Lemmatisation reduces words to their simple form (lemma) as in a dictionary form by

removing inflectional endings from a certain word (Korenius et al., 2004). The use of lemmatisation or stemming increases the accuracy of the algorithm (Hossain et al., 2021).

Tokenisation is a necessary stage of text data preprocessing in almost all machine learning tasks (Solangi et al., 2018). The need for this method is justified by the fact that machine learning models are able to accept only numeric data but not text data. Therefore, tokenisation is almost always used to turn text into tokens, which are words, phrases, and symbols. (Domala et al., 2021). Words are sentence markers, and sentences are paragraph markers (Hossain et al., 2021).

POS tagging is a grammatical classification that consists in assigning the words in the text to the corresponding parts of speech (Kumawat, Jain, 2015). This method helps the machine learning algorithm to distinguish when the same word plays a different role, used in one sentence as a verb and in another as a noun, for example. This is especially important when training the algorithm on a large corpus of text data.

All the steps described above relate to the cleaning and normalisation of text data. However, this is not enough to transfer data to machine learning algorithms - as mentioned earlier, the input data must be in numeric form. Therefore, the following stage, which is often mandatory after data cleaning, will be considered next: numerical representation of the text.

The most widespread techniques for extraction of numeric parameters from a plain text can be divided into two parts: first is based on the frequency of words such as bag-of-words (BoW model) and TF-IDF, and second is word embedding techniques such as GloVe, FastText and Word2Vec (Kowsari et al., 2019).

Researchers highlight that, nowadays, the simplest text vectorisation model is the bag-of-words model (Rani et al., 2022; Kowsari et al., 2019). The essence of the method is to identify all unique words in the text and convert them into a vector taking into account the frequency of words so the length of the vector is equal to the number of unique words. However, the BoW model has disadvantages and limitations, for example, grammar and word order are not taken into account, that is, semantic relationships between words are not taken into account, while this is an important factor for many machine learning tasks.

TF-IDF (term frequency-inverse document frequency) is also a very common, fairly simple and effective model. This method is considered an improved BOW model, where Term Frequency (TF) is the number of times a term appears in the document, and the inverse frequency of the IDF document measures the rarity term in the entire corpus (Mansour et al., 2022).

A group of more complex text vectorisation methods is called word embeddings. These models take into account the semantic meaning of words, assuming that closely spaced words are

interconnected. One of the powerful techniques was proposed by T. Mikolov as the Word2Vec model (Mikolov et al., 2013). Word2Vec is a neural network which combines continuous bag-of-words (CBOW) and Skip-gram architectures to create a high-dimension vector for each word (Kowsari et al., 2019). The achievement of the model is to identify relationships between words with similar semantic meanings.

### **2.1.2. Machine learning models for text processing**

As mentioned before, machine learning is a field of computer science that has revolutionised how social media data is processed and analysed. Regression and classification models are the most common and basic methods among machine learning techniques.

However, in the context of text analysis, logistic regression is not commonly used as a standalone technique. Instead, most researchers prefer to utilise it to establish relationships with additional features. This can be accomplished by categorising texts according to their topics (Zheng, 2012) or leveraging the results of NLP techniques (Rajpurkar et al., 2016). When Jain et al. (2021) experimented with logistic regression, it yielded poorer results than other techniques, such as XGboost and stacking.

Despite the limited use of logistic regression in text analysis, regression analysis can still be helpful in identifying factors that influence sentiment analysis polarity (Blasco et al., 2023). Moreover, a variant of regression analysis called group lasso regression can be especially effective when combined with clustering algorithms that belong to unsupervised machine learning techniques. Combining regression analysis with clustering makes it possible to identify groups of words that are more informative when used together rather than separately (Zheng et al., 2012). This approach can help gain deeper insights into the complexities of textual data and reveal hidden patterns that may not be visible through traditional analysis methods.

Also, machine learning is often used to solve the problem of text classification. Generally speaking, text classification is one of the classical problems in NLP and is a process of categorising texts into groups. When solving the problem of textual classification, it is crucial to understand the type of data and the purpose of the analysis. For example, the text classification system contains four different levels of scope that can be applied: at the document level, the algorithm obtains the relevant categories of a full text of the paper/document. Then, an algorithm can obtain the relevant categories in a portion of a document, in sentences and in even finer divisions - word combinations and phrases (Barnes et al., 2019). Another scope of text classification is the purpose of the analysis. Classification tasks are often used in sentiment analysis, categorisation of news, question answering, topic analysis and even spam detection (Cambria et al., 2022).

Authors note that text classification can be performed either through manual annotation or by automatic labelling. Automatic text classification is divided into two categories: Rule-based methods and ML – base methods. Rule-based classifiers define word patterns that are likely to belong to different classes. Next, a set of rules is built, in which the left part corresponds to the word pattern, and the right part corresponds to the class label. They are then used to classify text (Aggarwal, Zhai, 2012). This part of the work will focus on the ML approaches to classify text. Machine learning algorithms learn and train on the pre-labelled examples and inherent associations between texts and their labels (Cambria et al., 2022). Some of the most frequently used analytical techniques will be listed and described below.

Naive Bayes is a supervised learning algorithm based on the Bayes theorem widely used for classification problems (Aggarwal, Zhai, 2012). It assumes that all the predictors are independent of each other and therefore apply conditional probability. The logic is that probability of an event is defined as the probability of event A occurring given that event B has already occurred. The advantages of such an approach are its ease of implementation and speed in comparison with other techniques (Bhardwaj, 2020).

Another supervised learning algorithm that can be used for text classification is the Support Vector Machine (SVM). It is used for solving complex problems which use a comparatively big number of features in the model. Support Vector Machine (SVM) is a supervised learning algorithm and has been found to be very effective for a model with numerous features. It can be used as a non-linear and non-parametric technique for solving complex problems. In comparison, in Naïve Bayes, there are strong assumptions about the shape of the data distribution. Nevertheless, the SVM model is not ideal. The main disadvantages are a lack of transparency in results caused by a high number of dimensions and the memory complexity of the model (Barnes et al., 2019).

Furthermore, there is an approach based on Decision trees (Aggarwal, Zhai, 2012) called Random Forest. The RF randomly selects a set of features to choose the best split at each node. The random forest consists of many trees, and the overall prediction is calculated by averaging the predictions from individual DTs (Bhardwaj, 2020). Such algorithms can easily handle categorical features and are comparatively fast for both learning and prediction in comparison with other algorithms. However, there are some disadvantages, as more trees in the model increase the time complexity in the prediction step. Also, the tree is hard to interpret, and the model has a high possibility of overfitting (Barnes et al., 2019). All in all, almost all classifiers can be adapted to the case of text data. Some of the other classifiers include nearest neighbour classifiers and proximity-based classifiers as Rocchio's (Aggarwal, Zhai, 2012; Miron´czuk, Protasiewicz, 2018).

Taking into account all information from the previous chapter, we can make a conclusion that post popularity is considered a good proxy for measuring marketing strategy success in social media. Our data contains information about the number of likes, reposts, and comments on posts, which can be used to track popularity. Likes and comments are the most quantifiable components of any social media platform, including VK. However, there is no universal metric to describe "popularity", so the choice of the measure is a separate subject to study (Brunelli et al., 2021). For this reason, predicting post popularity becomes interesting not only from a business and marketing point of view but also from an academic perspective.

### 2.1.3. Evaluation metrics

An important part of the implementation of ML tasks is the choice of a metric for evaluating the model performance. The choice of a metric depends on many factors: the type of machine learning task, the amount of data, the balance of the sample, the scale of variables and others.

### 2.1.4. Metrics for text classification

The most widely used metrics for classification tasks are F1 score, Precision, Recall and Area Under Curve (AUC) (Minaee et al., 2021).

Precision (1) shows the number of positive samples correctly classified to the total number of positively classified samples in a class (Tharwat, 2020).

$$Precision = TP / (TP + FP) \quad (1)$$

Recall (2), also known as sensitivity, is defined as:

$$Recall = TP / (TP + FN) \quad (2)$$

Recall explains what proportion of objects of a positive class out of all objects of a positive class the algorithm found.

where,

True Positive (TP): Number of positive samples correctly classified;

True Negative (TN): Number of negative samples correctly classified;

False Positive (FP): Number of negative samples incorrectly classified;

False Negative (FN): Number of positive samples incorrectly classified;

In the case of a multi-class classification task, precision and recall should be calculated separately for each class, analyse the individual performance on class labels, or average the values to get the overall precision and recall (Minaee et al., 2021).

The F1 score (3) is the harmonic mean of the precision and recall, controlling the balance between them:

$$F1\ score = 2 * Recall * Precision / (Recall + Precision) \quad (3)$$

F1 score can take values from 0 to 1, where 1 is the best result. The closer the precision and recall values, the higher the F-score is (Basu, Murthy, 2012).

The area under a ROC curve (AUC)

The area under the receiver operating characteristic (ROC) curve is widely used to estimate the predictive accuracy of distributional models (Lobo et al., 2008). Often ROC-curve is used for visualisation, and in order to assess the quality of the classifier, AUC is calculated. The AUC is equivalent to the probability of assigning more weight to a positive class than to a negative one (Fawcett, 2006). The metric has its advantages and disadvantages, but at the moment, it is one of the most common metrics for binary classification tasks, which in addition, is used not only in machine learning but also in medicine and radiology.

### 2.1.5. Metrics for Regression

Traditionally, to assess the adequacy of regression machine learning models, a set of the following metrics is used: R<sup>2</sup>, MSE, RMSE, MAE, and MAPE. R-squared (4) or coefficient of determination is the ratio of the variance that can be explained by the linear model to the total variance (Plevris et al., 2022). The coefficient can take values from 0 to 1, while a value above 0.8 recognises that the model results are reliable. The disadvantage of the coefficient of determination is that it increases with the number of variables, even if they do not increase the predictive power. In order to eliminate this disadvantage, an adjusted R-squared coefficient is applied.

$$R^2 = 1 - \frac{\sum_{i=1}^n (P_i - A_i)^2}{\sum_{i=1}^n (A_i - A_{mean})^2} \quad (4)$$

The Mean Squared Error (MSE) is the most commonly used measure in regression (James et al., 2021). MSE (5) measures the standard error of forecasts, respectively, the lower the value, the better the model. It must be taken into account that MSE is scale dependent and sensitive to outliers.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 \quad (5)$$

The Root Mean Squared Error (RMSE) is preferred to the MSE as it is on the same scale as the data (Hyndman and Koehler, 2006). The metric is calculated as the square root of MSE.

The Mean Absolute Error (MAE) is much less sensitive to outliers, and that is why it is preferred by authors (Armstrong, 2001). MAE (6) measures the difference between two continuous variables (Naser and Alavi, 2021).

$$MAE = \frac{\sum_{i=1}^n |y_i - x_i|}{n} \quad (6)$$

The Mean Absolute Percentage Error (MAPE) measures the extent of error of the forecast in percentage. The limitation of this metric is data with the presence of zeros, due to the sensitivity to zeros, the metric will be unrepresentative, but in this case, weighted MAPE (7) could be applied. Due to the fact that the metric is expressed as a percentage, a clear advantage of MAPE is independence from the data scale.

$$MAPE = \frac{100}{n} \sum_{i=1}^n \frac{|E_i|}{|A_i|} \quad (7)$$

### 2.1.6. Cluster analysis

Clustering is a technique in unsupervised machine learning which groups a set of objects into clusters based on the similarity of information available in the dataset (Bratchell, 1989). As a result, data points in one cluster are more similar to each other than in other clusters. The application of cluster analysis can be found in various spheres and tasks. For example, using data from social networks it is possible to analyse user profiles and identify segments, later gain and provide marketing and operational departments of the company with insights (van Dam, van de Velden, 2015). This type of implementation of cluster analysis is the most suitable for our work. We have mentioned the word “similarity” several times, but what this means in the context of cluster analysis. In many techniques this is based on the concept of distance. K-means and DBScan (Density Based Spatial Clustering of Applications with Noise) are two of the most popular clustering algorithms<sup>2</sup>. K-means is a centroid-based technique, which places all objects in the dataset into a known number of groups based on the location of the closest centroid to a data point. The neighbourhood between a centroid and an object is calculated based on euclidean distance. DBScan is a completely different algorithm, density based. According to this algorithm, each data object has a type (core, border and an outlier) depending on the density between the objects. So that core objects form a cluster, border – identify the ends of the cluster and outliers are eliminated the furthest ones and the noise. In comparison to DBScan, anomalous points in k-means will be assigned to the same cluster as “normal” data objects<sup>3</sup>.

<sup>2</sup> Sinha, “Difference between K-Means and DBScan Clustering| Geeks for Geeks org” URL: <https://www.geeksforgeeks.org/difference-between-k-means-and-DBScan-clustering/>

<sup>3</sup> Sinha, “Difference between K-Means and DBScan Clustering| Geeks for Geeks org” URL: <https://www.geeksforgeeks.org/difference-between-k-means-and-DBScan-clustering/>



## 2.2. Data description and transformations

### 2.2.1. Source of the data

The data source for the analysis is publicly available data downloaded from the VK API. It is an application programming interface that allows developers to access data and functionality from the VK social network. The VK API provides a wide range of methods for accessing data from the site, including user information, groups, and posts. VK API is available to users through a RESTful API interface, with responses returned in JSON format. Firstly, there is a need to obtain an API key, which is required for all API requests. VK also provides documentation and sample code to help get started with the API.

We gathered data from three official groups of GSOM: for bachelor enrollees (gsom\_abiturient), for master enrollees (gsom\_ma), and for the whole GSOM community (gsom.spbu). The uploading of data for the study began in October 2022, hence the data on subscribers groups analysed for the period from October 2022 to May 2023, regarding the information about the groups we have all the history since their creation in 2011. Data uploading from API is organised every week on the same day. To access the necessary data, we used several methods, each of them required to list all the fields that the user would like to extract. The first method (groups.getById) loaded general information on groups, such as their ID, name, contacts, and description. The second method (wall.get) returns a list of posts from the user's or community's wall. Since this method could only load 100 posts at a time, a cycle was coded to load all the posts. We collected posts from all three groups and all of their followers. The third method (groups.getMembers) provided us with a list of the groups' followers. That information allowed us to collect detailed information on each follower, such as their ID, name, and socio-demographic information, by utilising the users.get method. All the information was saved into three folders, one for each group. The data in all these folders are updated automatically on a weekly basis. Inside each folder is the same structure as follows:

- <group\_name>.json (e.g.gsom\_abiturient.json) is the file with the group description;
- members\_full\_group\_<group\_name>.json (e.g. members\_full\_group\_gsom\_ma.json) is the file with the full data for members of the group;
- members\_group\_<group\_name>.json (e.g. members\_group\_gsom\_abiturient.json) is the file with the list of the group's members;
- wall\_owner\_id\_<group\_id>.json (e.g. wall\_owner\_id\_23777199.json) is the file with the wall's data of the group;

- "walls" is the folder with the wall data of group members.

### 2.2.2. Data description

Each of the files listed above contains data with a nested structure. The number of variables in each file is different and reaches up to 250 attributes. An example of what the source data looks like is presented in Figure 4.

```

root
|-- activity: string (nullable = true)
|-- addresses: struct (nullable = true)
|   |-- count: long (nullable = true)
|   |-- is_enabled: boolean (nullable = true)
|   |-- main_address: struct (nullable = true)
|       |-- address: string (nullable = true)
|       |-- city: struct (nullable = true)
|           |-- id: long (nullable = true)
|           |-- title: string (nullable = true)
|       |-- country: struct (nullable = true)
|           |-- id: long (nullable = true)
|           |-- title: string (nullable = true)
|       |-- id: long (nullable = true)
|       |-- title: string (nullable = true)
|       |-- work_info_status: string (nullable = true)
|   |-- main_address_id: long (nullable = true)
|-- age_limits: long (nullable = true)
|-- city: struct (nullable = true)
|   |-- id: long (nullable = true)
|   |-- title: string (nullable = true)
|-- contacts: array (nullable = true)
|   |-- element: struct (containsNull = true)
|       |-- desc: string (nullable = true)
|       |-- email: string (nullable = true)
|       |-- phone: string (nullable = true)
|       |-- user_id: long (nullable = true)

```

**Figure 4** Source data structure example (file with group information).

The data downloaded from the API is redundant for the purposes of our analysis, so only those attributes that are valuable for the current study were selected from each folder. Next, we will describe in detail the variables that will be involved in further analysis.

The main information that we need is located in two folders: the file with the wall's data of the group and the file with the full data for members of the group. From the first file, the following information is taken.

- Text of the post
- Number of views of the post
- Number of reposts
- Number of likes
- Number of comments

- Date of the post, which is in the format (Year - Month - Date - Time)
- Group
- Presence of a photo in the post
- Day of the week
- Time of the day

All of the variables mentioned in this part are self-explainable. However, some of them are still worth commenting on. For example, the number of likes, reposts, comments, and total number of reviews have an integer data type. "Group" is a categorical variable reflecting whether the post belongs to one of the three groups: *gsom\_abiturient* (undergraduate applicants), *gsom\_ma* (masters applicants group), and *gsom\_spbu* (general group of GSOM). The day of the week is also a categorical variable, which is coded as 0 - Monday, and 6 - Sunday. Variable "Time of the day" has four categories: morning - if the post was published from 5:30 a.m till 11:59 a.m, day - time of publishing from 12 p.m till 4 p.m, evening - post time from 4 p.m till 12 p.m and night - from 12 a.m till 5:30 a.m. These two variables were converted from a variable containing information about the date of the post. Finally, the last variable is a dummy, where 1 - a photo is attached to the post, and 0 - the post contains only text. Here the unit of observation is a post in one of the three possible groups. The total number of posts is 4,360. All reposts from one group to another are excluded from the analysis. Despite this, some posts may contain the same text in three groups. They were also excluded, and the first one to post of the group was kept in our analysis.

Data for group members will be used to build dashboards, reflecting our vision of the target audience of the three groups under study. The data contain the following groups of indicators:

*Demographic data indicators*

- Sex
- Date of birth
- Year of high school graduation
- School type
- City of living

### *Behavioural data indicators*

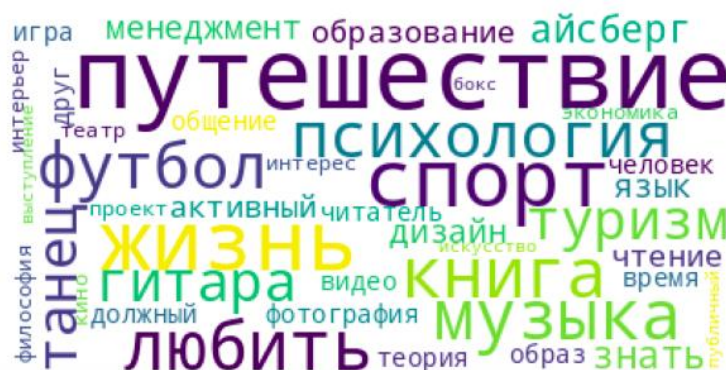
- Connection device
- Date of the last activity session in VK
- Interests

### **2.2.3. Exploratory Data Analysis**

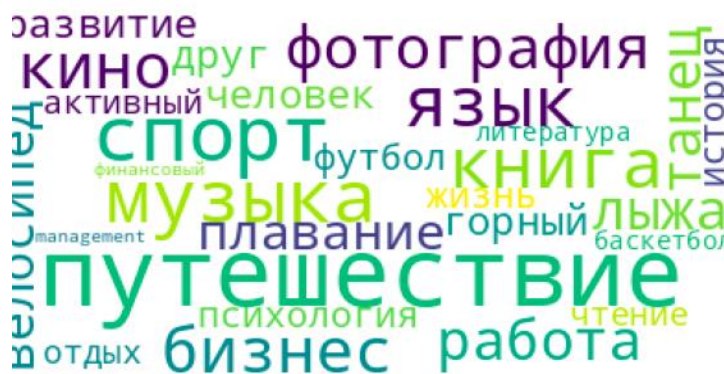
Our study is divided into two parts, in which we separately examine the personal data of group subscribers, as well as the data of the groups themselves.

To clarify the target audience of GSOM groups, not only demographics (gender, average age of subscribers, and region of residence) are of interest, but also data that could tell us about the behaviour of potential subscribers. For example, on the wall, VK users can write about their interests. We decided to take a closer look at it and compare the interests of subscribers of undergraduate and graduate groups and the general GSOM group (Fig. 5, 6; Annex 1). The common interests of all three groups are the same. Subscribers identify "travel", "music", and "books" as their main interests.

Interestingly, undergraduate group subscribers list more interests than graduate group subscribers. Furthermore, among the subscribers of the master's group, there is a more specific wording of areas directly related to work, such as: "management", "business", and "work". Whereas undergraduate group subscribers more often use broader formulations such as: "life", "psychology", "philosophy", "design", "economics", and some others. This suggests that the graduate group audience is more mature and interested in specific areas. Another peculiarity is that the subscribers of the master's group more often mention foreign languages in their interests. For the marketing department, this can be interesting because master's programs offer foreign languages as an elective, while this is not expected in master's programs at other universities.

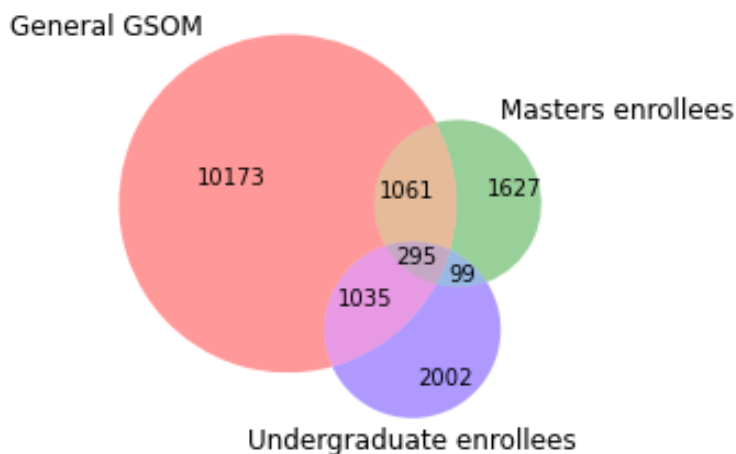


**Figures 5** A word cloud of undergraduate group subscribers' interests



**Figures 6** A word cloud of graduate group subscribers' interests

Also, we were able to compare the number of subscribers in three groups: undergraduate students, graduate students, as well as the general GSOM group (Fig. 7). According to the data for 2023, the largest number of subscribers (12,564 people) is in the general group. Then there are 3,431 subscribers to the undergraduate group and 3,082 subscribers to the graduate/masters group. At the same time, about 50% of undergraduate and graduate group subscribers are also subscribed to the general group. Three hundred ninety-four people are both undergraduate and graduate group subscribers. Most likely, these are GSOM professors and students who have decided to continue their education within the walls of GSOM.



**Figure 7** Distribution of subscribers by groups

Let us move on to evaluating audience engagement among the three groups. In this part, the unit of observation is the post in the group. By May 2023, after cleaning from duplicates and reposts, the group for undergraduate applicants has reached 516 posts, masters applicants - 1250 posts and the general GSOM group - 1997. The table with descriptive statistics (Table 1) shows that the groups have a relatively large number of views of the posts, but the average number of likes is less, around 0,1% of the number of views. This indicates a low rate of audience engagement because the current performance of VK is much lower than the higher education social media

engagement rate benchmark in such social media as Twitter<sup>4\*</sup> (0,058), Instagram<sup>\*5</sup> (2,58%), Facebook\* (0,145%) and Tik-Tok (16,26%)<sup>6</sup>. The closest to VK social media are Facebook and Twitter, but still, the data indicate extremely low engagement of the current audience with the content that is posted in the group. The average for a group of undergraduate applicants is 16 likes and four reposts per post, whereas the average number of likes and reposts per post for master's group applicants is 6 and 1, respectively. Nevertheless, in comparison with the total number of subscribers in the groups, the audience of undergraduate and graduate groups is more active than in the general GSOM group. Interestingly, people tend to comment more in the graduate applicants' group and general group than in undergraduate applicants.

**Table 1** Descriptive statistics of the posts of the 3 GSOM groups

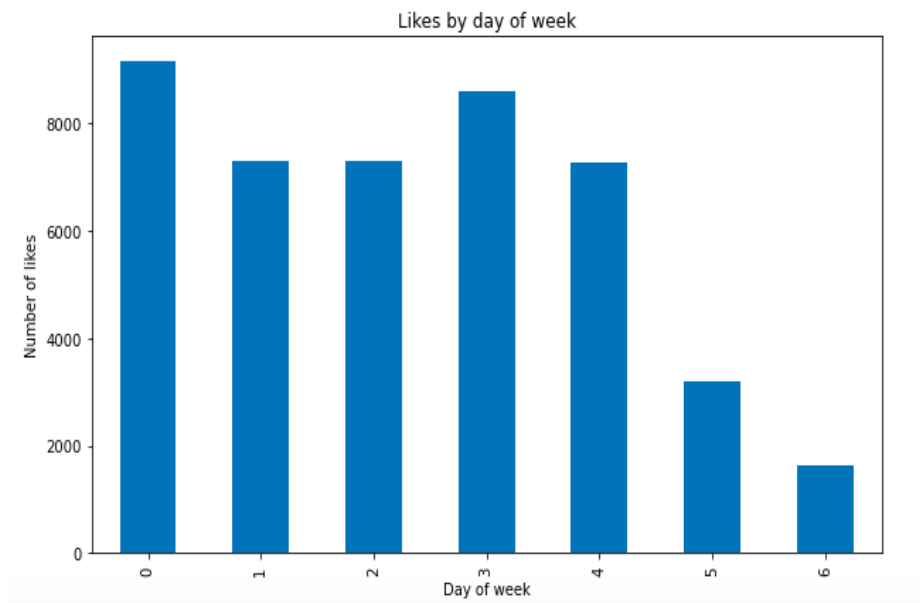
<b>Group</b>	<b>The mean number of views</b>	<b>The mean number of reposts</b>	<b>The mean number of likes</b>	<b>Number of comments</b>	<b>% of posts with photo</b>
gsom_abiturient	1198,7	3,75	16,64	185	29,94%
gsom_ma	576,2	1,5	6,30	513	23,22%
gsom_spbu	1449,0	2,5	11,09	612	22,86%

The number of comments and likes can be heterogeneous and depends on when the post was published on social media. Figure 9 describes the number of likes in posts depending on the day of the week it was published, where 0 - is Monday and 6- is Sunday. For instance, posts downloaded on Monday get more likes than any other day of the week. Thursday and Friday are closing the top three. Also, it is necessary to take into account the time of day when the post was published. According to the number of likes, the most suitable time for the post is in the evening, from 16 to midnight (Fig. 8, 9). The most "successful" days of the week in terms of the number of comments on posts are also considered to be Monday and Thursday. Unlike the likes, there were more comments on Wednesday's posts. The publication time is comparable to the posts that got more likes (Annex 2).

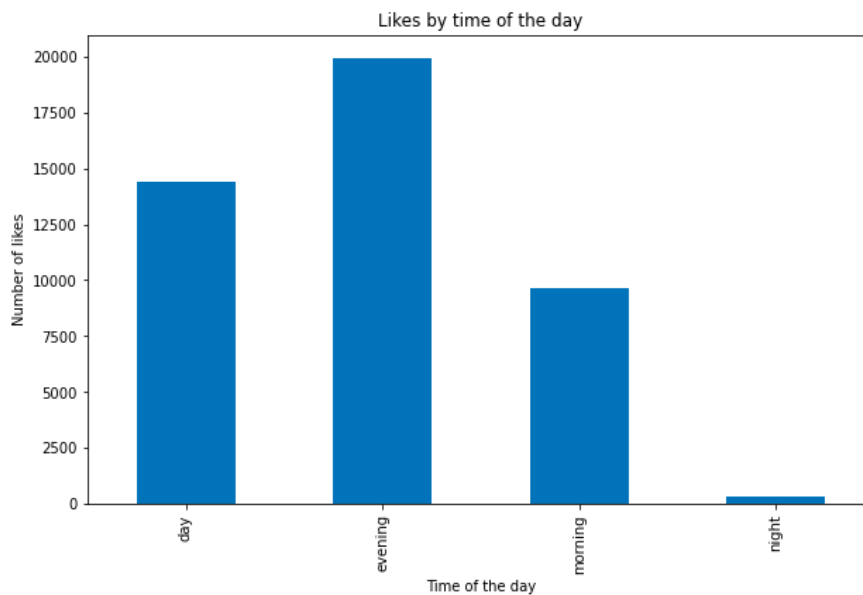
<sup>4</sup> \* Twitter признан экстремистской организацией на территории РФ.

<sup>5</sup> \* Meta Platforms Inc. признана экстремистской организацией на территории РФ.

<sup>6</sup> Feehan, "2023 Social Media Industry Benchmark Report | Rival IQ." URL: <https://www.rivaliq.com/blog/social-media-industry-benchmark-report/#title-higher-ed>



**Figure 8** Number of likes on posts depending on the day of the week (0 - Monday, 6 - Sunday)

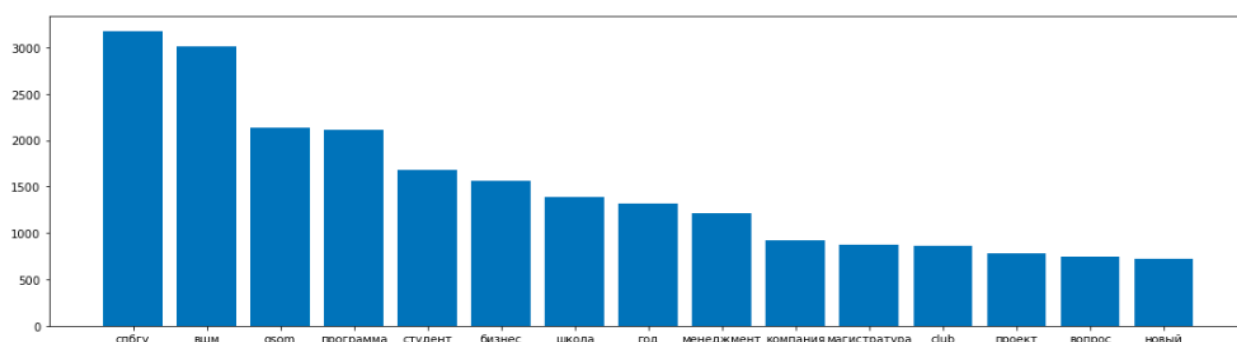


**Figure 9** Number of likes on posts depending on the time of the day

Let us analyse the popularity of posts over time. The Annex demonstrates the plots of reposts and likes of all three groups from 2012 to 2023 (Annex 3). The number of reposts began to increase from the end of 2013, remaining at the same level (less than ten reposts) until 2021. In 2021 there was a sharp jump, which has remained until 2023. The standard number of reposts, except for rare outliers with 100 reposts, is 20. Posts in all three groups typically gain up to 100 likes. Since 2016-2017, they have remained at the same level, with 30 to 50 likes. Also, we have searched and analysed the content of the top 5 most popular posts based on likes, reposts, and comments (Annex 4). The maximum number of likes a post could get is 474. The remaining four popular posts received likes in the range between 200 and 374. It is worth noting that all of the

most "liked" posts were related to the news about GSOM's inclusion in global or Russian ratings. These posts also received the largest number of reposts. For example, the content of such posts was that GSOM is "in the top 25 business schools" or "in the top 1% of business schools in the world". The two most popular posts have the same tag - "#GSOMachievements". The top 5 posts with the highest number of comments represent news about the admissions campaign. For example, posts about the placement of students, admission tests, student ID ceremonies and other information gained from 20 to 37 comments. An analysis of the most popular posts also showed that the highest audience engagement by the number of likes was in 2017-2018 and by the number of comments in 2020-2021.

Before building the models, and diving deeper into the analysis of the posts, the most frequent words in the posts were derived (Fig. 10). As can be seen, the words directly related to the university and the department - "spbu", "gsom", and "program" - have the greatest number of mentions. They will need to be excluded or removed using the min-max df parameters when applying tf-df in the preprocessing of the text of the posts to exclude their influence on the engagement.



**Figure 10** Top-15 most popular words in the posts of 3 groups

## 2.3. Dashboards for tracking main estimators

### 2.3.1. Tool description

There are a lot of different tools for building dashboards, such as Power BI or Tableau. There is also a cloud-based data visualisation platform that is available on the Russian market, unlike the ones listed before, from Yandex - Yandex.DataLens<sup>7</sup>. It is a powerful tool that offers a wide range of features for data exploration, analysis, and presentation. One of the key benefits of Yandex.DataLens is the ability to handle big data with ease. The platform is designed to work with large, complex datasets like ours and has a wide variety of source connections. This makes it ideal for the purposes of our study.

<sup>7</sup> Official website of Yandex.DataLens. URL: <https://cloud.yandex.ru/services/datalens>



To connect our dataset to Yandex.DataLens, we used another service provided by Yandex - Yandex Query. It is an interactive data visualisation service that enables real-time streaming queries to structured and partially structured data using Yandex Query Language, a dialect of SQL<sup>8</sup>. The interconnectivity of the two tools allows for smooth and easy data updates as well as does not require any data preprocessing.

Yandex.DataLens also offers a wide range of visualisation options, including charts, graphs, tables, and maps. These highly interactive visualisations allow users to explore data in real-time and drill down into specific areas of interest. This makes it easy to identify data trends, patterns, and anomalies and communicate insights to stakeholders.

Overall, Yandex.DataLens is a powerful and flexible platform for building dashboards and reports from complex datasets from different sources. Its ability to handle big data and its wide range of visualisation options are the main reasons to use this tool for our study.

### 2.3.2. Dashboard description

As the first step, an interactive prototype of the dashboard was created in Figma. The example of the dashboard prototype for the community for bachelor enrollees is presented in Figure 11. Initially, it displayed visualisations that represented socio-demographic information on the group's followers as well as the main metrics, such as the number of followers, average number of likes, reposts, and other indicators.

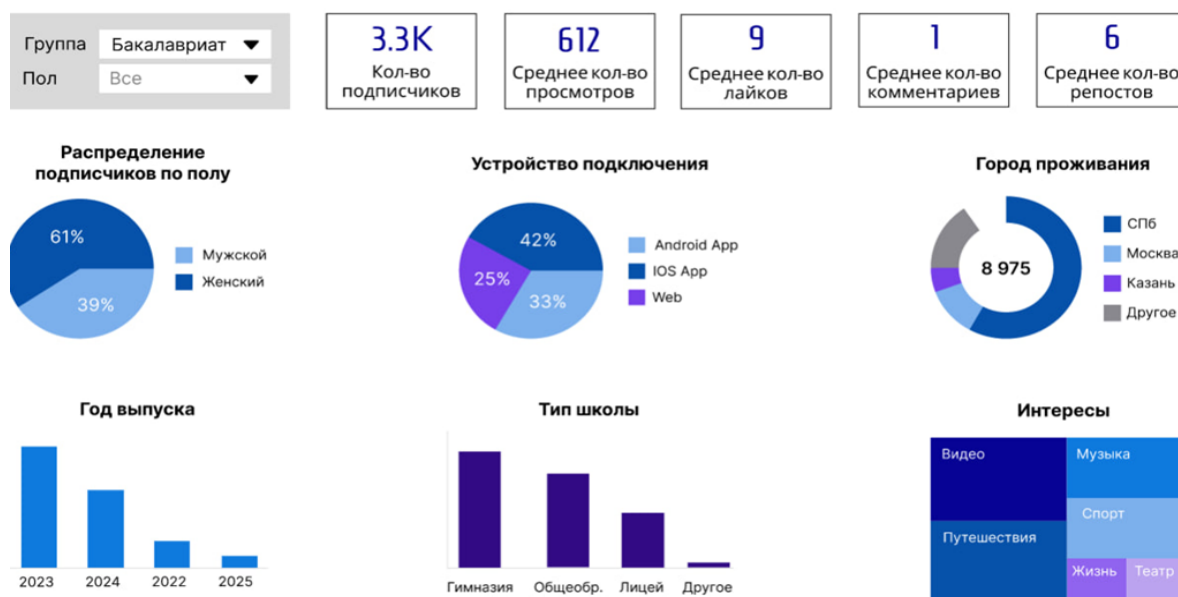


Figure 11 Prototype of the dashboard

<sup>8</sup> Official website of Yandex Query. URL: <https://cloud.yandex.ru/services/query>

After consultation with the marketing department, it was decided to leave only the essential information on the groups' followers. The dashboard will serve the marketing department as a tool for tracking the target audience. For instance, one of the other features of the dashboard is a display of the proportion of people who connect from a laptop or mobile phone. This information can be used for optimising the content for different platforms and devices. Furthermore, the dashboard includes a filter - for the gender of members.

The information represented in all the groups is the same, with the only difference being that for the groups for bachelor enrollees, we can see the types of schools the enrollees have finished instead of the university, as is the case for the other two groups. The example of the final dashboard on the community for bachelor enrollees is presented in Figure 12. The screenshots of the other two groups' dashboards can be found in Annex (Annex 5).



**Figure 12** The dashboard of the group for bachelor enrollees

### 2.3.3. Managerial use

The dashboard will be an essential tool for the marketing department of GSOM for several reasons. First of all, the dashboard will provide deeper insights into the behaviour and preferences of different audience segments. For example, it will allow the department to see who is interested in applying to GSOM and their socio-demographic characteristics. This information will be invaluable for tailoring the content to better resonate with the target audience, increasing engagement, and ultimately driving conversions. Secondly, the marketing department can see how active the audience is and the flow of newcomers from month to month.

Overall, the dashboard is a powerful and versatile tool that will enable the marketing department to make data-driven decisions and optimise the groups' content for maximum impact.

By leveraging the insights the dashboard provides, the department will be able to identify growth opportunities and ultimately achieve its marketing goals.

## **Summary of Chapter 2**

In this chapter, the methodologies for processing text data based on the NLP approach were considered: the necessity of the method was justified, the main components and their order were determined, and the purpose of each step was described. Then the proposed preprocessing technique was tested on the texts of the posts of the GSOM VK group for subsequent use as a feature in the machine learning model. Further, the data structure from the VK API was analysed in detail, and methods and tools for working with data were described. Exploratory data analysis was also conducted to obtain a comprehensive understanding of the available data and their characteristics. As a result, it was revealed that the interests of subscribers are concentrated around such areas as travel, photography, cinema, music, sports and business.

Continuing to analyse the target audience, we have created a dashboard, which allows us to monitor the performance and socio-demographic characteristics of the audience in 3 groups, taking into account the number of active and inactive users. The dashboard will be a useful tool for the marketing department both for defining the target audience and content planning.

In the next chapter, we will discuss the application of the methods described above. In particular, we are using TF-IDF for cleaning and processing data and various models, including regression and classification methods, that are used in our research.

## Chapter 3. Building Machine Learning models for VK data processing

### 3.1. Dataset preparation for model building

#### 3.1.1. Data transformations and data cleaning

Previously we mentioned the redundancy of some data and vice versa, the problem of missing values, as well as the fact that some variables are challenging to interpret. All these problems and many others need to be solved with the help of data preprocessing. The primary data tool used in this study is the Python programming language, and the main libraries for working with data. Accordingly, all the following steps described with data manipulation are done using Python.

First of all, we cleared the data from users with inactive accounts, assuming that the inactive one can be considered the one who logged into the social network more than one year ago. Furthermore, analysing the posts of the groups, it was discovered that in the early stages of the existence of the GSOM group in VK, they were unofficial, and anyone could post their questions, comments and notes as a post on the group's wall. In this regard, based on the difference between the id of the author and the owner of the post, all posts made not on behalf of the official GSOM community were deleted because this data is irrelevant for the purposes of the current study.

A significant part of the work with data in our study is devoted to the analysis of text data. In this regard, text data preprocessing was an integral part of the work. Data preprocessing was carried out according to the methodology described in paragraph 2.1.1 Neural Language Processing (NLP) techniques for text preprocessing.

Further, we will consider an example of text processing of group posts. For a more detailed description, let us give an example of how the initial text of posts looks in the dataset:

*“📅Календарь мероприятий для абитуриентов бакалавриата \n \n1 марта в 17:00 \nОнлайн-презентация программы «Менеджмент» \n🔗Регистрация: <https://vk.cc/ckADRW>\n \n2 марта в 17:00\nОнлайн-презентация программы «Государственное и муниципальное управление»\n🔗Регистрация: <https://vk.cc/ckADTy>\n \n3 марта в 17:00\nОнлайн-презентация программы «Международный менеджмент» \n🔗Регистрация: <https://vk.cc/ckADUS>\n \n25 марта в 12:00\nДень открытых дверей программ бакалавриата в кампусе «Михайловская дача» \n🔗Регистрация: <https://vk.cc/ckADWy>”*

Evidently, text in this initial state could not be used for analysis without preprocessing because the following problems are apparent: the presence of emojis, numbers, symbols, links, also frequently repeated words that do not carry meaning and different registers. To eliminate these

problems, a function (Fig. 13) was written for preprocessing the text of posts, taking into account its features.

```
def preprocessing(sentence, as_list=False):
    s = re.sub('[^а-яА-Яа-зА-З]+', ' ', sentence).strip().lower()
    s = re.sub('ё', 'е', s)
    function_words = {'INTJ', 'PRCL', 'CONJ', 'PREP'}
    lemmatized_words = list(map(lambda word: MORPH.parse(word)[0], s.split()))
    result = []
    for word in lemmatized_words:
        if word.tag.POS not in function_words:
            result.append(word.normal_form)
    result = [w for w in result if w not in STOPWORDS]
    if as_list:
        return result
    else:
        return ' '.join(result)
```

**Figure 13** Text preprocessing function

The function contains the reduction of all words to a single lowercase, part of speech tagging, lemmatization, clearing stopwords, and removing characters, emojis and punctuation. After applying this preprocessing function to the text of the post given in the example earlier, the result was as follows:

*“календарь мероприятие абитуриент бакалавриат март онлайн презентация программа менеджмент регистрация https март онлайн презентация программа государственный муниципальный управление регистрация март онлайн презентация программа международный менеджмент регистрация март день открытый дверь программа бакалавриат кампус михайловский дача регистрация”*

The resulting text can be used for vectorisation and subsequent application in machine learning models since it does not contain noise and redundant features.

### **3.1.2. Train-test split**

Before building any model, the dataset has to be divided into samples. Typically, it is divided into train and test samples. In some cases, the initial dataset can be divided into three parts instead of two: train, test, and validation. This procedure is taken in order to assess the prediction quality of the model. The training sample is used to build the initial model, then "training" it by adding several features and adjusting the model's parameters. A validation sample is usually used to test different models and compare them. The test sample is used at the end of the analysis to evaluate the model performance on the absolutely unseen data and understand how the model will work in real life.

When building a model, as well as splitting the dataset, it is necessary to take into account the specifics of the analysed data. As we are aggregating the data from the posts of groups of applicants, we should take into account the duration and schedule of admission campaigns in the university and divide the sample not randomly, but the date sequence should be maintained. Our data is also split into three parts: train, validation and test (Table 2.)

**Table 2** Training-validation-test split

Dataset	Time period	Number of observations
Train dataset	2011-03-21 - 2021-03-31	2744
Validation dataset	2021-04-01 - 2022-03-31	401
Test dataset	2022-04-01 - 2023-05-24	618

### **3.2. Modelling the dependencies between likes and post characteristics**

One of the important pieces of knowledge that we want to get is what influences the number of likes on posts, as it is one of the primary metrics of user activity in interacting with content. Posts have quite a lot of characteristics, and there may be completely non-obvious connections between the time and the day of the week of posting and even the presence of emojis in the text and a higher number of likes. In this regard, first of all, in our research, we will build a model to determine the relationship between the number of likes and the characteristics of the post, such as the presence of a photo, the time and day of the post, and tonality. We will also try to find out which words and phrases impact increasing user activity most.

#### **3.2.1. Selection of the target variable**

The social network VK has a lot of data about the groups: we were able to unload the number of likes, comments, reposts and views of the post. However, among the literature we have studied, there is no single correct model to follow to estimate the text's influence on the post's popularity, nor are there any rules for how the dependent variable - the popularity of posts or engagement- can be measured. This diversity of approaches provides room for our own experiments.

As a target variable to measure engagement, it is possible to take the absolute number of likes and comments, depending on the original dataset's data (Cordero-Gutiérrez and Lahuerta-

Otero, 2020). In addition to absolute values, a composite index can be used to determine the dependent variable. For example, the sum of likes and comments is divided by the number of users (Purba et al., 2021; Yost et al., 2021). However, as Brunelli et al. (2021) note, such a measurement is standard for normalisation across various individual profiles and is not meaningful and effective in modelling the behaviour of a single profile or, as in our case, just three groups. Instead of dividing by the number of followers, some authors suggest dividing the average number of comments and likes of the post by the number of views. Such an index is appropriate and could be used in our studies; however, our dataset has some limitations. In VK, the number of views in the post appeared only in 2016, whereas groups for undergraduate and graduate applicants have been publishing official information since 2012. Thus, the data on the number of views from 2012 to 2016 contain many missing values, replaced by 0, which makes it impossible to compile an index in cases where the posts collected likes and comments but not views. It should be noted that using comments or reposts as a targeting variable is not possible in the case of our data. The reason for this is the critically low number of comments and reposts, and the fact that most of them are made by people associated with the GSOM, such as employees. The main reason for the low number of comments and reposts we believe is the official tone of the communities and the narrowly focused content that reports only important news, events, and programs, but does not contain entertainment content or different headings that would increase activity under the posts.

Another possible approach is to formalise the problem of predicting post popularity as a binary classification problem and turn the variable of possible interest, for example, the number of likes, into a dummy: 1 - the like is present in the post, 0 - the post has gained no likes at all. This approach does not limit us in the data; we can use it for all periods presented from 2012 to 2023, and it allows us to try another machine learning model called logistic regression and experiment with building a dependent variable, using median, mean or a specific number of likes as the border of defining the dummy variable.

### **3.2.2. Sentiment analysis**

Before proceeding directly to building regression and classification models, we would turn to one of the popular and effective text analysis tools that can provide additional valuable information about the data. VK is a big social media platform where posts are the most common way to share information. One of the main components of any post is its text, which can be written in different ways and have a different impact on the audience. To better understand and analyse the content of social media posts, sentiment analysis is commonly used. Sentiment analysis is a technique that involves identifying and extracting emotions, opinions, and attitudes expressed in the text. In our analysis, we use Bidirectional Encoder Representations from the Transformers

(BERT) model, because it is one of the most powerful NLP algorithms. It provides a better understanding of words and sentences in the context, which is why it is perfect for sentiment analysis. BERT is a language modelling system which is pre-trained with a huge corpus and huge processing power and is used in such tasks as translators or chatbots.

The first step in performing sentiment analysis is to identify the language used in the post. In all the posts of the three analysed groups, the language used was identified as Russian, which was an expected result. After identifying the language, the posts were labelled based on their tonality, which can be negative, neutral, positive, or speech. If the model was not able to identify the tonality of a post, it was labelled as "skip". Additionally, sentiment probability was defined as a number between zero and one, which reflects the level of certainty in the tonality assessment. It was observed that most of the posts analysed had a neutral tonality. This could be due to the official nature of the analysed texts (Fig. 14). Additionally, the result of the sentiment analysis can be used as a variable in future models, as it gives additional information about posts - text tonality.

language	sentiment	probability
ru	neutral	0.998448
ru	neutral	0.998482
ru	neutral	0.997841
ru	neutral	0.998150

**Figure 14** Example of sentiment analysis results

### 3.2.3. Feature selection

The basis for building dependencies between posts and user activity expressed in the number of likes in our work are words and phrases from the text since, in the study, we made the assumption that these features are the most influential.

Along with the text, we decided to include in our model additional features. For example, learning from the EDA part, it could be stated that the day of the week and time of posting can influence the engagement rate of the followers. Furthermore, it is better to take into account the type of group from which a post was published as they had different amounts of subscribers and activity within the groups varies. All these variables are categorical and were added to the model one by one as binary variables. Moreover, nowadays, it is popular to include in the post some media files such as photos or videos. It is believed that these kinds of posts engage the audience more and make the feed visually more pleasing. We also decided to add to our model binary



variables reflecting the presence of a photo or video in the published post. A complete list of features about the posts available for analysis according to the data structure from VK API:

- Text of the post
- Time of the day (morning, day and evening)
- Day of the week
- Month
- Presence of the photo in the post
- Presence of the video in the post
- Text tonality

However, the choice of features should be justified by the performance of the models. In connection with this, experiments were carried out during which features were excluded one at a time, and the impact on target metrics was calculated. Since the study will use both regression models and binary classification models, the RandomForest model was used to select features in both cases. The choice of the final combination of features was based on changing the target metrics, namely AUC ROC for the binary classification model and WMAPE for the regression model. In both cases, the level of change of the metric was chosen empirically by less than 0.01 in order to make a decision about its insignificance for the model. As a result, the combination of features that lead to the best performance of the model is the text, time of day, day of the week and the name of the group where the post is taken from.

#### 3.2.4. Binary classification task

To build a binary classification model, first of all, we need to convert our target variable, "number of likes", into a binary one. The distribution of likes in the dataset is reflected in table number 3.

**Table 3** Descriptive statistics about the number of posts' likes

Descriptive statistics	Number of likes
Min	0
25%	3
50%	7
75%	15
Max	474

Empirically based on several experiments with different levels of thresholds, the first quartile was chosen as a threshold value to create a binary variable "like presence" taking values 0 and 1. As a result, the ratio of the variable with a value of 0 to 1 in the train set was distributed as 45% and 65%, respectively. Hence the data is nearly balanced. The models used in the study for a binary classification problem were Logistic Regression, RandomForest and Light Gradient Boosting Machine (LGBM). As described in the previous paragraph about the choice of features, we selected a set of variables that participated in the experiments. However, for each model, we conducted an experiment on two sets of variables: exclusively on text data and text data with the addition of selected features (time of day, day of the week, and the name of the group).

### 3.2.5. Baseline model

As a baseline solution for modelling the dependencies between likes and post characteristics, logistic regression will be used. The results of logistic regression can be seen in Table 4. One of the primary metrics here is a ROC-AUC score that shows how much the model is capable of distinguishing between classes: 69% of the data in the second case with additional features, the model predicts the presence or absence of like in the post correctly, which can be treated as a good performance. There is a noticeable difference between the metrics in the first and second cases: the model with additional features showed higher results. The recall has a really high level, demonstrating the high sensitivity of the estimator. In general, even the baseline model has already shown quite a high performance.

**Table 4** Logistic regression model performance on test data

<b>Evaluation metric</b>	<b>Only text features</b>	<b>Text with additional features</b>
ROC-AUC	0.55	0.69
Accuracy	0.85	0.87
Precision	0.87	0.89
Recall	0.96	0.97
F1-score	0.92	0.93

### 3.2.6. RandomForest Classifier

The more advanced Random Forest model is based on decision trees, which require the selection of the most optimal parameters, not only to improve the performance of the model but also to prevent overfitting.

First of all, GridSearch with Time-series cross-validation was launched to select the optimal hyperparameters for the model. The use of time-series cross-validation is justified by the nature of the data used in our study. Time series data cannot be cross-validated directly since mixing data will lead to a violation of the chronological order and, most likely, model overfitting. GridSearchCV is an approach to parameter search provided by the scikit-learn library in Python. By "fitting" parameters to the dataset, all possible combinations of parameter values are evaluated, and the best combination is stored. The input parameters in our case were `n_estimators = [50, 100, 150, 200]`, `max_depth = [2, 3, 5, 7, 10]` (Fig. 15).

```
param_grid_forest = dict(
    n_estimators=[50, 100, 150, 200],
    max_depth=[2, 3, 5, 7, 10]
)

grid_search_forest = GridSearchCV(
    estimator=forest,
    param_grid=param_grid_forest,
    cv=tscv,
    n_jobs=-1).fit(X_train, y_train)

print("The best hyperparameters are ", grid_search_forest.best_params_)
```

**Figure 15** GridSearchCV for Random Forest

As a result of applying GridSearchSV, the most optimal number of estimators for RandomForestClassifier turned out to be 50 for the text-only dataset and 200 for the second case with additional features, and the maximum depth of the tree was 10. So, the model was built with the optimal hyperparameters found and a random state equal to 42. The results of the metrics are presented in Table 5. As expected, the more advanced model gave better results. Recall equal to one gives an optimistic result that all positive cases the estimator correctly predicted over all the positive cases in the data, but at the same time, the accuracy is 0.87, which means that some observations from the negative class were labelled as positive class.

**Table 5** RandomForest model performance on test data

<b>Evaluation metric</b>	<b>Only text features</b>	<b>Text with additional features</b>
ROC-AUC score	0.58	0.76
Accuracy	0.87	0.87
Precision	0.87	0.87
Recall	0.99	1.00
F1- score	0.93	0.93

The third model used for the binary classification problem in our study is Light Gradient Boosting Machine, which is a more powerful model also based on decision trees. Similar to the previous model, GridSearchCV was first applied, which revealed the following most optimal hyperparameters: the number of estimators equal to 100 for the text-only dataset and 200 for the second case with additional features, and the maximum depth of the tree was 10 for both. The performance of the LGBM model on the test data turned out to be slightly weaker than in the RandomForest model. However, the table shows metrics only for the test dataset, while on the validation dataset, the metrics are almost the same, and on the training dataset, LGBM performance metrics are higher (Annex 6). When choosing the best model for our study, we will focus on LGBM since it is a more efficient, stable and powerful model.

**Table 6** LGBM model performance on test data

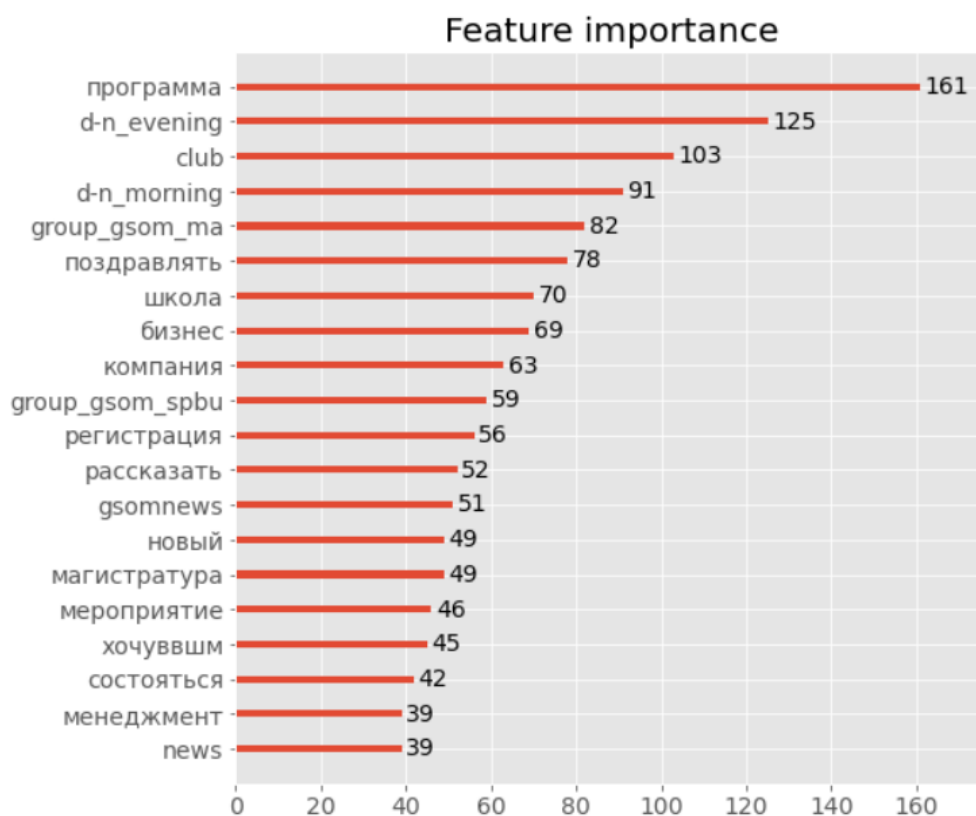
<b>Evaluation metric</b>	<b>Only text features</b>	<b>Text with additional features</b>
ROC-AUC score	0.58	0.64
Accuracy	0.83	0.85
Precision	0.87	0.89
Recall	0.95	0.95
F1-score	0.91	0.92

In general, the results of the models are quite high, and the models are able to predict with great accuracy whether a post will have less than three likes or more. Table 7 presents the results of binary classification models on a test set with additional features. However, it is important to note that all models also performed well on a text-only dataset, which means that the main ability of the models to classify posts is based on the words and phrases contained in the text of the posts and not on the time of the post or a specific day of the week.

**Table 7** Classification models performance comparison

Evaluation metric	Logistic Regression	Random Forest	LGBM
ROC-AUC score	0.69	0.76	0.64
Accuracy	0.87	0.87	0.85
Precision	0.89	0.87	0.89
Recall	0.97	1.00	0.95
F1-score	0.93	0.93	0.92

An important part of researching simulation results is the importance of features. Figure 16 depicts the feature importance according to the results of the LGBM model. Analysing the graph, we can conclude that higher activity of users is manifested in posts about the description of programs, registration for events, congratulations and a news section. It is also worth noting that the master's program is more interesting for users relative to the undergraduate program, according to the results of the model. In addition to the words in the posts, the number of likes is also affected by the time of the post - the morning and evening of the day are the most suitable time for posting. However, if we return to the "Feature Selection" part, where the contribution of each variable to the performance of the model was analysed, then the setting time is still much less influential in relation to the words and phrases contained in the text. Therefore, the emphasis in the marketing strategy should be on the content and topics of posts in the GSOM VK groups.



**Figure 16** Feature importance graphic for LGBM model.

### 3.2.7. Regression

As an experiment in model work, we decided to add another class of models, namely regression models. Random Forest, XGboost and LGBM regression models were built to explain the relationship between words and two-word phrases in posts and the number of likes on those posts. The same additional features were included in models as ones in the Binary Classification task. Also, the same algorithm - GridSearch with time series cross-validation was also launched to select the optimal hyperparameters for the model. The procedure and the output example were shown in the previous part. As a result of applying GridSearchCV, the most optimal number of estimators for RandomForestRegressor turned out to be 200, and the maximum depth of the tree was 10. So, the model was built with the optimal hyperparameters found and a random state equal to 42. For the XGboost model, the hyperparameters are slightly different: the learning rate is added as it determines the step size at each iteration while moving to the minimum of the loss function (Murphy, 2012). According to the GridSearchCV, the model was built with the following input hyperparameters: learning rate = 0.01, maximum depth = 4 and 100 estimators. The LGBM model was applied with default hyperparameters. The summary of the performance of the regression models is presented in Table 8.

**Table 8** Classification models performance comparison

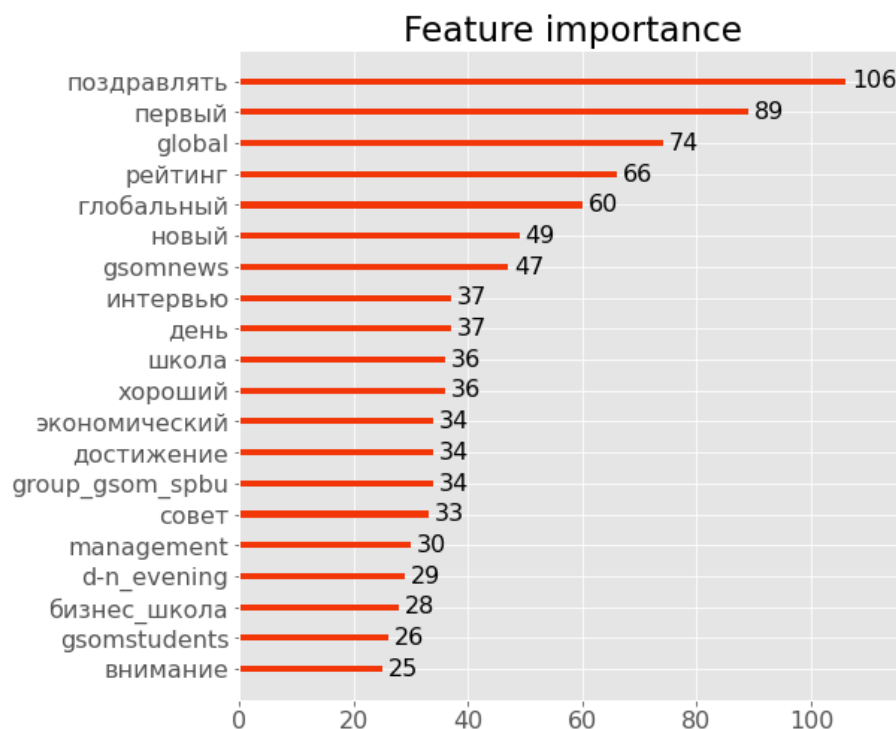
<b>Evaluation metric</b>	<b>Random Forest with additional features</b>	<b>XGboost with additional features</b>	<b>LGBM with additional features</b>
RMSE	17.12	18.352	18.235
MAE	9.449	9.921	10.989
WMAPE	0.583	0.611	0.678
R2	0.13	0.001	0.014

The set of metrics is different for Classification task models and traditional Regression because, in the second case, the dependent variable is a numerical one - the number of likes in the post. To assess the adequacy of regression machine learning models, a set of the following metrics is used: R2, MSE, RMSE, MAE, and MAPE.

Evaluation metrics are barely different from one model to another: RMSE, MAE and MAPE are slightly higher in XGBoost and LGBM than in Random Forest. However Random Forest Model has the highest R2 score based on the test sample. All in all, R2 is still critically low. Firstly, we thought that models could really be suitable for building the dependence of likes on the text of posts, but the text alone is not enough, and it is necessary to introduce additional independent variables. However, after including additional features to the models, their performance indeed increased but very slightly.

At the moment, the most effective model is LGBM. In this regard, it is on this model that we study the influence of the words of posts on the number of likes. We have built and visualised feature importance, and the results indicate that additional features such as type of group and time of posting in the evening have an influence on the engagement of the audience. As can be seen from Figure 17, the most influential words are "rating", "congratulate", and "first". This makes sense since, most likely, the posts are related to the fact that GSOM took place in some international university rankings. Among the words which appeared to be triggering and appealing for likes is also information about any kind of news and activities. However, we could not take

these results seriously as the performance of the models is poor. Combining various features, we tuned parameters that improved the models. Comparing models built on regression tasks with classification tasks turns out that they are less suitable for our analysis.



**Figure 17** Feature importance graphic for the LGBM model

### 3.3. Cluster analysis

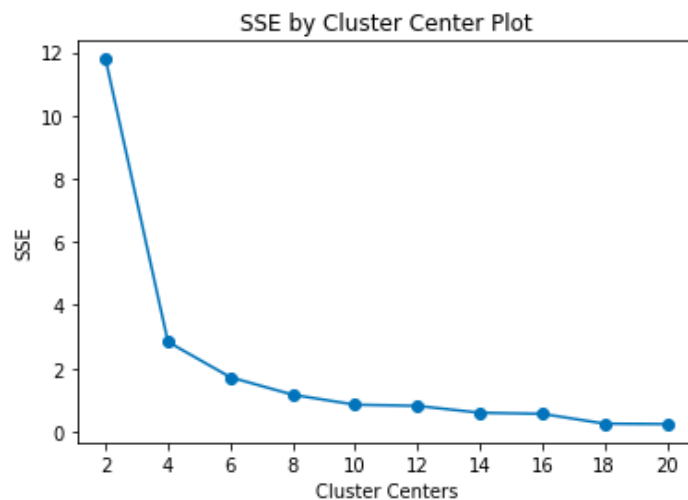
Analysis of the posts was able to tell which factors and words increase user engagement. To deepen our understanding of the target audience, we decided to do a cluster analysis of users subscribed to VK groups. Segmentation helps to bring “customers” together and not to analyse each user individually but to select marketing strategies for groups and build communications with them. For example, a group can be subscribed to by enrollees, parents of incoming students, and teachers - all these imply various methods of communication. We also assumed that information from users' pages (interests, posts and reposts on their own pages) could be signals about people's preferences. In this way, we could find out the preferences of segments.

It was decided to conduct several experiments with clustering. Firstly, different numbers of hyperparameters were tried. For example, in the first iteration, clusters were built by gender and age of subscribers separately for each of the groups: undergraduate applicants, graduate applicants, and subscribers of the general GSOM group. In the second iteration, clustering by three groups was conducted according to 3 parameters: gender, age, and city (three categories - St. Petersburg, Moscow, and other cities). Also, we have tried two methods for clustering: k-means and DBScan.

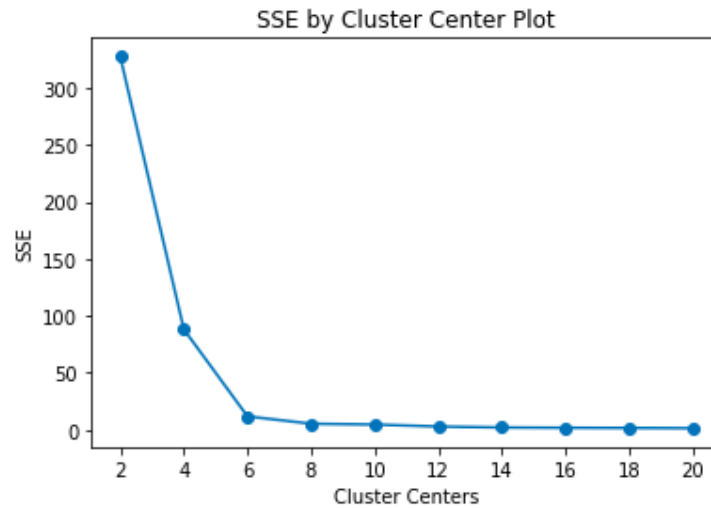


The first partitions all the points in the sample space into  $K$  groups of similarity, which is usually measured using Euclidean Distance. The second technique is a set of density-connected objects that is maximal regarding density-reachability. The difference is that in k-means, every object is assigned to a cluster, while in DBScan, each object not contained in some cluster is considered to be noise. The following results will be described only for the k-means algorithm, since k-means gave, as a result, clusters with clearer boundaries by age and accurate distribution by city, while the DBScan algorithm formed 2 clusters (out of 6) that were with mixed data by age and city, hardly interpretable.

In our analysis, we use variables which are not directly comparable (sex, city, age). Years of life have a greater range than categories of sex and city of living (because the field with the wider range of values likely has greater distances between values). As it may end up being the primary driver of what defines clusters, standardisation is needed to make the relative weight of each variable equal. So, the parameter age was standardised, while dummy variables were created for each of the categories of city and sex and included in the model. Then with the help of elbow-method, we were able to determine the optimal number of clusters with models of 2 hyperparameters - sex, years old (Figures 18, 19; Annex7), and three hyperparameters - sex, years old, city of living (Annex 8, 9) for all of 3 groups. It turned out that for the model with two hyperparameters for all three groups, the optimal number of clusters equals 4, while for the model with three hyperparameters - the optimal number is 6.



**Figure 18** Elbow method for defining the optimal number of clusters for gsom\_abiturient group, k-means with two parameters: sex, years old.



**Figure 19** Elbow method for defining the optimal number of clusters for gsom\_abiturient group, k-means with three parameters: sex, years old, city of the living.

After fitting k-means and matching the results with data, we were able to attach to the resulting clusters the following data reflecting GSOM subscribers' preferences: interests, posts, and reposts collected from personal pages. All text entries were preprocessed using TF-IDF, as in the analysis of the VK group posts themselves.

Results indicate that the interests of men and women subscribed to GSOM groups differ. Business interests related to “investments”, “real estate” (Table 9), “internet technology”, “design”, and “programming” (Annex 10) are more common among men subscribers. Among women, some general interests are more widespread: “psychology”, “public relations” (Annex 10), “sustainable development”, and “public speaking relations” (Table 9). Among all clusters within three groups (undergraduate enrollees, graduate/masters enrollees, subscribers of the general group), there are words - "foreign languages", "reading books", "travel" as well as interests indicative of an active lifestyle of subscribers: sports (“soccer”, “taekwondo”, “dancing”, etc.) and various hobbies (“drawing”, “photography”).

**Table 9** Interests of undergraduate enrollees group, clustering method: k-means, parameters: sex, years old

Cluster 1	Cluster 2	Cluster 3	Cluster 4
men, 15 - 38 years	men, 39 - 63 years	women, 35-83 years	women, 15-27 years

works education taekwondo movie music vocals singing scandals intrigue consequences investment investments real estate creation	space robotics business	languages texts public relations universities educational projects internationalisation higher education trips learning Finnish medicine	love dance trips gastronomy winemaking Latin American dancing sustainable development ecology public performance psychology business
--	-------------------------------	--	---

Subscribers among the three groups may have overlapped, and it is also worth mentioning that a small number of users indicated their interests on their own pages. That is why the results of the analysis of the 6 clusters, taking into account the differences in the city of residence, could only be plotted among subscribers of the general GSOM group because it has the largest number of subscribers (Table 10). However, as seen in Table 10, the results are not very different from those described above. The interests of subscribers by city correspond to the distribution by age and gender. Also, thanks to the segmentation of users, we were able to identify accounts whose interests include the word “work”. It turned out that these users are engaged in writing term and dissertation papers for money, and they are not the target audience of the groups studied. These accounts appear in the graduate enrollees group and in the general group (Annex 10, 11) and appear to be both men and women but in the same age gap (15-34 years old).

**Table 10** Interests of general GSOM group subscribers, clustering method: k-means, parameters: sex, years old, city of the living

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
men Other cities 15 - 98 years	men SPb 15-97 years	women Moscow 18-58 years	women SPb 15 -83 years	men Moscow 17-64 years	women Other cities 15 - 78 years

music	music	trips	trips	music	art
trips	art	art	psychology	art	trips
sport	sport	Internet	People	trips	dancing
psychology	literature	phones	music	social	psychology
Internet	movie	mobile	dancing	development	culture
football	economy	computers	works	urban	books
health	story	connected	hiking	market	yoga
family	dancing	technique	nature	business	languages
business	new	gastronomy	summer	policy	self-development
marketing,	development	winemaking	smiles	gadget	music
tourism	business	dancing	children	medium	innovation
seminars	high	sustainable	books	entertainment	People
trainings	Internet	development	adventures	Web	tasty food
education	beautiful	ecology	walks	development	
movie					

Based on the data of subscribers' posts on their profile pages, we also made an attempt to analyse the most frequent words for each of the six clusters. However, as a result, the words collected from users' pages turned out to be too general and not focused on specific topics, so for any cluster, this analysis did not provide additional information that could complement the characterization of subscriber segments (Annex 12). However, despite the lack of additional information, this analysis also yielded a result, namely that no segment of subscribers concentrated on specific topics, and subscribers showed interest in a variety of topics from different types of groups.

### Summary of Chapter 3

In this chapter, we selected features for likes and post dependency models, analysed post sentiment, and finally provided many experiments with models, trying to explain the relationship between post text and the number of likes from users in GSOM VK communities. As a result of the sentiment analysis, it was revealed that the vast majority of posts are neutral in nature, and the rest are positive. This result was quite expected due to the official manner of communication in the GSOM communities. After conducting a series of experiments, we have arrived at some preliminary conclusions regarding our research. First of all, we have determined that regression analysis is not suitable for our purposes. As a result of experimenting with classification models in an attempt to model the relationship between the characteristics of posts and the level of activity of users, it was found that the subject of the post is of the greatest importance and not the time or

day of posting. The most interesting topics of the posts were presentations of programs, especially master's programs, registration for events, and the news.

What is more, we have tried various methods while conducting cluster analysis and tried to segment users by their preferences. Six subscriber segments were identified based on parameters such as gender, age and city. For each segment, areas of interest are defined, which will allow the set up of targeted advertising, introduce new headings in GSOM groups and focus on target segments according to their interests.

## Conclusion

During our research, many experiments have been conducted using different models and sets of parameters. The third chapter describes all the models and their results in detail. This section will focus on the conclusions drawn from the models' results.

First of all, a substantial number of followers are not active and irrelevant to the analysis, they are not the target audience. For example, the pages that provide services for writing papers or fulfilling different tests for students cannot be considered a target audience of GSOM. It is important to remember that fact when measuring engagement and maybe use an alternative metric instead of standard ER that takes into account the total number of followers. For example, the number of likes divided by the number of views can be a good substitute. This finding highlights the need to promote higher education programs and not let things take their course, as we indicated in our work before. That is why the launch of target advertisement in VK that the marketing department already plans is necessary. Our findings made during this research can help set up the target advertisement campaign more efficiently. The marketing department can use the information derived from cluster analysis to tailor targeting ads content to the interests of the specific cluster of users. In particular, we would advise focusing on extracurricular activities provided by the university in correspondence with the interests of the targeted cluster. Also, we would recommend adding the university ratings in the ads as it is important for enrollees, according to the results of our research.

What is more, we would recommend the marketing department try to "bring the groups back to life" not only by acquiring but also by keeping the followers who are real people interested in enrolling in GSOM. For this purpose, there is a need to make content that is interesting for the target audience. We found out that while the presence of video content matters, it is not the primary factor when we look at the followers' reactions. It is the text of the post that is more important. So, we would recommend focusing on the content, in particular, to tailor it to the audience's preferences that can be tracked on the dashboard. For example, as the dashboard and cluster analysis show, the most widespread interests among our followers are travel and sport. Hence, the marketing department should write about student exchange, languages and sports opportunities that GSOM provides to attract the attention of the target audience. Moreover, during our research, we discovered some "trigger" words that significantly influence engagement. Based on this information, we would recommend writing about GSOM events, clubs, and news, as well as business topics and partner companies. Also, the best posting time is in the morning and evening hours, at this time, the posts are more likely to be seen and reacted to. All the recommendations

described above should help the marketing departments achieve their goals and attract prospective enrollees to VK groups.

However, it is worthy of note that our research has some limitations. First of all, the type of data used is time series, hence, there is a risk of structural changes in the data. For instance, a significant shift in the number of likes (from tens to hundreds) could make our models less relevant. Secondly, the amount of data collected is not that big (three thousand posts from all three groups), which could involve a risk of model overfitting and limit our ability to derive more detailed conclusions from the data. That is why we recommend continuing research after applying all the recommendations we provided. Further steps could be gathering more data - at least from two complete cycles of admission campaigns (one before applying the recommendations and one - after) to test the new content efficiency and check the target advertisement results. Moreover, it would be interesting to look at the data of competitors' groups' activity to gain some new insights and find out how GSOM can stand out among other universities. Also, some A/B tests can be conducted when testing the new way of creating the group content. Another area for improvement is looking at the differences between engaged and not engaged followers. Furthermore, qualitative interviews could be conducted with several group participants - this would allow for generating hypotheses for the ML task.

## Reference list

1. Aggarwal, Charu C., and Cheng Zhai. "A Survey of Text Classification Algorithms." Springer EBooks, August 1, 2012: 163–222. [https://doi.org/10.1007/978-1-4614-3223-4\\_6](https://doi.org/10.1007/978-1-4614-3223-4_6).
2. Araslanov, Egor, Evgeniy Komotskiy, and Ebenezer Agbozo. "Assessing the Impact of Text Preprocessing in Sentiment Analysis of Short Social Network Messages in the Russian Language." 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), October 26, 2020. <https://doi.org/10.1109/icdabi51230.2020.9325654>.
3. Armstrong, J. Scott. "Evaluating Forecasting Methods." International Series in Operations Research & Management Science, 2001: 443–72. [https://doi.org/10.1007/978-0-306-47630-3\\_20](https://doi.org/10.1007/978-0-306-47630-3_20).
4. Batrinca, Bogdan, and Philip Treleaven. "Social Media Analytics: A Survey of Techniques, Tools and Platforms." *AI & Society* 30, no. 1, February 1, 2015: 89–116. <https://doi.org/10.1007/s00146-014-0549-4>.
5. Basu, Tanmay, and C. Siva Ram Murthy. "A Feature Selection Method for Improved Document Classification." *Lecture Notes in Computer Science*, December 15, 2012: 296–305. [https://doi.org/10.1007/978-3-642-35527-1\\_25](https://doi.org/10.1007/978-3-642-35527-1_25).
6. Bawden, David, and Lyn Robinson. "The Dark Side of Information: Overload, Anxiety and Other Paradoxes and Pathologies." *Journal of Information Science* 35, no. 2, April 1, 2009: 180–91. <https://doi.org/10.1177/0165551508095781>.
7. Bawden, David, and Robinson, Lyn. "Information Overload: An Overview." In *Oxford Encyclopedia of Political Decision Making*. Oxford: Oxford University Press, 2020. doi: 10.1093/acrefore/9780190228637.013.1360.
8. Belfin, R. V., E. Grace Mary Kanaga, and Suman Kundu. "Application of Machine Learning in the Social Network." *Recent Advances in Hybrid Metaheuristics for Data Clustering*, June 5, 2020: 61–83. <https://doi.org/10.1002/9781119551621.ch4>.
9. Bhardwaj, Aditya. "Sentiment Analysis and Text Classification for Social Media Contents Using Machine Learning Techniques." *Social Science Research Network* (November 23, 2020). <https://doi.org/10.2139/ssrn.3735851>.



10. Blasco, Raquel Lozano, Marta Mira Aladrén, and Mercedes Gil Lamata. "Social Media Influence on Young People and Children: Analysis on Instagram, Twitter and YouTube." *Comunicar* 31, no. 74 (January 1, 2023): 125–137. <https://doi.org/10.3916/c74-2023-10>.
11. Bratchell, N. "Cluster analysis." *Chemometrics and Intelligent Laboratory Systems*, January 1989: 105-125, [https://doi.org/10.1016/0169-7439\(87\)80054-0](https://doi.org/10.1016/0169-7439(87)80054-0)
12. Buckley, Stephen M., Markus Ettl, Prateek Jain, Ronny Luss, Marek Petrik, Rajesh Kumar Ravi, and Chitra Venkatramani. "Social Media and Customer Behavior Analytics for Personalized Customer Engagements." *IBM Journal of Research and Development* 58, no. 5, November 21, 2014: 7:1-7:12. <https://doi.org/10.1147/jrd.2014.2344515>.
13. Burkov, Andriy. *The Hundred-Page Machine Learning Book*, 2019.
14. Chapleo, Chris, and Helen O'Sullivan. "Contemporary Thought in Higher Education Marketing." *Journal of Marketing for Higher Education* 27, no. 2, December 7, 2017: 159–61. <https://doi.org/10.1080/08841241.2017.1406255>.
15. Calus, Vlad. "The Biggest Social Media Challenges According to 80+ SMMs | Planable." *Planable*, February 3, 2023.
16. Chaudhry, Rayed. "Maximising Enrollment with Targeted Social Media Strategies for Higher Education." *ThinkOrion*, February 17, 2023. <https://www.thinkorion.com/blog/social-media-marketing-for-higher-education>.
17. Constantinides, Efthymios, and Marc C. Zinck Stagno. "Potential of the Social Media as Instruments of Higher Education Marketing: A Segmentation Study." *Journal of Marketing for Higher Education* 21, no. 1 June 1, 2011: 7–24. <https://doi.org/10.1080/08841241.2011.573593>.
18. Cordero-Gutiérrez, Rebeca, and Eva Lahuerta-Otero. "Social Media Advertising Efficiency on Higher Education Programs." *Spanish Journal of Marketing - ESIC* 24, no. 2, May 14, 2020: 247–62. <https://doi.org/10.1108/sjme-09-2019-0075>.
19. Domala, Jayashree, Manmohan Dogra, and Anuradha Srinivasaraghavan. "Lyrics Inducer Using Bidirectional Long Short-Term Memory Networks." *Algorithms for Intelligent Systems*, January 1, 2021. [https://doi.org/10.1007/978-981-16-3246-4\\_2](https://doi.org/10.1007/978-981-16-3246-4_2).

20. Dutot, Vincent, and Elaine Mosconi. "Understanding Factors of Disengagement within a Virtual Community: An Exploratory Study." *Journal of Decision Systems* 25, no. 3, June 17, 2016: 227–43. <https://doi.org/10.1080/12460125.2016.1187547>.
21. Fawcett, Tom. "An Introduction to ROC Analysis." *Pattern Recognition Letters*, no. 8 (2006): 861–74. <https://doi.org/10.1016/j.patrec.2005.10.010>.
22. Gareth, James; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert. "An Introduction to Statistical Learning." *Springer Texts in Statistics*, 2021. <https://doi.org/10.1007/978-1-0716-1418-1>.
23. Géron, Aurélien. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. "O'Reilly Media, Inc.," 2019.
24. Grinev, Maxim and Grineva, Maria. "Information Overload in Social Media Streams and the Approaches to Solve It." Paper presented at the 21st International World Wide Web Conference, 2012.
25. Hai, Tao, Jincheng Zhou, Dayang N. A. Jawawi, Xin Xiao Zheng, Surjeet Dalal, Cresantus Biamba, Edeh Michael Onyema, and Noble C. Anumbe. "Machine Learning Prospects in Social Media and Cloud Data Mining and Analytics." *Research Square (Research Square)*, August 26, 2022. <https://doi.org/10.21203/rs.3.rs-1988715/v1>.
26. Hasan, Rakibul, Maisha Maliha, and M. Arifuzzaman. "Sentiment Analysis with NLP on Twitter Data." 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), July 11, 2019. <https://doi.org/10.1109/ic4me247184.2019.9036670>.
27. Harrison-Walker, L. J. "Customer prioritisation in higher education: targeting "right" students for long-term profitability." *Journal of Marketing for Higher Education* 20, no. 2, 2010: 191–208. doi:10.1080/08841241.2010.526355.
28. Hossain, Md. Tazmim, Md. Arafat Rahman Talukder, and Nusrat Jahan. "Social Networking Sites Data Analysis Using NLP and ML to Predict Depression." *International Conference on Computing, Communication and Networking Technologies*, July 6, 2021. <https://doi.org/10.1109/iccent51525.2021.9579916>.

29. Hyndman, Rob J., Koehler, Anne B. "Another Look at Measures of Forecast Accuracy." *International Journal of Forecasting* 22, no. 4 (2006): 679–88. <https://doi.org/10.1016/j.ijforecast.2006.03.001>.
30. Jain, Mini, Peya Mowar, Ruchika Goel, and Dinesh Kumar Vishwakarma. "Clickbait in Social Media: Detection and Analysis of the Bait." *Conference on Information Sciences and Systems* (March 24, 2021). <https://doi.org/10.1109/ciss50987.2021.9400293>.
31. Jain, Varsha, Emmanuel Mogaji, Himani Sharma, and Anantha S. Babbili. "A Multi-Stakeholder Perspective of Relationship Marketing in Higher Education Institutions." *Journal of Marketing for Higher Education*, 2022: 1–19. doi: 10.1080/08841241.2022.2034201.
32. Koltay, Tibor. "Information Overload in a Data-Intensive World." *Advanced Information and Knowledge Processing*, July 1, 2017: 197–217. [https://doi.org/10.1007/978-3-319-59090-5\\_10](https://doi.org/10.1007/978-3-319-59090-5_10).
33. Korenius, Tuomo, Jorma Laurikkala, Kalervo Järvelin, and Martti Juhola. "Stemming and Lemmatization in the Clustering of Finnish Text Documents." *Conference on Information and Knowledge Management*, November 13, 2004. <https://doi.org/10.1145/1031171.1031285>.
34. Kowsari, Kamran, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura E. Barnes, and Donald E. Brown. "Text Classification Algorithms: A Survey." *Information* 10, no. 4. April 17, 2019: 150. <https://doi.org/10.3390/info10040150>.
35. Kumawat Deepika, Jain Vinesh. "POS tagging approaches: a comparison." *International Journal of Computer Applications* (0975 –8887) Volume 118–No. 6, May 2015: 32-38.
36. Mansour A., Mohammad J., Kravchenko Y. "Text Vectorization Method Based on Concept Mining Using Clustering Techniques." *VI International Conference on Information Technologies in Engineering Education (Inforino)*, 2022: 1-10. 10.1109/Inforino53888.2022.9782908.
37. Lobo, Jorge M., Alberto Jiménez-Valverde, Raimundo Real. "AUC: A Misleading Measure of the Performance of Predictive Distribution Models." *Global Ecology and Biogeography*, 17, no. 2, 2008: 145–51. <https://doi.org/10.1111/j.1466-8238.2007.00358.x>.

38. Maresova, Petra, Jan Hruška, and Kamil Kuca. "Social Media University Branding." *Education Sciences* 10, no. 3, March 1, 2020: 74. <https://doi.org/10.3390/educsci10030074>.
39. Maringe, Felix and Gibbs, Paul. *Marketing Higher Education: Theory and Practice*. Maidenhead, England; New York: McGraw Hill/Open University Press, 2008.
40. Marino, Vittoria. "Social Media Mix in the University Communication Plan: A Bridge Towards Public Engagement: Structured Abstract." *Developments in Marketing Science: Proceedings of the Academy of Marketing Science*, January 1, 2016: 275–81. [https://doi.org/10.1007/978-3-319-29877-1\\_57](https://doi.org/10.1007/978-3-319-29877-1_57).
41. Mejova Yelena. "Sentiment analysis: an overview". Computer Science Department, University of Iowa, 2009: 1–34.
42. Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*, 2013: 3111-3119.
43. Minaee, Shervin, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. "Deep Learning--Based Text Classification." *ACM Computing Surveys* 54, no. 3. April 17, 2021: 1–40. <https://doi.org/10.1145/3439726>.
44. Mooney R-J, Nahm U-Y, Mooney R-J. "Text mining with information extraction. In: Daelemans W, du Plessis T, Snyman C, Teck L (eds) *Multilingualism and electronic language management: proceedings of the 4th international MIDP colloquium*". Bloemfontein, Van Schaik, South Africa, 2003: 141–160.
45. Mushketova, Natalia S., Elizaveta Bydanova, and Gilles Rouet. "National Strategy for Promotion of Russian Universities in the World Market of Education Services." *International Journal of Educational Management*, January 5, 2018. <https://doi.org/10.1108/ijem-10-2016-0207>.
46. Murphy, Kevin P. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.
47. Naser, M., Alavi, A. "Error Metrics and Performance Fitness Indicators for Artificial Intelligence and Machine Learning in Engineering and Sciences." *Architecture, Structure and Construction*, November 2021. <https://doi.org/10.1007/s44150-021-00015-8>
48. Plevris, Vagelis, Solorzano, German, Bakas, Nikolaos P., Seghier, Ben, El Amine, Mohamed. "Investigation of performance metrics in regression analysis and machine

- learning-based prediction models.” 8th European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS 2022), June 2022. <https://doi.org/10.23967/eccomas.2022.155>.
49. Rani D., Kumar R., Chauhan N. "Study and Comparison of Vectorization Techniques Used in Text Classification." 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2022: 1-6. [10.1109/ICCCNT54827.2022.9984608](https://doi.org/10.1109/ICCCNT54827.2022.9984608).
  50. Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang. "SQuAD: 100,000+ Questions for Machine Comprehension of Text." ArXiv (Cornell University) (June 16, 2016). <https://doi.org/10.18653/v1/d16-1264>.
  51. Sinha, Adwitiya, and Fatih Kayaalp. "Sentiment Analysis with Machine Learning Methods on Social Media." *Advances in Distributed Computing and Artificial Intelligence Journal* 9, no. 3, September 17, 2020: 5–15. <https://doi.org/10.14201/adcaij202093515>.
  52. Singh Jasmeet, Gupta Vishal. "Text Stemming." *ACM Computing Surveys* 49, no. 3, September 16, 2016: 1–46. <https://doi.org/10.1145/2975608>.
  53. Solangi, Yasir Ahmed, Zulfiqar Ali Solangi, Samreen Aarain, Amna Abro, Ghulam Ali Mallah, and Asadullah Shah. "Review on Natural Language Processing (NLP) and Its Toolkits for Opinion Mining and Sentiment Analysis." 2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences (ICETAS), November 1, 2018. <https://doi.org/10.1109/icetas.2018.8629198>.
  54. Tharwat, Alaa. "Classification Assessment Methods." *Applied Computing and Informatics* 17, no. 1, January 4, 2021: 168–92. <https://doi.org/10.1016/j.aci.2018.08.003>.
  55. Van Dam, Jan-Willem, Van de Velden, Michel. "Online profiling and clustering of Facebook users." *Decision Support Systems*, 2015: 60-72. <https://doi.org/10.1016/j.dss.2014.12.001>
  56. Van Knippenberg, Daan, Dahlander, Linus, Haas, Martine R., and George, Gerard. "Information, Attention, and Decision Making." *Academy of Management Journal* 58, no. 3, 2015: 649-657. doi:10.5465/amj.2015.4003.
  57. Vukatana, Kreshnik, Gjergji Mulla, Ran Liu, and Xhulio Mitre. "Analysing User Opinions on Content and Social Media Apps to Online Marketing: Evidence from Albania." *Journal*

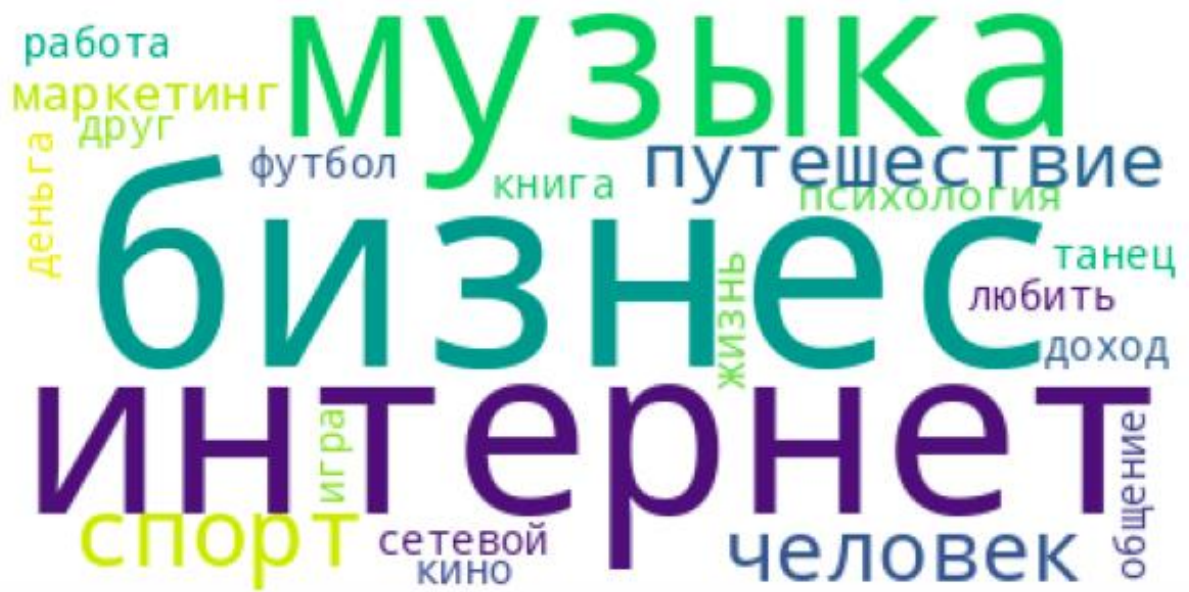
- of Eastern European and Central Asian Research 9, no. 6 December 3, 2022: 1072–82. <https://doi.org/10.15549/jeecar.v9i6.965>.
58. Yost, Elizabeth, Tingting Zhang, and Ruoxi Qi. "The Power of Engagement: Understanding Active Social Media Engagement and the Impact on Sales in the Hospitality Industry." *Journal of Hospitality and Tourism Management* 46 March 1, 2021: 83–95. <https://doi.org/10.1016/j.jhtm.2020.10.008>.
59. Zailskaite-Jakste, Ligita, and Rita Kuvykaite. "Implementation of Communication in Social Media by Promoting Studies at Higher Education Institutions." *The Engineering Economics* 23, no. 2, April 18, 2012. <https://doi.org/10.5755/j01.ee.23.2.1550>.
60. Zerres, Christopher. "Social Media Marketing." Springer EBooks, January 1, 2020: 1–18. [https://doi.org/10.1007/978-3-658-28973-7\\_32-1](https://doi.org/10.1007/978-3-658-28973-7_32-1)
61. Zheng, Wen-Bin, Yuntao Qian, and Minchao Ye. "A Grouped Structure-Based Regularized Regression Model for Text Categorization." *Journal of Software* 7, no. 9 (January 9, 2012). <https://doi.org/10.4304/jsw.7.9.2119-2124>.

**Internet resources:**

62. Chaudhry, Rayed. "Maximising Enrollment with Targeted Social Media Strategies for Higher Education." ThinkOrion, February 17, 2023. URL: <https://www.thinkorion.com/blog/social-media-marketing-for-higher-education>. (Access date: 10.02.2023)
63. Calus, Vlad. "The Biggest Social Media Challenges According to 80+ SMMs | Planable." Planable, February 3, 2023. URL: <https://planable.io/blog/social-media-challenges/#:~:text=The%20biggest%20social%20media%20challenge,their%20strategies%20to%20keep%20up>. (Access date: 10.02.2023)
64. Feehan, Blair. "2023 Social Media Industry Benchmark Report | Rival IQ." URL: <https://www.rivaliq.com/blog/social-media-industry-benchmark-report/#title-higher-ed> (Access date: 10.02.2023)
65. Massimiliano Viola, Luca Brunelli, and Gian Antonio Susto. "Instagram Images and Videos Popularity Prediction: a Deep Learning-Based Approach". 2021. URL: <https://ceur-ws.org/Vol-3102/paper2.pdf> (Access date: 20.03.2023)

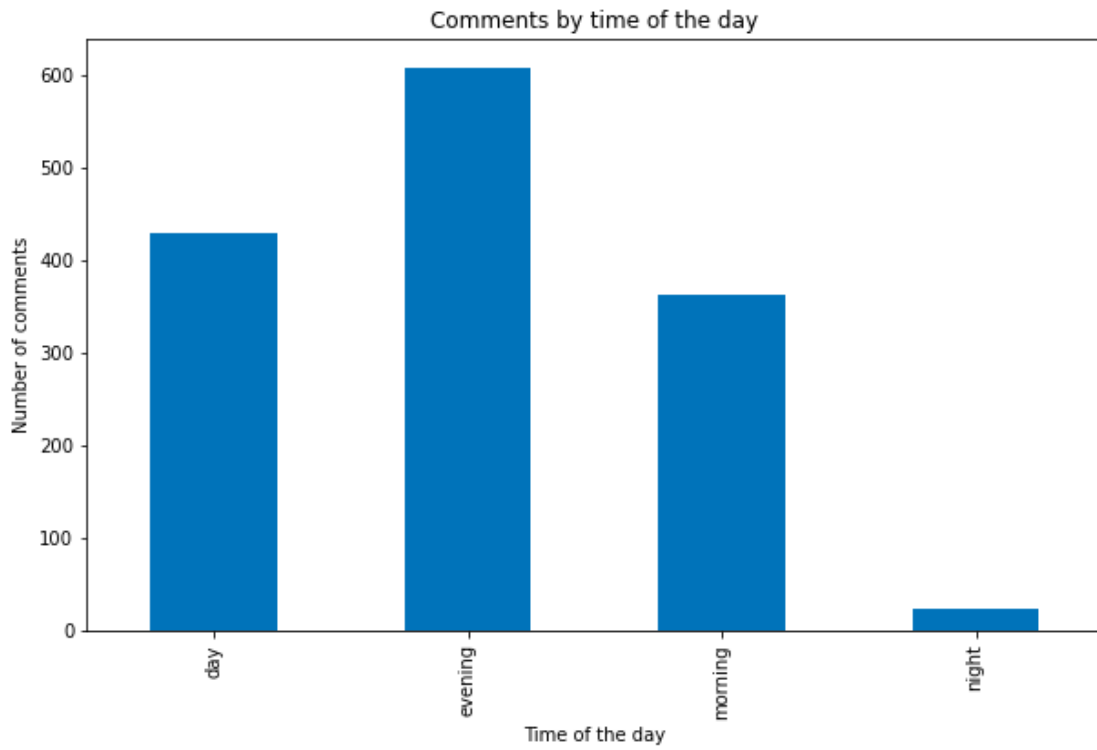
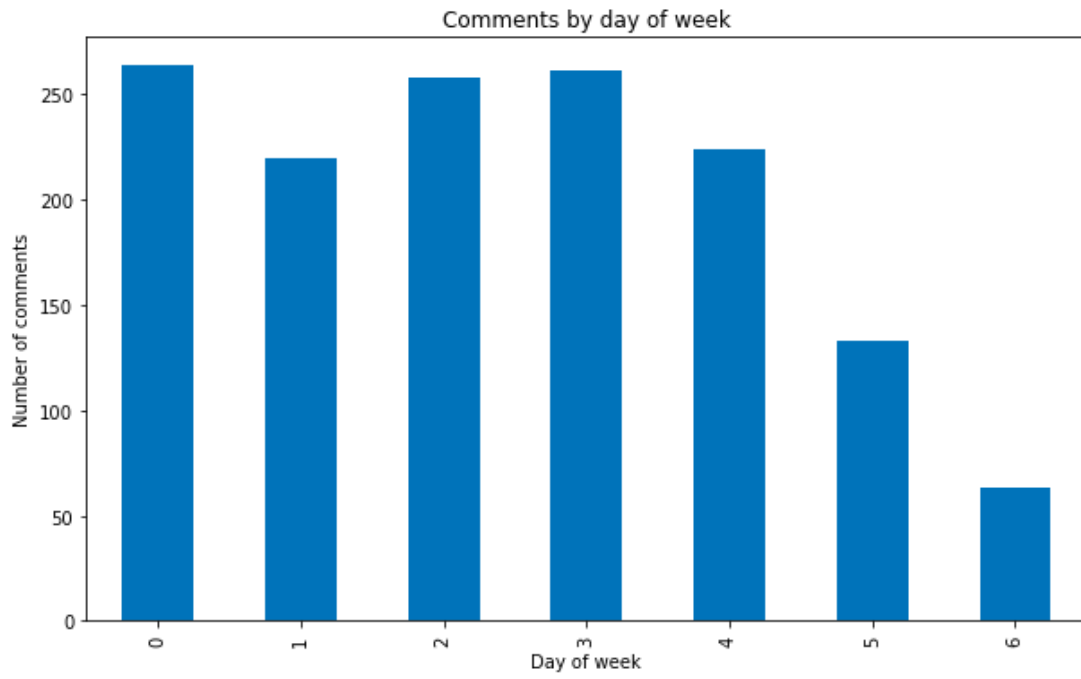
66. "How Universities Use Social Media for Marketing | Full Fabric," n.d. URL: <https://www.fullfabric.com/articles/how-universities-use-social-media-for-marketing>  
(Access date: 20.03.2023)
67. Official website of Yandex.DataLens. URL: <https://cloud.yandex.ru/services/datalens>
68. Official website of Yandex Query. URL: <https://cloud.yandex.ru/services/query>
69. Rogers, Jenny. "How Universities Use Social Media for Marketing | Full Fabric," n.d. URL: <https://www.fullfabric.com/articles/how-universities-use-social-media-for-marketing>. (Access date: 10.02.2023)
70. Sinha, Disha "Difference between K-Means and DBScan Clustering| Geeks for Geeks" URL: <https://www.geeksforgeeks.org/difference-between-k-means-and-DBScan-clustering/> (Access date: 10.02.2023)
71. VK Press Office. "Vkontakte podvela itogy pervogo kvartala 2022 goda". April, 2022. URL: <https://vk.com/main.php?subdir=press&subsubdir=q1-2022-results>. (Access date: 03.03.2023)

A word cloud of general GSOM group subscribers' interests

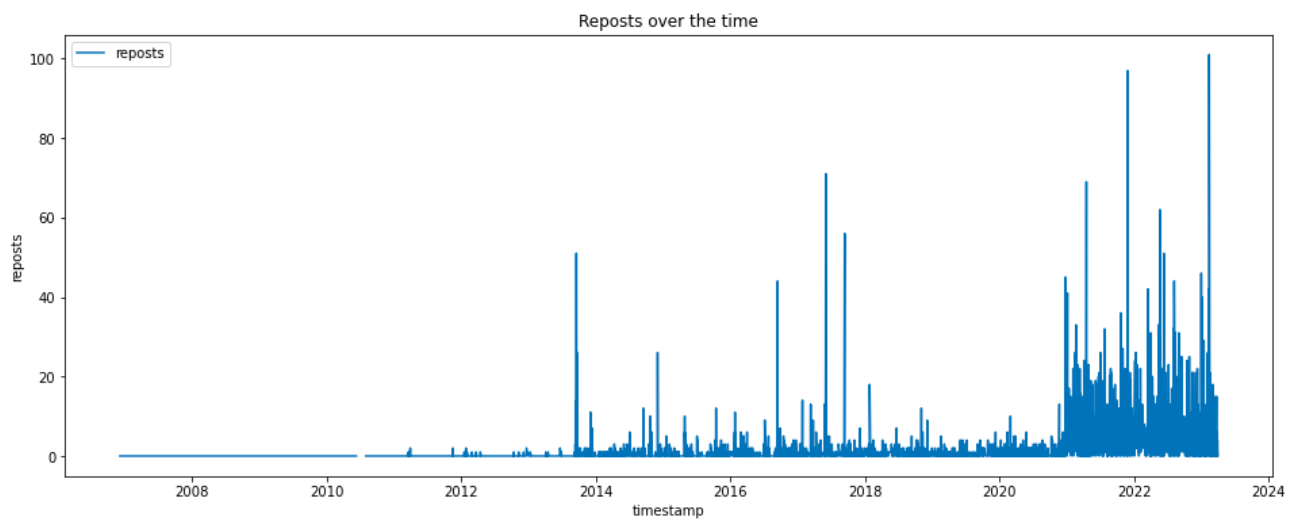
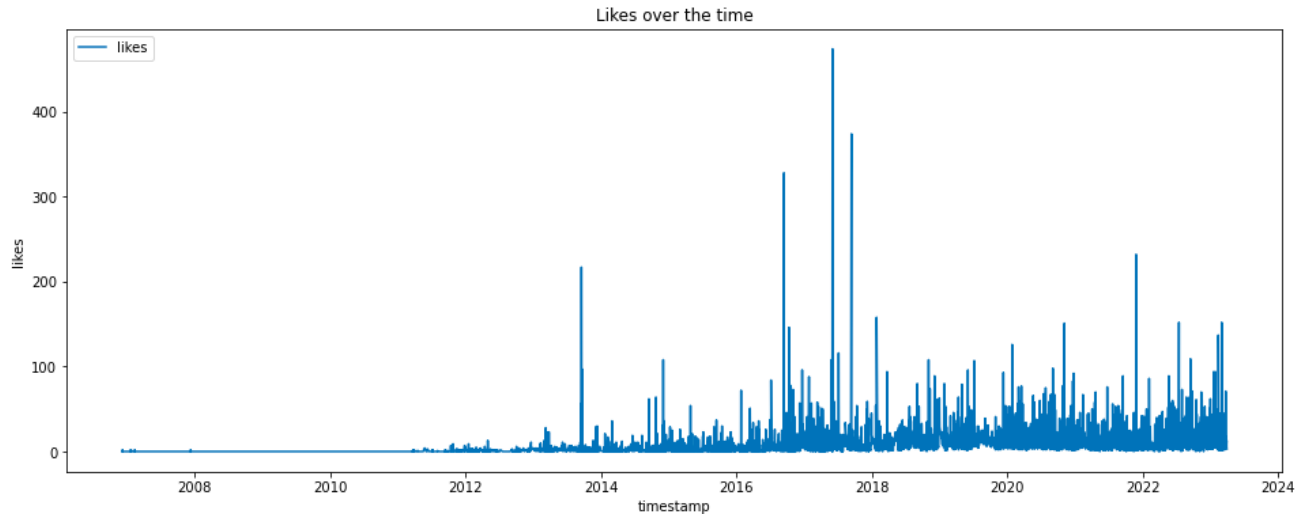




The number of comments on posts depending on the day of the week and time of the day



Number of likes and reposts in the GSOM groups over time: from 2012 till 2023

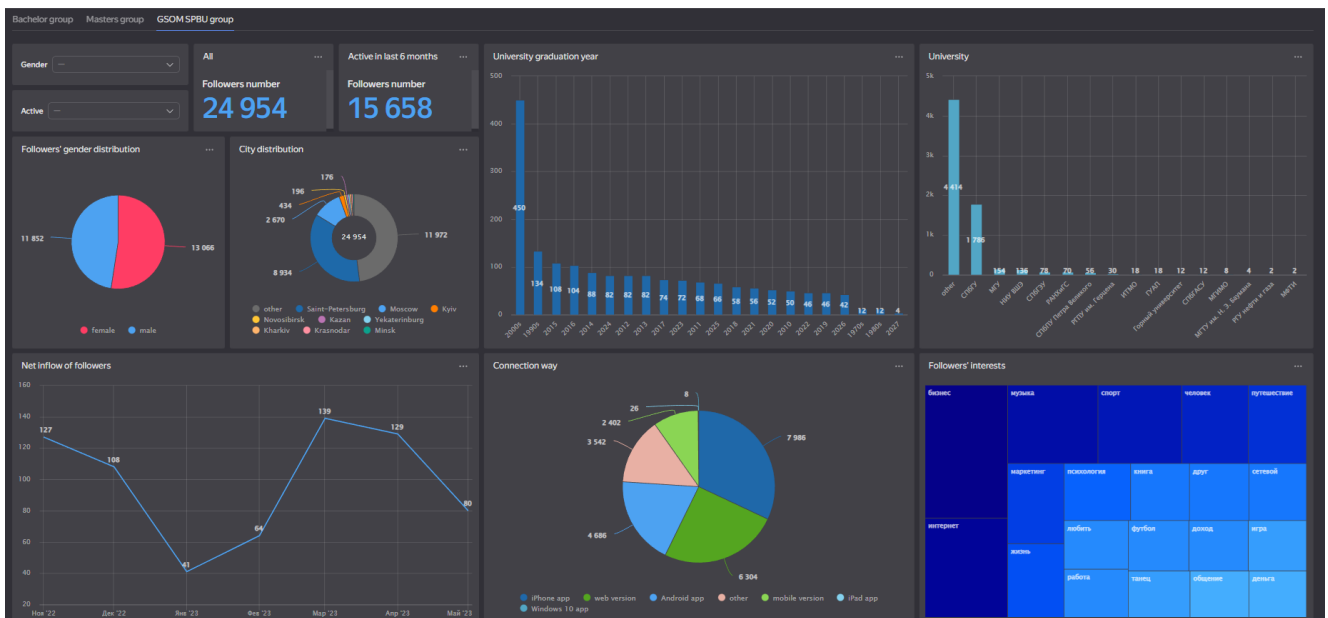


## Top-5 posts by the number of likes and comments

	items_date	items_text	items_comments_count	items_likes_count	items_reposts_count
0	2017-06-01 08:15:24	#GSOMachievements #GSOMsuccess\n\nПервый день ...	0	474	71
1	2017-09-11 10:51:38	#GSOMachievements\n\n ⚡ ВШМ СПбГУ в топ-25 лучши...	0	374	56
2	2016-09-12 09:07:38	ЕЩЕ ВЫШЕ!\n\nМагистерская программа ВШМ СПбГУ ...	0	328	44
3	2021-11-24 13:25:09	ВШМ СПбГУ вошла в 1% лучших бизнес-школ мира! ...	0	232	97
4	2013-09-16 14:39:08	ВШМ СПбГУ стала первой российской бизнес-школо...	0	217	51

	items_date	items_text	items_comments_count	items_likes_count	items_reposts_count
0	2016-06-24 11:41:35	#Абитуриенту #Важно\n\nДорогие будущие студент...	35	5	0
1	2021-08-25 07:11:00	Бить первокурсником волнительно: новая среда, ...	32	49	22
2	2019-06-26 13:01:00	5 июля в "Михайловской даче" пройдёт церемония...	24	17	0
3	2018-07-23 12:11:40	BREAKING NEWS 🔥\n\nРаспределение на потоки для...	22	10	0
4	2017-05-15 14:01:04	#GSOMadmission\n\n23 мая 2017 состоится заключ...	17	2	0

Dashboards for the master's enrollees group and the general GSOM group



Comparison of metrics on the train dataset of the RandomForest and LGBM models

<b>Evaluation metric</b>	<b>RandomForest</b>	<b>LGBM</b>
ROC-AUC score	0.94	0.99
Accuracy	0.77	0.96
Precision	0.75	0.97
Recall	1.00	0.98
F1-score	0.86	0.97

## Function for determining the optimal number of clusters

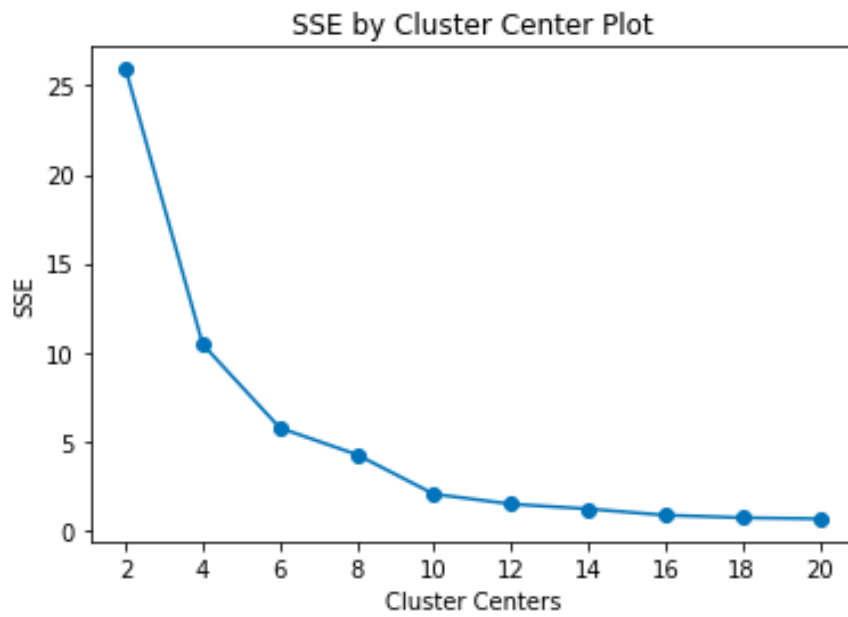
```
from sklearn.cluster import MiniBatchKMeans
def find_optimal_clusters(df, max_k):
    iters = range(2, max_k+1, 2)

    sse = []
    for k in iters:
        sse.append(MiniBatchKMeans(n_clusters=k, init_size=1024, batch_size=2048, random_state=20).fit(df).inertia_)
        print('Fit {} clusters'.format(k))

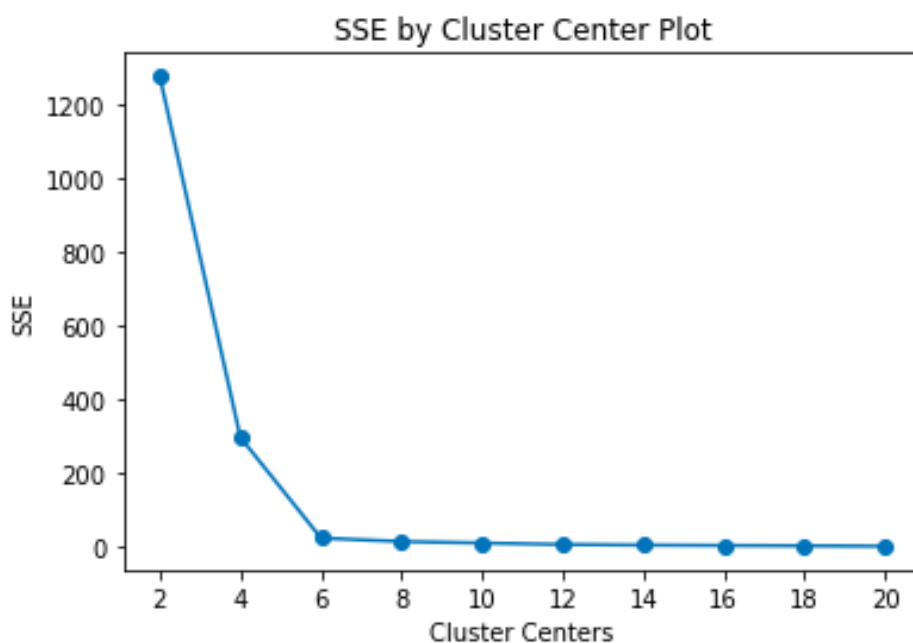
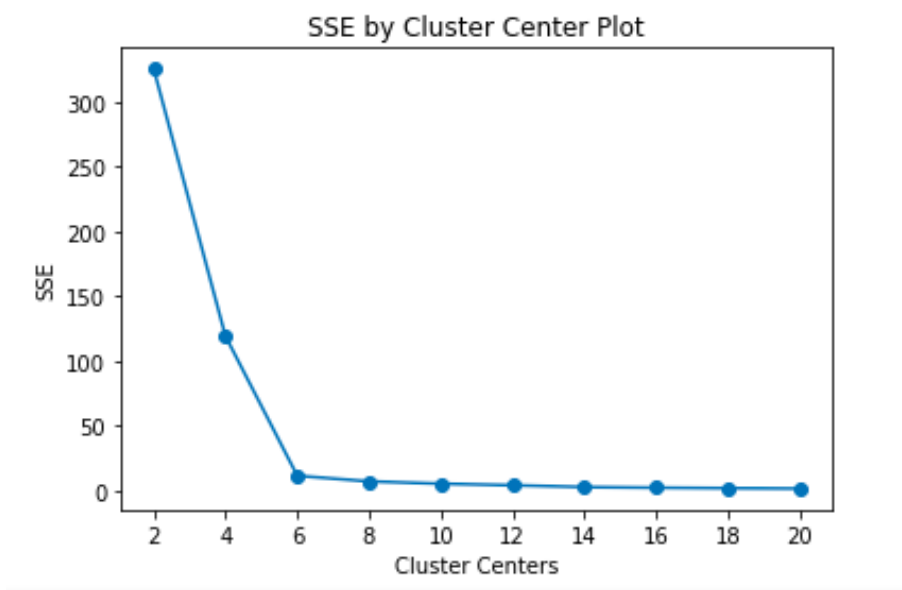
    f, ax = plt.subplots(1, 1)
    ax.plot(iters, sse, marker='o')
    ax.set_xlabel('Cluster Centers')
    ax.set_xticks(iters)
    ax.set_xticklabels(iters)
    ax.set_ylabel('SSE')
    ax.set_title('SSE by Cluster Center Plot')

find_optimal_clusters(df_profile, 20)
```

Elbow-method for defining optimal number of clusters for gsom\_spbu group (general group of GSOM), k-means with 2 parameters: sex, years old



Elbow-method for defining optimal number of clusters for gsom\_ma group (graduate/masters enrollees) and gsom\_spbu group (general group of GSOM), k-means with 3 parameters: sex, years old, city of living.





Interests of graduate enrollees group, clustering method: k-means, parameters: sex, years old

Cluster 1	Cluster 2	Cluster 3	Cluster 4
men, 35 - 63 years old	men, 16 - 34 years old	women, 36-59 years old	women, 16-35 years old
swimming billiards formula soccer bicycle marathon hobbies life hitchhiking people travelling photography	works soccer swimming hockey basketball hiking numismatics foreign languages travels geography history programming education	languages sports texts public relations universities educational projects internationalization higher education books travel sailing equestrian	south korea real friends books movies people smiles words cool sneakers headphones skateboard black and white mug pictures coffee decoupage drawing origami book reading

Interests of general GSOM groups, clustering method: k-means, parameters: sex, years old

Cluster 1	Cluster 2	Cluster 3	Cluster 4
men, 36 - 98 years old	women, 35 - 83 years old	men, 15-35 years old	women, 15-34 years old
literature business movies music sports art new psychology development active recreation high internet beautiful travel	travels people dancing books music children psychology hikes nature sports yoga walks marketing movies	development music sports social design advertising philosophy travel sailing internet urban chemistry	dancing works travel art culture music people psychology self-development innovation delicious food love

## Most frequent words from profiles' posts of general GSOM group subscribers

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
men Other cities 15 - 98 years	men SPb 15-97 years	women Moscow 18-58 years	women SPb 15 -83 years	men Moscow 17-64 years	women Other cities 15 - 78 years
The Gift play Russia thousand love	Apartment Petersburg Russia love the case	The book Love lucky learn world	Love happiness group voice congratulations	Moscow Russia love city the world	Love Size price Question city

Source code for data analysis in the study

<https://github.com/DariaDoroshkova/MasterThesis.git>