International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

Examination of 'Interests' and 'Activities' of Social Network users

Andrei V. Plotnikov

Abstract*: The present study relates to the analysis of attribute data related to users of the social network VK. The general population N = 52,614 users is the intersection of audiences from two communities for social media marketing. Based on the collected statistics on the "interests" attribute, one can compile a generalized portrait of an IT specialist and online marketer: this is a man aged about 30 years old, not married, or who defines his family status as "everything is complicated". He speaks an average of two languages, works for an organization, or studies at a university. He has about 370 followers on VK. The result based on the data from the field 'activities' is very close to the data from the field 'interests', and gives a similar picture of the generalized portrait of a specialist. As part of the study, the authors have learned how to segment users into the users that identify themselves as 'IT specialists or online marketers', and 'other' users, using machine learning methods.

Keywords*: Digital Economy, Social Media Marketing, Social Networks, User Behavior, Online Behavior.

I. INTRODUCTION

The concept of the digital economy is spreading all over the world, including Russia. [1] In 2017, the Government of the Russian Federation developed and approved a project for creating conditions for Russia's transition to the digital economy. Five areas were chosen within the program: personnel and education, information infrastructure, information security, formation of research competencies and technological reserves, and statutory regulation. [2]

Researchers identify the following areas in the digital economy [3]: studies based on large data arrays (Big Data) [4, 5] using machine learning algorithms [6]; Internet of things [7, 8], the use of cloud technologies [9], e-commerce [10], social networks [11], and bilateral markets [12], independence of geographical location [13, 14], management features, and innovative activity incentives.

Thus, the digital economy has four specific features: the irrelevance of the geographical location or the absence of the need to do business/work in one geographical location, the key role of platforms, the importance of network effects, and the use of big data. These features distinguish it from previous concepts of economics. This study relates to the digital economy since the object of the study is the VK social network, which complies with the principles of the digital economy.

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
Andrei V. Plotnikov*, Perm National Research Polytechnic University, Perm, Russia.

II. ONLINE COMMUNICATIONS

Consumer targeting based on demographic profiles is the preferred approach for finding a narrow audience for subsequent settings of advertising. Targeted advertising based on data from social networks is a marketing field, featuring increased activity in recent years. Advertising plays an important role in business because it can attract new customers without directly interacting with them. The advertising industry in social networks is developing rapidly, inventing new ideas for its clients in order to attract new consumers and, at the same time, support the existing ones. With the advent of the Internet, the continuous development of technology, a growing database of documents, and the number of users make advertising more significant.

Advertising on social networks has shifted the focus from general or traditional advertising without segmenting to targeted advertising using various methods of user attribute targeting (geography, gender, bad habits, etc.).

Marketing communications perform several functions [15]:

1. Help organizations add value to their products or services. Marketing communications can inform potential consumers about the availability of goods, their availability and after-sales service, functions, and advantages. Moreover, organizations use marketing communications not only to attract potential customers but also to improve relations with the existing customers to receive feedback.
2. Emphasize the attributes of service providers and goods;
3. Describe the benefits of the consumption of goods and services;
4. Use metaphors to convey value to the consumer;
5. Facilitate the involvement of consumers in the production and process of using goods.
6. Help consumers evaluate service offerings and distinguish offers from suppliers.
7. Determine the appropriate time and place of sales promotion.

It is important to apply a comprehensive approach to marketing communications and take into account trends: using the Internet to interact with customers on social networks and using Internet marketing as a set of methods that are used on the Internet to promote or transmit communications about a company's product and services to potential customers or the target audience. Online marketing communications have become important in the last decade, as information and communication technologies have spread to the general public. As a result of universal accessibility, online marketing communications are cheaper than traditional means.

Since the use of information and communication technologies greatly influences the way enterprises work, information is transmitted to business partners, internal communication is organized, and customers communicate with each other. In addition, given the possibility of segmenting consumers on the network, gender segmentation of users appears in the online environment, and the user interacts with various marketing incentives. In marketing research in general and in modeling consumer behavior, in particular, when studying control variables, socio-demographic characteristics of consumers are taken into account, where gender is a key variable in segmentation.

Social networks are an important component in building online communications with potential and existing customers. S. Roy [16] in his work notes that the analysis of social networks shows which groups are interconnected, and also reveals common attributes that can be used for segmentation, and then for building communication based on attributes. The measurement of the relationships and interactions between people, groups, organizations, and other related entities is very important in marketing communications and is the subject of interdisciplinary research. Social networks such as Twitter, Facebook, LinkedIn are very large with millions of vertices and billions of edges, and in order to collect meaningful information from these tightly connected graphs and a huge amount of data, it is important to find the right network topology, as well as to analyze various network parameters. N.S. Faber [17] in his article notes that sociologists are increasingly recognizing the potential of social network analysis, which explains behavior, depending on its social structure. Social network analysis is a valuable tool for exploring some of the central mechanisms that underlie intragroup and intergroup behavior. The paper emphasizes the general relevance of this scientific approach and describes the prerequisites, generation, and application of cross-sectional, as well as network statistics. At the same time, the authors strive to provide a general introduction for researchers new to this approach, demonstrating the potential and limitations of analyzing social networks for various fields. Shih-Chih Chen and Chieh-Peng Lin [18] argue that social media marketing campaigns influence user satisfaction through perceived value and social identity. Social identification by user attributes will be analyzed in the study.

A serious barrier to marketing communications in social networks is spam. People's attitudes to online privacy and measures taken to protect themselves from spam are transforming into closed social networks. This is manifested in the introversion of user accounts, and advertisers may not be able to find the target audience due to the lack of necessary account attributes. This study deals with analyzing attribute data of users of the VK social network, namely those who themselves are directly related to online marketing and social media marketing. Their attributes will be studied.

III. PROPOSED METHODOLOGY

A. General description

The general population is N = 52,614 users (the intersection of audiences from two communities for SMM: "Cerebro Trager" and "Target Hunter"). Intersecting user audiences are more representative than audiences that are members of only one of the two communities. Some users hide pages, and some whose pages are open do not fill in some attribute values. For example, a place of work, political preferences, or favorite music, etc. Thus, one cannot collect values from each user. Audience collecting dates were the following: Summer 2019.

B. Algorithm

As a research base, the Vkontakte social network (international name: VK), the Russian social network, has been chosen.

The connection data will be received using the VK API tool. This is an interface that allows receiving information from the vk.com database using http requests to a special server.

The Python libraries have been used in the study:

pymorphy2 for lemmatization of words. This is the process of putting the word form in a lemma – its normal (vocabulary) form.

NumPy is the homogeneous multidimensional array.

WmdSimilarity is the determination of the similarity of documents.

The following words have been chosen as keywords that will set the topic: 'marketing', 'Internet', 'SMM', 'SEO', 'contextual', 'advertising', 'Yandex', 'advertising', 'Google', 'promotion' , 'targeting' , 'target', 'cerebro', 'targeted', 'tergetologist', 'marketer', 'optimization' (in Russian).

All the necessary libraries are imported. Next, a service procedure that prepares a list of words for the lemmatizer is performed. The procedure that selects nouns and Latin alphabet from the given list and leads them to normal form (lemmatization) is defined.

The TF-IDF method is used, and then without obtaining the desired result, the presence of this word in the dictionary of the pre-trained model RUS_VEC (fastText algorithm from Facebook, trained on Atraneum) is verified.

The procedure for calculating the accuracy of classification into several classes is the following.

The data file is read and the records of users who logged in no later than 60 days before today are cut off since the users who visit rarely the social network may be inactive in their intentions, as well as unrepresentative for research.

After filtering, the number of users is N = 52,614. These users have the following attribute values filled in: 'about' – 10.8 %, 'activities' – 15.6 %, 'bdate' – 75.7 %, 'books' – 9.4 %, 'career_group_id' – 14.9 %, 'career_position' – 9.9 %, 'city' – 82.6 %, 'domain' – 100.0 %, 'education' – 0.0 %, 'exports' – 0.0 %, 'faculty' – 45.1 %, 'followers_count' – 87.5 %, 'friend_status' – 100.0 %, 'games' – 5.6 %, 'home_town' – 29.8 %, 'id' – 100.0 %, 'interests' – 15.5 %, 'military' – 0.0 %, 'occupation' – 68.0 %, 'personal_alcohol' – 29.7 %, 'personal_langs' – 34.6 %, 'personal_life_main' – 29.6 %, 'personal_people_main' – 29.7 %, 'personal_political' – 10.1 %, 'personal_religions' – 9.6 %, 'personal_smoking' – 29.8 %, 'relation' – 45.1 %, 'sex' – 100.0 %, 'site' – 25.7 %, 'university' – 45.1 %.

The following fields of interest are then selected: 'bdate', 'university', 'faculty', 'followers_count', 'personal_langs', 'city', 'occupation', 'relation', 'sex', 'interests', and 'activities'. As a key field, 'interests' (a) & 'activities' (b) are used.

The remaining fields are processed: replace NaN, categorize.

The pre-trained model (the fastText algorithm from Facebook, trained on the Atraneum, has been used) is imported.

The pattern that describes the proximity to the field of interest 'interests' is formed, then the same pattern for the field 'activities' is formed.

The 'distance' column, the distance from the pattern to the interests of each user, is formed. The smaller the number is, the closer the user's interests to Internet marketing are.

IV. RESULT ANALYSIS

The data analysis has revealed the following: the number of "Internet marketers" or IT specialists, as expected, is relatively small. The majority of users have interests that are far from Internet marketing and SEO.

First, the method based on the vectorization of words from 'interests' using the TF-IDF method is used, and then the resulting vectors are clustered using the k-means method. Thus, it turns out that if a person wrote that he/she has interests from online marketing to esotericism, then to which cluster should he/she be attributed to? Thus, the data on 'interests' are absolutely not unified, i.e., everyone writes what they want. To predict any specific interest is impossible due to the infinite number of their classes. Perhaps, in this case, the method is not optimal. The authors tried to group all interests into several clusters according to the maximum similarity of words. The result was clearly mixed.

For example,

187 Internet advertising, marketing, self-development '1'

193 Marketing, photo, fashion, psychology '2'

667 Marketing, self-development, travel '1'

728 Internet commerce. Information Technology. Business '0'

344 Digital marketing, advertising, self-development '1'

550 Music '0'

415 Family, SMM, business '0'

551 Music, literature, vocals, education, science, digital '3'

In other words, here the sphere of Internet marketing/marketing, advertising, and business falls immediately under three clusters: '0', '1', '2'.

After an unsatisfactory result, the pre-trained fastText model from Facebook, trained on the Atraneum, is imported. The pattern that describes the proximity to the field of interest 'interests' is formed, and then the same pattern for the field 'activities' is formed. The marking carried out according to the WMDistance parameter, the distance between the generalized vector of the given keywords and the vector of interests from VK, is added.

The result was much better than that obtained by the TF-IDF method. A distance of 0.0 corresponds to the distance to an identical vector, a distance of 10.0 – to an infinitely distant one.

One should determine which users would be assigned to the group of "Internet marketers and IT specialists." The distance value 1.00 will be considered borderline (Figure 1). A social portrait of a user with interests close to Internet marketing and with a type of activity close to Internet marketing will be drawn up. One should determine how many of these users are in the research sample.

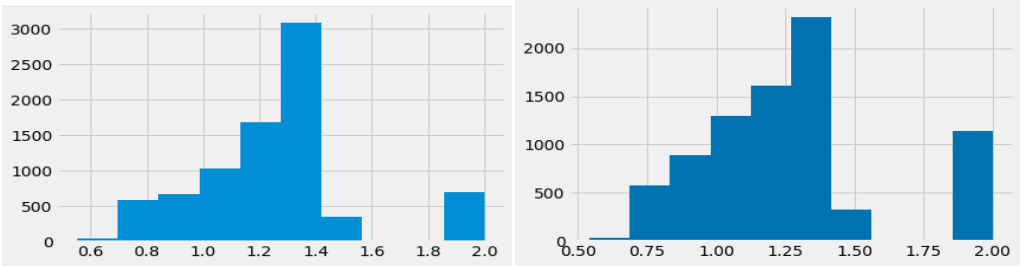

Fig. 1 a, b: Distribution of the users by the distance from the Internet marketing field and search engine optimization. By indicated interests – on the left (a); by activity – on the right (b).

According to 'interests', there are 1,414 people, or 17 % of the total number of users in the sample (Figure 1a), and according to 'activities', there are 1,700 people, or almost 21 % of the total number of users in the second sample (Figure 1b). Now, one should make a comparison to find out men or women are in the majority among the specialists in the field of Internet marketing and search engine optimization (Figure 2 a, b).

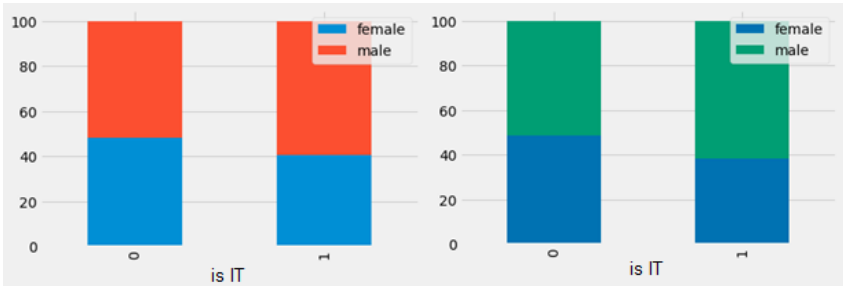

Fig. 2 a, b: Gender differences according to the indicated 'interests' – on the left; the indicated activities – on the right

It can be seen that among the Internet marketers and SEO specialists in displayed 'interests' (Figure 2a) men predominate (575 F and 839 M), while the gender

composition of the unexpressed group is approximately the same (3,257 F and 3,491 M). In Figure 2b, men also prevail by type of activity (651 F and 1,053 M), and the gender composition of the unexpressed group is approximately the same (3,162 F and 3,334 M).

The average age in groups should be defined.

The average age for the 'interests' sample is less by three years (born in 1988). A similar result was obtained in the sample by type of activity. It should be borne in mind that not all participants in either group indicated the age.

The number of followers for the sample by 'interests' (Figure 3a) and the sample by type of activity (Figure 3b) should be determined.

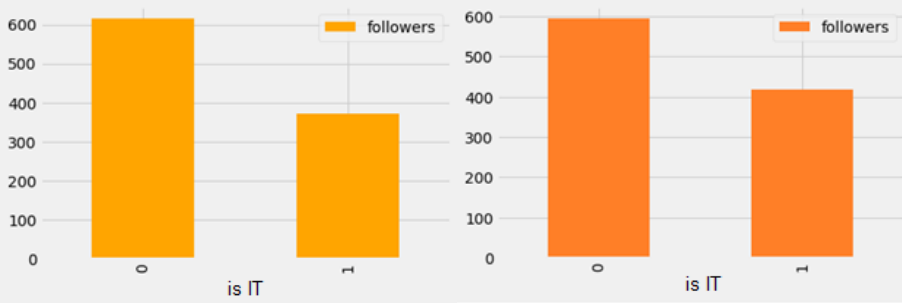

Fig. 3 a, b: Distribution of the number of followers of the sample 'interests' (a) and sample by type of activity (b)

The number of followers for two samples (371 (a); 418 (b)) is significantly lower than that for users with other interests (not IT, not online marketing = 615 (a); 595 (b))

The kind of the composition of groups in the context of formation (Figure 4a, b) should be determined.

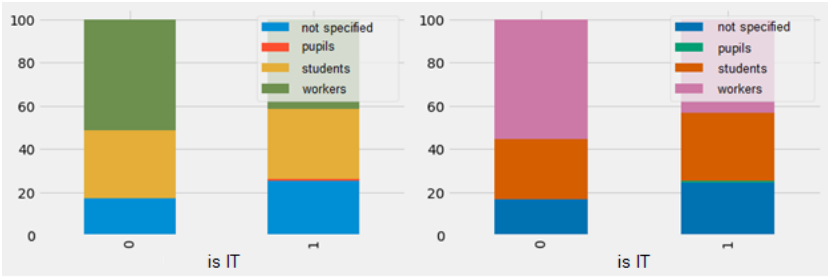

Fig. 4 a, b: Distribution of samples by education (a – 'interests'; b – sample by type of activity)

(a, 0): not specified = 1,159; pupils = 14; students = 2,101; workers = 3,474.

(a, 1): not specified = 359; pupils = 10; students = 460; workers = 585.

(b, 0): not specified = 1083; pupils = 7; students = 1,812; workers = 3,594.

(b, 1): not specified = 419; pupils = 10; students = 538; workers = 737.

Next, the language skills of the declared groups (Figure 5a, b) should be considered.

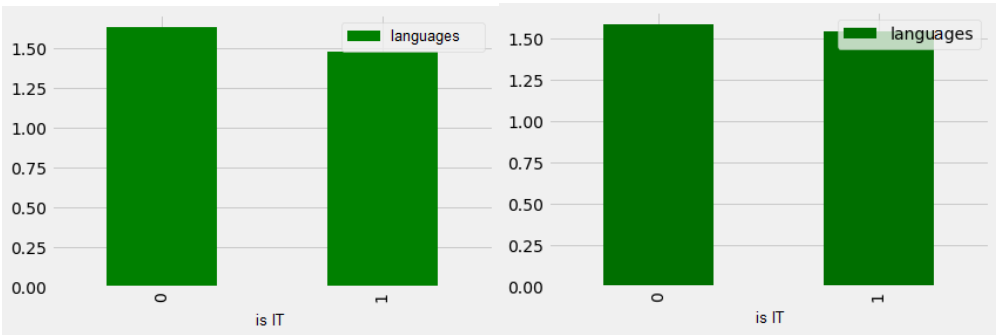

Fig. 5 a, b: Number of declared languages by groups

According to the results of analyzing the samples for indicating the language of speech, the following average values have been revealed: (a, 0) = 1.63; (a, 1) = 1.47; (b, 0) = 1.59; (b, 1) = 1.54.

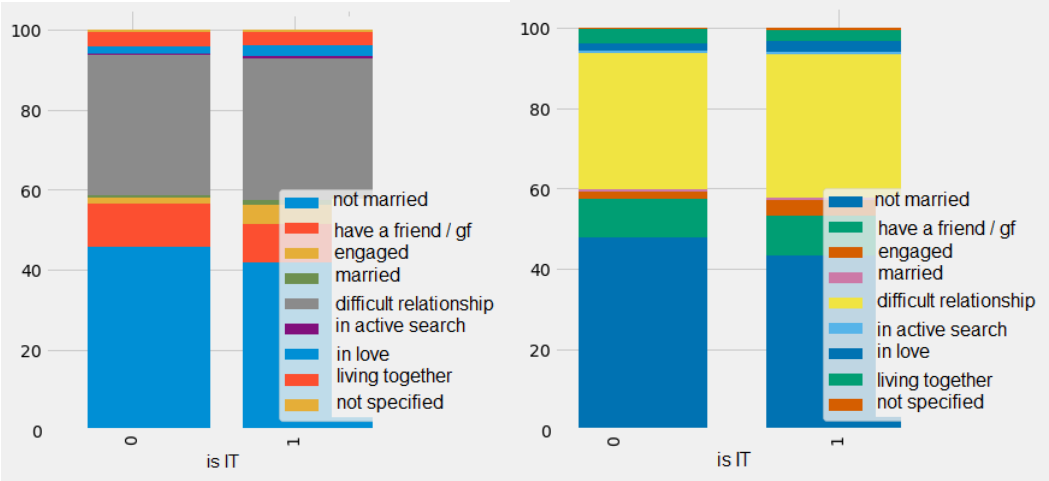

Fig. 6 a, b: Marital status

Table 1: Marital status

	interests		activities			interests		activities	
	others	IT	others	IT		others	IT	others	IT
not married	3,086	590	3,093	739	in active search	31	8	31	7
have a friend / gf	717	138	642	167	in love	122	38	115	46
engaged	105	65	103	66	living together	245	47	229	47
married	39	16	45	11	not specified	35	10	29	12
difficult relationship	2,368	502	2,209	609					

The marital status of users of different groups practically does not change from group to group (Figure 6a, b; Table 1).

V. CONCLUSION

Thus, based on the statistics on the “interests” attribute, a generalized portrait of an IT specialist and an online marketer can be compiled: this is a man aged about 30 years old, not married, or defines his matrimonial status as “everything is complicated”. He speaks an average of two languages, works for an organization, or studies at a university. He has about 370 followers on VK.

Based on the statistics, the data obtained for the 'activities' field are very close to the data of the 'interests' field and give a similar picture of the generalized portrait of the “internet marketer”: this is a man aged about 30 years old, not married, or defines his matrimonial status as “everything is complicated". He speaks an average of two languages, works for an organization, or studies at a university. He has around 400 followers on VK.

The main problem in the study is that a significant obstacle is the closed profiles of users of social networks. Thus, in introverted (accounts with empty attributes) accounts, it is impossible to find information for analysis due to the lack of the necessary account attributes.

Thus, it turned out that in the framework of the study, researchers had learned how to segment users into those identifying themselves as IT specialists or online marketers and ‘other’ users (who did not identify themselves in a social network) using machine learning methods. The user identification was carried out on the basis of the "interests" and "activities" fields, but it is not possible to predict, using the machine learning methods, the likelihood that a user can be assigned to IT or online marketing if he/she did not fill in the "interests" and "activities" fields. In other words, the model is not enough for a reliable classification of features, and the quality of the available features is too low due to large omissions.

ACKNOWLEDGMENT

The study was supported by a grant from the President of the Russian Federation for state support for research by young Russian scientists – candidates of science (project MK-698.2019.6).

REFERENCES

1. D.D. Burkaltseva, O.A. Guk, A.S. Tyulin, “Economic development based on the use of digital technologies”. Security concerns. Materials of the 4th international scientific-practical conference. Ministry of Education and Science of the Russian Federation; V.I. Vernadsky Crimean Federal University; Institute of economics and management; Department of Business Informatics and Mathematical Modeling, 2018, pp. 7-9.
2. The digital economy. The Program. Available at: https://data-economy.ru
3. C. D'Souza, D. Williams, “The digital economy”. Bank of Canada Review, 2017, pp. 5-18.
4. S. John Walker, (2014). Big data: A revolution that will transform how we live, work, and think.
5. I.A.T. Hashem, I. Yaqoob, N.B. Anuar, S. Mokhtar, A. Gani, S.U. Khan, “The rise of “big data” on cloud computing: Review and open research issues”, Information systems, 47, 2015, pp. 98-115.
6. S. Athey, “The impact of machine learning on economics”. In The economics of artificial intelligence: An agenda. University of Chicago Press, 2018.
7. J. Gubbi, R. Buyya, S. Marusic, M. Palaniswami, “Internet of Things (IoT): A vision, architectural elements, and future directions”. Future generation computer systems, 29(7), 2013, pp. 1645-1660.
8. K. Ashton, “That ‘internet of things’ thing”. RFID journal, 22(7), 2009, pp. 97-114.
9. C. Stergiou, K.E. Psannis, B.G. Kim, B. Gupta, “Secure integration of IoT and cloud computing”. Future Generation Computer Systems, 78, 2018, pp. 964-975.

10. Z.R. Andam, (2014). "e-Commerce and e-Business". Asia and Pacific Training Centre for Information and Communication Technology for Development.
11. R.I. Dunbar, V. Arnaboldi, M. Conti, A. Passarella, (2015). The structure of online social networks mirrors those in the offline world. Social networks, 43: 39-47.
12. J.C. Rochet, J. Tirole, "Platform competition in two-sided markets", Journal of the European economic association, 1(4), 2003, pp. 990-1029.
13. P. Popiel, "Boundaryless" in the creative economy: assessing freelancing on Upwork" Critical Studies in Media Communication, 34(3), 2017, pp. 220-233.
14. L. Mishel, "Despite Freelancers Union/Upwork claim, Freelancing is not becoming Americans' Main Source of Income", Economic Policy Institute Briefing Paper, 2015, p. 415.
15. V.L. Purcarea, I.R. Gheorghe, C.M. Gheorghe, "Uncovering the online marketing mix communication for health care services". Procedia Economics and Finance, 26, 2015, pp. 1020-1025.
16. P. Dey, S. Roy, "Social network analysis. In Advanced methods for complex network analysis". IGI Global, 2016.
17. R. Wölfer, N.S. Faber, M. Hewstone, "Social network analysis in the science of groups: Cross-sectional and longitudinal applications for studying intra-and intergroup behavior". Group Dynamics: Theory, Research, and Practice, 19(1), 2015, p. 45.
18. S.C. Chen, C.P. Lin, "Understanding the effect of social media marketing activities: The mediation of social identification, perceived value, and satisfaction". Technological Forecasting and Social Change, 140, 2019, pp. 22-32.