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

IN-DEPTH ANALYSIS OF PUBLISHERS IN TRAVEL AFFILIATE MARKETING BASED ON AVIASALES DATA

Master's Thesis by the 2nd year students:

Makarkina Irina Daniilovna

Moiseeva Anastasiia Nikolaevna

Soldaeva Nataliia Aleksandrovna

Academic advisor:
Elvira V. Strakhovich, Associate Professor

Saint Petersburg
2021

ЗАЯВЛЕНИЕ О САМОСТОЯТЕЛЬНОМ ХАРАКТЕРЕ ВЫПОЛНЕНИЯ ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

Мы, Макаркина Ирина Данииловна, студентка второго курса магистратуры направления 38.04.02 «Менеджмент», Моисеева Анастасия Николаевна, студентка второго курса магистратуры направления 38.04.02 «Менеджмент» и Солдаева Наталия Александровна, студентка второго курса магистратуры направления 38.04.02 «Менеджмент» заявляем, что в нашей магистерской диссертации на тему «Глубинный анализ аффилиатов в сфере путешествий на основе данных компании «Авиасейлс»», 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)».

Макаркина И.Д

Моисеева А.Н.

Солдаева Н.А.

01.06.2021

STATEMENT ABOUT THE INDEPENDENT CHARACTER OF

THE MASTER THESIS

We, Makarkina Irina, second year master student, program 38.04.02 «Management», Moiseeva Anastasiia, second year master student, program 38.04.02 «Management», Soldaeva Nataliia, second year master student, program 38.04.02 «Management», state that our master thesis on the topic «In-Depth Analysis of Publishers in Travel Affiliate Marketing Based on Aviasales Data», 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)».

Makarkina I.

Moiseeva A.

Soldaeva N.

01.06.2021

АННОТАЦИЯ

Авторы	Макаркина Ирина Данииловна Моисеева Анастасия Николаевна Солдаева Наталия Александровна
Название ВКР	Глубинный анализ аффилиатов в сфере путешествий на основе данных компании «Авиасейлс»
Образовательная программа	38.04.02 «Менеджмент»
Направление подготовки	Бизнес-аналитика и большие данные
Год	2021
Научный руководитель	Страхович Эльвира Витаутасовна
Описание цели, задач и основных результатов	Данная ВКР представляет углубленный анализ партнеров-аффилиатов компании Aviasales. Основные задачи включают анализ веб-контента и категорий сайтов-аффилиатов с применением методов машинного обучения и обработки естественного языка, а также исследование современных особенностей аффилиатного маркетинга в сфере путешествий. С теоретической точки зрения работа вносит вклад в ограниченный набор исследований, посвященных теме аффилиатного маркетинга. С точки зрения бизнеса работа позволяет ответить на стратегические запросы компании Aviasales и предлагает модель классификации аффилиатов и практические рекомендации для менеджмента.
Ключевые слова	Аффилиатный маркетинг, большие данные, машинное обучение, обработка естественного языка, Aviasales, сфера путешествий

ABSTRACT

Master Student's Names	Makarkina Irina Moiseeva Anastasiia Soldaeva Nataliia
Master Thesis Title	In-Depth Analysis of Publishers in Travel Affiliate Marketing Based on Aviasales Data
Educational Program	38.04.02 «Management»
Main field of study	Business Analytics and Big Data
Year	2021
Academic Advisor’s Name	Elvira V. Strakhovich
Description of the goal, tasks and main results	The paper provides an in-depth analysis of the affiliate partners of the Aviasales company. Main tasks include the analysis of the web content and categories of the affiliates with the application of Machine Learning and Natural Language processing tools as well as the investigation of the peculiarities of the modern affiliate marketing. From the academic point of view, the paper contributes to the limited number of research devoted to the sphere of affiliate marketing. From the business perspective the thesis allows to respond to strategic questions of Aviasales company as well as offers affiliates classification model and practical managerial recommendations.
Keywords	affiliate marketing, Big Data, Machine Learning (ML), Natural Language processing (NLP), Aviasales, travel industry

TABLE OF CONTENTS

INTRODUCTION

With the rapid development of modern technologies comes the rapid change in consumer behaviour as well as development of both digital marketing in general and digital marketing tools in particular. Indeed, digitalisation leads to more knowledgeable and demanding consumers and diminishes the span of consumers' attention, which means that sellers need to find more and more exquisite ways to stand out. Moreover, the internet-based marketing approaches galore have switched consumers' attention to pieces of advice, opinions and various kinds of suggestions expressed online. At the same time on-line marketing usually requires much less investments than traditional marketing and together with a thought through digital strategy can bring substantial results. Therefore, the rise of such digital instruments as affiliate marketing is an expected and understandable phenomenon.

Affiliate marketing is the type of online marketing where a commission is earned by a third party (affiliate) for the promotion of other people's or companies' products (merchant). Affiliate marketing defined as performance marketing and associate marketing. Affiliate marketing integrates with three parties: Advertiser/Merchant, Publisher/Affiliate, and Consumer/Buyer. Merchants can be selling any type of company's products like electronics, books, clothing, and air tickets online or could be insurance companies selling policies etc. Publishers are the ones who forward advertiser's products or services through its website, blog or social media account. Consumers are the party with the major share of power. They represent a very prominent part of this cycle since they are the ones attracted by the advertisements and making an action. Clicks from publisher's website to advertiser's website and purchases of the product afterwards are called conversion.

Affiliate marketing experienced rapid growth in 2010 and has been popular ever since, remaining one of the most popular ways of mutual collaboration aimed to drive sales, increase brand awareness and generate passive income. Nowadays, affiliate marketing has turned into a billion-dollar industry and more and more companies are considering using it in their practice. In 2016, U.S. retailers spent $4.7 billion on affiliate marketing. Astoundingly, by 2020 U.S. affiliate marketing spend is expected to rise to $6.8 billion.

Therefore, it can be said that affiliate marketing has truly become the buzzword of the 21st century and, thus, research connected to it will contribute to the development of modern

marketing strategies. This thesis will contribute in both theoretical aspects of affiliate marketing, as it will address a substantial research gap, and a practical or empirical one, as the investigation is based on the data from the existing company.

Research gap

Despite affiliate marketing being a popular and even trendy topic in the business community, the academic literature devoted to it is quite limited. Researchers have devoted their attention to the economic benefits of affiliate programs, studied remuneration mechanics and the issues of trust in the context of affiliates. However, the number of academic papers is modest and, thus, as a result of conducted literature review several research gaps have been identified.

Although a couple of authors attempted to classify affiliates, those endeavors have been based on the guts and experience and were not supported by case studies or in-depth analysis. Even more, the main object of the research in academic literature is, in fact, predominantly advertisers and not the affiliates. Another issue is generalisation: only a few papers are considering affiliate programs through the prism of a certain industry. The situation can be explained by the fact that though many businesses are interested in understanding the affiliate marketing tool, they prefer to keep the results of the analysis in-private as it can be considered as a source of competitive advantage.

Aviasales partnership

As it was mentioned, lack of case studies and industry specific analysis for the major problems of current affiliate marketing research. That is why the partnership with Aviasales in terms of this thesis increases the contribution of this paper into the global affiliate marketing knowledge. It is necessary to mention that Aviasales is the largest Russian flight tickets metasearch, otherwise known as aggregator engine. The basic mechanism behind every metasearch engine is that the aggregator sends queries to several search engines, but the results come into one list categorised based on where they came from. Therefore, from the very idea of how the company works and what they do, the topic of affiliate marketing and its characteristics is of a particular interest.

Research questions

In terms of this study we will focus on questions of how companies choose affiliate programs: namely, whether the content or type of their website influences the number of programs or the type of the affiliate verticals they participate in.

The thesis will be divided into two parts. Firstly, the hidden patterns of the data will be investigated and the unifying characteristics of affiliates will be identified. Secondly, in accordance with Aviasales requirements the main types of the affiliates considered will be content sites, service sites and cashback and promo code sites. Thus, the main analysis will be provided in terms of this viewpoint. Furthermore, peculiarities of English and Russian language sites will be investigated. The partnership with Aviasales will allow to test hypotheses on real-life data and to find both business and academic insights.

Based on the research gaps mentioned above, the following formulation of the thesis is: 'In-Depth Analysis of Publishers in Travel Affiliate Marketing Based on Aviasales Data'. It naturally divides into two sub-questions:

1. Which types of websites most often participate in the affiliate programs? In how many affiliate programs?
2. Is there any specific pattern between the type of an affiliate and a vertical of advertiser?

Aim of the paper and object of the research

The aim of this paper is to study in-depth which characteristics possess the affiliate part of the affiliate marketing and how they can be used by Aviasales. This will lead to investigation of the quality of such forms of partnership as affiliate marketing. In particular, the main focus will be on the compatibility of the industry between the company that offers an affiliate program and the type of an affiliate. And Aviasales data, thus, will be the object of the research. The business result of the research will be the managerial insights that will allow Aviasales to adjust its marketing strategy to each class identified.

Big Data application to the case

To be able to carry out the intended analysis Big Data and Machine Learning tools are required.

The term “Big Data” refers to data that is so large or complex that it’s difficult or impossible to process using traditional methods. The act of accessing and storing large amounts of information for analytics has been around a long time.

The use of Big Data analytics is flexible and, hence, applicable to various fields. With the use of Big Data a lot there has been enormous growth in multiple industries, for example, banking and travelling.

The reason behind popularisation in such a short time span is that not only does humanity generate data nowadays in a manner that has been unseen before, but also the value derived from Big Data is essential to almost any business. Big Data is used in the questions of cost and time reduction, new product development and smart decision making.

Therefore, the importance of Big Data nowadays cannot be overestimated, and this research will be including different algorithms of Big Data to provide a unique and substantial value to the topic of affiliate marketing itself and to the particular case of Aviasales company.

In particular, since the part that provokes questions the most is the content of websites and is closely related to the language itself, the application of Natural Language Processing (NLP) algorithms is required.

According to Anthony Pesce, Natural Language Processing is a field that covers the structure of human language and focused on computer understanding and manipulation of it. Moreover, due to language being an essential part of human interactions and lives in general the applications for those algorithms are almost non-limited. Despite understanding languages and semantics being relatively inconsiderable and not burdensome for most people, machines have a different perception of this matter. The considerable amount of unstructured data, semantic context that oftenly is absent and lack of formal and clear rules makes it for computers genuinely difficult to understand human speech or substitute them for routine and repetitive tasks. However, the NLP is essential in human lives, since a lot of technologies daily used by people are made on algorithms of the NLP, such as email filters, smart assistants, language translation, digital phone calls and even Internet search. That is the reason Machine Learning and Artificial Intelligence (AI) are acquiring consideration and momentum, with stronger human reliance on processing frameworks to impart and perform given tasks.

Moreover, both unsupervised and supervised machine learning tools will be applied in order to answer the formulated research questions and respond to the business needs of Aviasales. Thus, unsupervised learning, meaning the usage of unlabelled data, will allow to

study the data, find the hidden data patterns and point out certain site clusters. Therefore, the clustering approach will help to group similar affiliates. Supervised learning approach such as classification, implying already labelled data, will allow to respond to the business needs and analyze affiliates in the viewpoint that responds to the demands of Aviasales.

Thesis outline

The paper will be structured as the following: firstly, the literature review will be introduced. It will cover the definition of affiliate marketing and structure of affiliate marketing system, then will give the overview of the previous attempts to group affiliates and then discuss the main directions of affiliate research such as online trust. The specifics of affiliate marketing in the travel industry will also be presented as well as the application of Machine Learning (ML) tools to this field.

In the second chapter the empirical analysis of affiliate marketing strategy based on the travel company Aviasales will be introduced, consisting of data description and methodology of the study. First, unsupervised learning in a form of cluserization aimed at finding hidden patterns in the data will be performed. Second, classification of the websites in accordance with Aviasales business goals will be performed. For classification purposes three models: Linear Support Vector Classifier, Gradient Boosting and CatBoost models will be built and compared to achieve the best possible results. Lastly, the conclusion and the area of further research in this sphere will be presented as well as the managerial recommendations for Aviasales.

Overall, in order to fully exploit the potential of affiliate programs as an internet promotion tool, companies and the industry of online marketing in general should comprehend the distinct attributes and peculiarities identified with it, the demeanor of customers towards it and the manners in which buyers' are affected by its substance. The findings are expected to be useful for Aviasales and marketing managers in general in their evaluation and management of affiliate programs.

Chapter 1. LITERATURE REVIEW

1.1. Affiliate marketing literature review

The academic research on affiliate marketing has been quite limited. The major discussion revolves around the definition of affiliate marketing itself, issue of trust and

publisher viewpoint. Substantial gap in analysis of affiliates and representation of the mechanism from affiliated point of view exists.

1.1.1. The notion and mechanism of affiliate marketing

Before diving into the underlying issues of affiliate marketing it is important to study what affiliate marketing actually is, who are the main parties involved and how this instrument actually functions.

First of all, affiliate marketing can be attributed to Internet based marketing among other tools like search engine marketing, email marketing, social media and influencer marketing, content marketing etc (Olbrich et al., 2019). Affiliate marketing assumes that an affiliate (a third party) is paid for every visitor that comes to a merchandiser (advertiser or publisher) website from hyperlinks published by this third party. Thus, the three main participants of the process are:

1. Merchandiser / advertiser – a party that is willing to sell its products or services via online devices of applications;
2. Buyer – an individual or a company that ends up buying the product or service;
3. Affiliate / publisher – An intermediary that uses its website or app to publish a hyperlink that leads to the merchandiser's website (Dwivedi, 2017).

It is also important to note that affiliate marketing cannot exist without actual customers. They are the ones generating revenue streams and engaging in long-term relationships with affiliates. Therefore, based on the information presented above, schematically affiliate marketing can be presented in a following way (Figure 1):

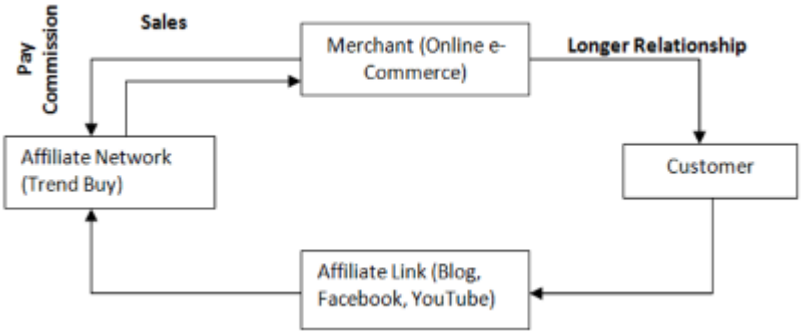


Figure 1 Affiliate Marketing Framework (Source: Suresh et. al, 2018)

Moreover, popularity of affiliate marketing can partly be attributed to its underlying concept that requires minimal or, in some cases, even zero expenses from affiliates. In most cases the affiliate's compensation implies the form of Cost Per Click (CPC) or Cost Per Action (CPA). This means that the money is paid by the advertiser for a certain customer's activity: click, a form fulfillment or an acquisition of a product. Thus, in such a scheme, the income of an affiliate largely depends on the traffic and leads its site generates to a publisher website, which in its turn allows the publisher to partially control the quality of the channel. To trace the traffic affiliate mechanisms usually use cookies or utm-tags allowing to define consumer actions via scripted code (Tsvetkova, 2021).

Thus, it can be said that most affiliate programs are commission based, though the payment agreement may depend on the nature of the industry to which the advertiser belongs as the nature of the business and importance of affiliate marketing in a certain context differs. For example, the retail industry contributes 43% of the complete affiliate promoting market income followed by telecom and media, travel and recreation areas, which contribute 24% and 16% respectively (Wang et. al, 2014). Moreover, different business domains have different customer acquisition costs that influence the amount of the affiliates' commission. Average commissions vary from 1 to 15% percent (Haq, 2012).

It is important to note that according to Affiliate Marketing Statistics (AFFSTAT, 2016) the major part of partner advertisers advance Business-to-Consumer (B2C) offers. Namely, 81.4% choose projects connected to B2C, while remaining 18.6% spotlight their endeavors on Business-to-Business (B2B) items. The data on possible overlap is not presented in the report.

In general, it can be said that affiliate marketing in recent years has truly become one of the most widespread digital instruments. As indicated by a Forrester report appointed by Rakuten (2016), 81% of promoters and 84% of publishers use affiliate marketing. Furthermore, according to Statista (2019), in the United States (U.S.) alone affiliate marketing expenditures are expected to triple in comparison to the last decade and reach $8.2 billion by 2022. Worldwide, it was assessed in the Awin Report (2021), that in 2017 promoters put $13 billion in partner advertising.

1.1.2. Approaches to affiliate categorization

Despite affiliates being one of the core parts of the affiliate marketing system, attempts to study, classify or group them are quite limited. Academic papers only briefly mention possible ways to divide publishers into manageable groups.

Thus, Goldschmidt et al. (2003) proposed to divide affiliates based on traffic generated by them. The following categories were described:

- Hobby sites: widespread type of affiliates that usually generates quite moderate traffic of less than 10 000 visitors per month. Hobby sites usually contain a mix of relevant to the particular topic information and author's personal information.
- Vertical sites: sites devoted to a chosen topic, for example, beauty, dating etc with traffic between hobby sites and super-affiliates. They provide in-depth information on the subject and usually have a focused audience.
- Super-affiliates: sites bringing most traffic (more than 50 000 visitors per month) and profits, however, relatively unfocused as they try to appeal to a wider audience. An example of this type of site are mass media sites.

Another classification developed by the Internet Advertising Bureau (IAB, 2016) also revolved around the issue of traffic, nevertheless, emphasized the different aspect. The main idea was to take into account not the general capacity to generate traffic but the instrument chosen to promote the advertiser. Thus the affiliates were divided into:

- Reward sites: the affiliate offers a reward or a bonus for a consumer that buys through its link.
- Content sites and blogs: such sites provide unique content on the topics that cover audience's interests
- E-mail sites: this type of publishers actively use own databases and newsletters in order to attract consumers
- Comparison sites: site that contains mechanisms allowing users to compare offers on certain products or services
- Retargeting sites: sites that track visitors interests and actively use digital instruments to re-engage them
- Pay-per-click affiliates: sites that use custom landing pages and keywords to stimulate purchase
- Voucher and deal sites: sites offering coupons or various discounts as a compliment for the purchase

- Social sites: affiliates that actively use social networks to generate traffic

Both classifications approach is limited and does not provide affiliate managers with in-depth information. Thus, Goldschmidt's approach can be considered as quite general, while Internet Advertising Bureau's (IAB) approach (IAB, 2016) does not account for the fact that affiliates can use multiple traffic generation strategies simultaneously.

It is important to note that both approaches are based on the authors' own experience and knowledge of the field and are not supported by empirical studies.

1.1.3. Main directions of affiliate marketing research

Overall, affiliate marketing includes a lot of various aspects and peculiarities. Therefore, the researchers have fragmentally discussed different issues: from the structure of the affiliate systems to the appropriate forms of commissions.

Namely, several studies have been devoted to the economic effect of affiliate marketing. For example, Mican (2008) stated that the usage of affiliate marketing can increase sales of the advertiser. Edelman et al. (2014) attributed this to the fact that affiliate marketing systems allow an advertiser to attract and manage a vast amount of miscellaneous websites without substantial money investments. Nevertheless, Akcura (2010) emphasized that excessive usage of affiliate schemes, though positively influences profits, undermines consumers' loyalty and may result in a loss of customers in the long term.

Attempts to give recommendations on affiliate marketing management were given by Ivkovic et al. (2010), who discussed general requirements for affiliate programs: modern software, constant technical support availability, adequate pricing, clearly stated commission policy. Discussion on commission and payment mechanisms was also introduced in Libai et al. (2003) paper, where pay per conversion and pay per lead were studied. The findings stated that the choice of payment depends on external factors like the number of affiliates participating in the program in general. At the same time Iva (2008) found that pay per sale mechanism to be the most popular one in the study of Croatian hotels affiliate programs.

Moreover, Bhatnagar et al. (2001), pointed out the importance of understanding consumer search behavior. This means that affiliates that participate in affiliate marketing programs have to respond to the needs of the consumer and utilize Search Engine Optimization (SEO) practices in order to be maximally efficient. Moreover, Papatla et al.

(2002) stated that affiliate programs bring the most results when the businesses of a merchandiser and a publisher align.

A sufficient part of academic research on affiliate marketing is devoted to the issue of online trust as it influences consumer's decisions significantly. Moreover, many researchers point out that trust is an essential and initial requirement for sustained online demand and, therefore, for marketing mechanisms like affiliate programs.

Giving the definition, trust is the decision of an individual or a group to rely on another individual or a group (Hooghe, 2017). Thus, trust can be considered as the foundation of every relationship either between individuals or groups of people and, consequently, between buyers and advertisers. Moreover, for the following three characteristics necessary in order for trust to appear can be synthesized:

1. At least 2 participants should be engaged in the relationships – a trustor and a trustee;
2. The factor of vulnerability must be present as trust occurs in situations, where one of the actors bears some risks;
3. Trust is subjectable to the context of circumstances.

Online relationships in terms of this paper are defined as exchanges mediated by Internet-based channels. The management of such relationships that includes digital marketing strategies and, therefore, affiliate marketing strategy, presents both challenges and opportunities for modern companies. Namely, they are influenced by the technological advancements and constant development of e-commerce, social media, mobile, AI and augmented reality (Supermetrics, 2021). Moreover, the context in which online relationships exist is quite peculiar. Thus, while shopping at a physical store, a customer is exposed to aromas, audios, visuals and sometimes faces and voices (Drugău-Constantin, 2018). However, an online shopper is deprived of almost all of the external factors, apart from visuals. Therefore, trust in such situation becomes a natural requirement for success.

Furthermore, while shopping online consumers are exposed to several additional risks, e.g. to be deceived or hacked, to disclose personal data, to choose the wrong colour, size or model etc (Lakshmi et al, 2019; Racherla et al, 2012). Nevertheless, data collected by BrightLocal (2018) indicates that each year more and more people are willing to fully trust online sources as well as reviews and opinions expressed on-line, which favors the development of affiliate marketing programs. Namely, 93% of consumers admit that their purchase decisions were influenced by online reviews (Podium, 2021).

According to Agyei et al (2020) trust significantly influences customer engagement, which in its turn increases loyalty. At the same time four main type of trust can be identified, namely, trust in a service provider (willingness to trust that an entity can deliver its promise), trust in a regulator (sense of security and assurance in protection by government and legislation), economy-based trust (trust based on calculation of economic benefits and risks), and information-based trust (involving knowledge of the other party) (Agyei et al, 2020). It is important to point out that trust in a service provider and trust in a regulator mainly drive customer engagement.

In terms of affiliate marketing, Danielle et al. (2009) pointed out that advertiser's success, especially in the travel industry, significantly depends on consumer acceptance of affiliate websites as it directly influences the number of generated leads. It is also important to note that affiliates can be considered as touchpoints that are often perceived by consumers as advertiser's brand representatives (Li et al., 2009). Therefore, fraudulent behaviour demonstrated by affiliates undermines the trust between a consumer and an affiliate, an affiliate and an advertiser and a consumer and an advertiser. This idea has also been reflected in Papatla et al. study (2002) that discussed the importance of inter-organizational trust and consumer loyalty in the context of affiliates.

Research provided by Gregory et al. (2014) revealed that certain characteristics of affiliates influence the degree of consumer trust. Thus, the more such websites show their competence and integrity by providing quality content and additional information on affiliate links, the better they attract consumers. Among trust-determining factors Gregory et al. (2014) also pointed out company size, website reputation, and web interface design. These findings repeated Duffy's (2005) statement that affiliates' critical factor of success is the ability to create appealing websites.

Thus, trust is an essential part of online relations on which affiliate systems are based. Consequently, the issue received attention from academics and researchers. Nevertheless, while the definition of trust has been theorized for quite a long time, the notion of online trust or trust in e-commerce is quite a recent and much more complex phenomenon that has its own peculiarities and, thus, is to be further explored. Another literature gap can be observed regarding the concept of trust or online trust specifically applied to the travel industry, which is the industry under question in terms of this paper. Hitherto this moment, there has been little systematic review of this body of work.

All in all, academic research on the issue of affiliate marketing is quite limited. Moreover, it mainly covers issues connected to an advertiser. In an attempt to close a research

gap throughout this thesis the main emphasis will be put on affiliates and their peculiarities. A brief overview of previous attempts to classify the publishers is presented further.

1.2. The context of the research

1.2.1. Company overview

Aviasales is the largest Russian flight tickets metasearch (Aviasales, 2021) founded in 2007 by Konstantin Kalinov. The term 'metasearch' implies that the company does not sell any tickets itself but finds and compares the best offers. A user then can be redirected into a partner's link, where he/she can buy the tickets. Thus, a user receives the information free of charge as the company profits from the partners commissions (Chernikova, 2014).

Headquartered in Phuket the company has offices both in Moscow and Saint Petersburg (Chernikova, 2014). Moreover, Aviasales actively operates on Kazakhstan's, Uzbekistan's, Belarusian, Ukrainian and Tanzanian markets (Forbes, 2020). Starting as a personal the company now has entered the list of top 10 most expensive companies in the Russian Internet. Thus, in 2020 it was evaluated by Forbes as a 180 million dollars company. Furthermore, 15 million people use the service every month (Aviasales, 2021).

Among its main competitors Skyscanner, Momondo, Kayak and Yandex.Avia can be named.

Aviasales has its own partnership program TravelPayouts. The program was founded in 2011 under the guidance of a company's founder Konstantin Kalinov and became popular in an instant (Baidin, 2018). In 2013 Artemiy Lebedev, a famous Russian designer, joined the program as a partner. Moreover, a little bit later large airports like Pulkovo and Sheremetyevo got involved as well.

According to its rules the affiliate can receive 50-70% from aviasales revenue from a ticket sold. Aviasales usually receives 2,2% commission (Aviasales, 2021). To the moment of this thesis being written Aviasales has already paid out 1 640 588 909 roubles.

1.2.2. Travel industry specifics and peculiarities of affiliate marketing usage

The concept of affiliate marketing has proven to be successful and widespread among various industries. Nevertheless, in terms of this paper it is important to consider affiliate marketing in the context of travel. Therefore, it is necessary to establish what the travel industry actually is and provide a brief overview of such context.

According to Cambridge Dictionary (2021), the verb “travel” implies the movement of people from one place to another. Thus, the travel industry is devoted to offering all types of assistance connected with such movement. This involves services like various types of transportation and transfer as well as assistance in accomodation, etc. Moreover, it is also important to note that despite travel and the tourism industries being quite similar, they reflect different scope. Basically, the tourism industry is relatively narrow and is devoted to satisfying the needs of the customers that head out to a different area for pleasure and joy (Revfine, n.d.). Hence, this implies that the travel industry is a more broad term than tourism since it covers a more extensive number of movement purposes and incorporates trips to non-vacationer locations.

According to Revfine (n.d.) under the category of travel industry the following businesses can be attributed: transportation, insurance, accomodation, food and restaurants, entertainment. Thus, the structure of travel industry can be presented in a following way (Table 1):

Travel industry				
Transportation industry	Insurance sector	Accomodation	Food and restaurants	Entertainment
means of transportation sectors (airline sector, water transport sector etc) rent of various means of transportation (car rent, bike rent, ferry rent etc) transfer coach travel	various types of insurances (visa insurance, medicine insurance, loss of luggage insurance, flight cancellation insurance)	various types of accommodation (hotels, hostels, camping, bed and breakfast, time-share accommodation, shared accommodation)	restaurants, cafés, nightclubs, catering services	adventure and amusement parks shopping malls casinos museums sightseeing tours

Table 1 Travel industry structure (Source: Revfine, n.d.)

It is important to note that sub-industries of the travel industry are very loosely defined and, thus, can also include legal aid on issues connected to travel, part of banking, namely, co-branded cards that allow to receive discounts and miles etc. Moreover, travel

agencies are a part of the industry as well, however they often operate across almost all presented sub-domains.

In general, setting aside COVID-19 pandemic that significantly destabilized many businesses, it can be said that the travel industry has demonstrated stable growth throughout the recent years. Figure 2 illustrates that the number of tourists steadily increases each year, despite several worldwide crises including SARS and Bird flu:

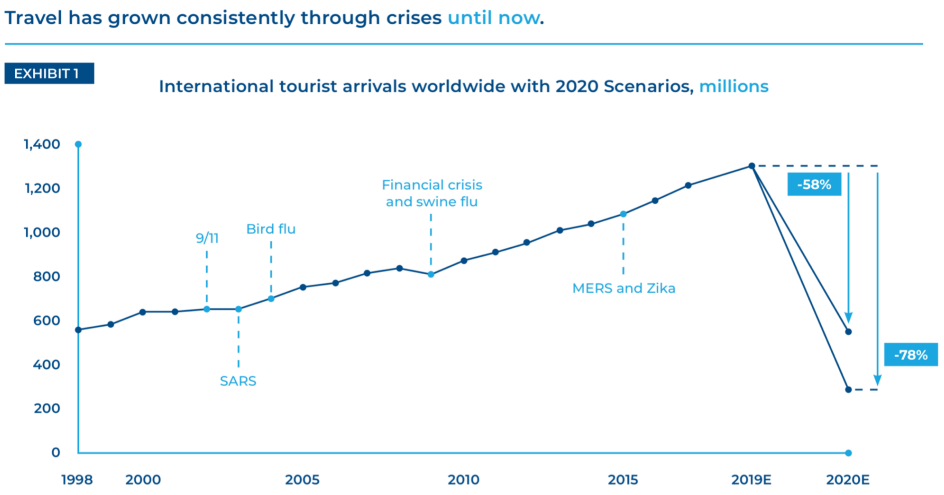


Figure 2 International tourist dynamics (Source: World Bank, UNWTO, 2020)

Thus, apart from digitalization and the coronavirus disease (COVID-19) situation, the industry itself can be considered quite attractive. In 2019, global leisure travel spend reached $4.715 billion, while the total travel industry, the one that includes all of the reasons people might have for travelling, archives 9.25 trillion U.S. dollars (Lock, 2020). Even looking at the country-scale the U.S.'s travel and tourism industry contribution was 580.7 billion U.S. dollars in 2019 which makes it the biggest index per country (Lock, 2020).

As for the peculiarities of industry that influence the adoption and spread of affiliate marketing among its players the following can be mentioned:

- As was discussed earlier the travel industry is connected with movement, meaning that during the service consumption prevailing amount of customers are forced to step outside their comfort zone and find themselves in a completely new environment about which their knowledge is limited. Therefore, to diminish such risks they tend to rely on reviews and recommendations (Oktadiana et al., 2018). Affiliate marketing

allows to reach and attract consumers through various travel blogs and opinion leaders.

- The industry is extremely competitive, as a lot of players compete for the same customer base and have similar business models and offerings. Moreover, the emergence of aggregators like Aviasales or Booking made information and comparison easily available for consumers that increased the pressure even more. Even more, due to technological development new business models like AirBnB emerged fastly and continued to flourish challenging the titans of movement (Sherwood, 2019). Thus, competitors are constantly looking for ways to attract customers, especially in a digital space and are prone to try affiliate marketing strategies that require relatively low investments.
- Technological factors have always been one of the most influential ones in terms of the travel industry. The way the customers interact with players of the travel industry is constantly changing along with digitalization. It is important to note that more and more travel services sales happen online. According to Online Travel Booking Statistics (2021) 70% of travellers research travel on their smartphone, 83% of the U.S. adults now prefer to book their travel online, 5% of the United Kingdom (the UK) travellers feel comfortable researching, planning and booking trips to new destinations using only their mobile, 82% of all travel bookings in 2018 were made online via a mobile app or website, without human interaction. Moreover, according to Deloitte (2019) technological innovations and breakthroughs such as artificial intelligence, mobile applications and the Internet of Things (IoT) will continue to enhance the experience of travelers, ease the process of traveling itself and eliminate some of the existing customers' pains, creating more necessity for players to be peculiar and visible online . Therefore, the business rapidly moves to digital and, thus, calls for online promotion strategies like affiliate marketing.

Therefore, the profitability, social aspects and the high level of digitalisation makes the travel industry an attractive market for affiliates and publishers. In addition to it, the Chief Executive Officer (CEO) of Affiliate Marketing Navigator, an award-winning outsourced affiliate management company, Geno Prussakov says that their study showed that a travel affiliate program ranks fourth in the top 20 leading affiliate niches (Travelpayouts Blog, 2017).

As COVID-19 situation has changed the travel industry significantly it is also important to assess affiliate marketing's existence in a "new normal" travel industry.

Taking into the consideration that in a majority of cases international mobility and traveling are still prohibited, there is an obvious sharp increase in domestic flights. For affiliate marketing that would mean switching from the international merchant or merchants offering services in other countries to domestic ones that may also require changes in content published by affiliates. However, it does not mean that the affiliates should completely eliminate all their international partners rather than simply changing the proportions. The cooperation with international partners may still be preserved but just focused more on future possibilities of travelling and future benefits of this cooperation for both parties.

Another trend caused by the recent epidemiological worldwide situation is the shift in travel purpose. Prior to COVID-19 such reasons for traveling as business would make more in proportion to leisure. The ability to work and organisee meetings and conferences remotely via applications like Microsoft Teams, Zoom, Webex etc., however, led to the business travels being unnecessary, while family holidays – still a desirable goal (Thaichon et al, 2019). The change can be seen in Figure 3:

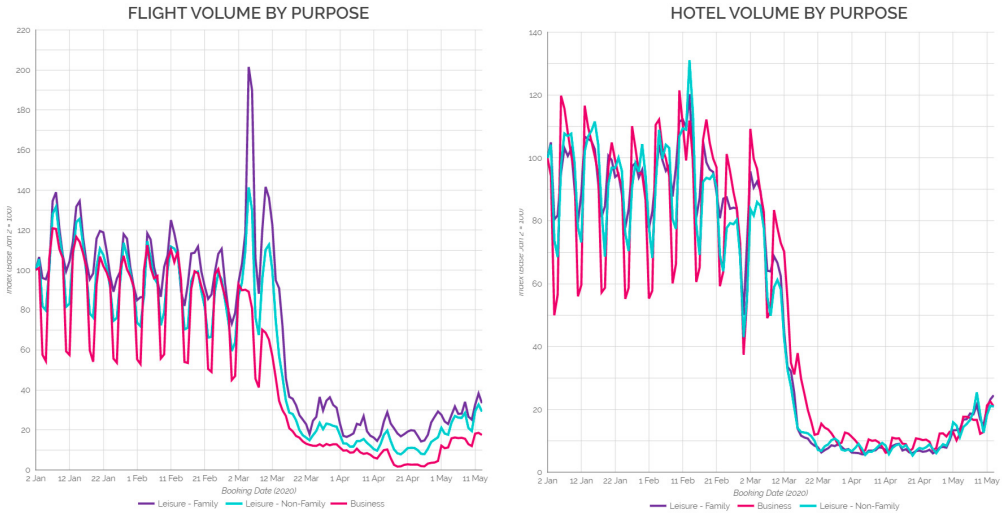


Figure 3 Flight and hotel volumes by purpose (Source: Novoselov, 2020)

The technology was not the only factor, however, making the changes happen, the organisational and behavioural changes in corporations also played a significant part. The recent year has shown that the level of productivity does not differ that much even if an employer works remotely but the operational costs associated with personnel decrease. Thus, it is expected that the number of business related travels will not grow substantially in the next few years.

Applying this information to affiliate marketing, this would probably mean that the amount of advertising and travel options suitable for businessmen will decrease, while family-oriented leisure promoting advertisers will take over the market. A new study by IdeaWorks (2021), a travel consultancy, projects that around 19-36% of business related travel traffic aboard the world's airlines prior to the pandemic will not achieve the numbers existing before the covid or it will take substantial time to do it. In light of the fact that the number of international travelers has decreased and most probably will not gain back its positions soon, domestic travel has become a lucrative niche.

Another change in travelers' behaviour is that due to increased prices and numerous restrictions pushed by a majority of transportation companies, there has been an increase in people's preference towards private means of transportation such as a rental car (Hooghe, 2017). This is an individual form of transport, which allows the tourist to reduce the amount of interaction with other people, while providing maximum flexibility in travel planning. One more possible change in the travel industry is the projected switch from customers' preferences from large hotels towards more private accommodations. Severe restrictions imposed on the large accommodation facilities and logical consequences such as closure of some objects where the social distance is impossible or limited number of rooms available will make people choose more private options. These changes in the industry influence the structure of affiliate marketing programs presumably changing the balance between the types of affiliates.

Pandemic situation influences market offerings and challenges its usual perception. An example of that can be the transformation of grounded aircrafts into pop-up restaurants (Porter, 2020). The emergence of the new products and services are expected to influence affiliate marketing in general and specifically content published by affiliates.

In order for all the players of the travel industry to survive in the current global economic crisis, they need to stimulate people to plan and book vacations. However, the problem of inability to predict the new governmental restrictions, the cancelation of many bookings and thus the substantial losses encourage many clients not to book in advance. In order to stimulate bookings a majority of travel-related companies around the world are changing the existing policies, such as last-minute cancellations. Stronger demand is expected for the winter season, with reservations for Thanksgiving, Christmas, and New Year being up by 38%, 40%, and 23%, respectively, compared to 2019. This, normally, will influence the cash flows of travel organisations, as they should have the option to repay

dropped bookings and protect the budget consistently. The decrease in revenues force industry players to appeal to cheaper means of marketing, especially affiliate systems.

All in all, the new existing initiatives happening in the travel industry and the prior to the COVID-19 established demand for the services in addition to the steady long-term growth of the market indicate that the affiliate marketing is still applicable to the travel sector and may be now more than ever.

1.3. Machine Learning usage in affiliate marketing

Affiliate marketing is one of the most suitable industries for the application of Machine Learning algorithms, taking into the account that this sphere hugely relies on a big amount of various data such as texts, videos etc.

Probably one of the biggest and yet the most concerning problems in modern affiliate marketing is the over saturation of the market and specific niches. Any player that is about to enter niches that used to be profitable and ludicrous now is to face the active competition, and therefore has to come up with some mechanisms that are to diversify it from all of the other competitors or to find new niches. To overcome both problems Machine Learning algorithms can be and better be used.

Despite affiliate marketing being a domain that can provide a curious researcher with a large amount of information, the application of Machine Learning techniques to the notion in academic literature is not significantly widespread. To the contrary the business application of Machine Learning in affiliate marketing is quite popular though companies rarely transform it into detailed published studies due to commercial reasons.

Nevertheless, the following domains of applications can be pointed out:

1. Personalization. According to literature research, if the affiliate appeals to the advertiser's target audience the chances of conversion are higher. Thus, Machine Learning techniques allow users to identify sites with similar content and group them in accordance with the potentially interested audience i.e. segment the audience. Moreover, Machine Learning algorithm based tool Smartlink allows a computer to automatically decide, which link is shown to which user at a certain period of time based on the cookie data (Dobyshuk, 2020).
2. Improvement of affiliate network management and automated campaign optimization. Usually affiliate networks consist of thousands of websites that are hard to track using simple tools like Microsoft Excel. Machine learning allows to get valuable insight on

which affiliates perform better than others and, thus, adjust both KPIs and the structure of the network. Moreover, BI integration tools help managers to easily analyze and manage the performance of their campaigns (Parker, n.d.). Thus, introduction of Machine Learning tools allows to achieve high quality analysis and improvement of campaigns' efficiency.

3. Noise reduction and broken links management. Affiliate websites are extremely prone to transformation into broken links due to problems with hostings and sites shutdown. This leads to the messiness in the database and may negatively influence affiliate marketing strategy as it can be based on the old data. Machine Learning allows to detect websites that are no longer working and remove them from the network.
4. Illegal marketing and fraud detection. Thus, Mackey et. al (2018) describes the study of the U.S. Department of Health and Human Services aimed at finding ways to control and tame opioid promotion through Twitter. Therefore, unsupervised Machine Learning was applied to receive clusters of sites involved. These were illegal online pharmacies, individual drug sellers and marketing affiliates. Affiliates participating in affiliate marketing programs can represent fraudulent schemes and, thus, significantly damage the advertiser's brand. Therefore, it is extremely important to be able to fastly detect and remove this type of websites.
5. Double counting detection. Affiliate marketing system is usually based on some kind of a user action resulting in affiliate's commision. However, in some cases, when a user registers through different channels this commission doubles or even triples as the same customer gets duplicated. Machine Learning is used to prevent double counting (BizAcuity, 2021).

Therefore, Machine Learning is actively used in affiliate marketing and is regarded as a tool helping to achieve business goals, however academic studies lack real-life cases and examples.

Chapter conclusion

This chapter provided an overview of the literature devoted to affiliate marketing, Machine Learning application to affiliate marketing as well as introduced the context of the research.

Thus, it can be said that though affiliate marketing is an extremely popular marketing tool the number of detailed studies devoted to it is limited. The authors studied the economic

effect of affiliate marketing, discussed remuneration schemes that might be used in terms of this tool, and devoted a lot of attention to the issue of trust. Though several researchers have tried to categorize the affiliates, their approaches were mainly based on general industry and marketing experience and did not include any calculations supporting the groups proposed.

All in all, application of Machine Learning tools into a real-life Aviasales case will allow us in terms of this thesis to contribute into the fund of existing research and introduce quality insights on the matter.

The next chapter will be devoted to the empirical study of the Aviasales affiliate marketing network data.

Chapter 2. EMPIRICAL PART

2.1. General research description and project plan

The aim of the empirical part of the project is to fulfill research and business objectives. Namely, it is important to find hidden data patterns as well as transform the data in accordance with Aviasales requirements.

Before diving into the details it is important to provide general outline of the project:

The data for the project is provided by Aviasales and its affiliate platform TravelPayouts. Two datasets are being used: first, the .pkl file with the main data on the affiliate urls (128 116 rows) and the second .xlsx file (303 223 rows) with data on which advertisers affiliates promote.

The main raw dataset contains 128 116 rows representing affiliates and includes solely the information on the affiliate website (Table 2):

Row	url	flag
5	bpponline.ru	direct advertiser
9	amondo.holiday	direct advertiser
12	akvaplan.com	direct advertiser
13	castrlnaurivierebasse.ft	direct advertiser
14	calnboard.ru	direct advertiser

Table 2 Example of the data in the initial dataset

(Source: compiled by authors based on Aviasales data)

Here, the ‘url’ column represents the website of the affiliate. ‘Flag’ columns represent the type of each website, which is an affiliate. It is important to note that though the data says ‘direct advertiser’ the affiliates are meant. In the data provided the company uses its internal classification from the viewpoint of the affiliate network owner that does not imply the term advertiser as merchandiser as was described in the literature review. Thus, in terms of this thesis main emphasis is made on the study of affiliates.

The main task is to carry out affiliates categorization in order to be able to answer research questions provided in the introduction and draw both academic and managerial insights on affiliate programs strategies.

It is important to mention that though the data is of a secondary type it is raw and, thus, demands preprocessing. Therefore, data preparation was the important initial step of the project. It included parsing, language detection, data cleansing and NLP preprocessing.

To solve the projected tasks it was first decided to use unsupervised learning, namely, K-means clustering to let the data speak for itself. Moreover, Principal component analysis (PCA) was also applied to handle the number of features in the data and simplify the analysis. Nevertheless, though the Russian language sites were showing clear patterns, English language sites were too bitty to form any sequence. Moreover, based on the acquired results Aviasales defined the priority categories needed to be discerned to fulfill business goals: content sites, cashback or promo codes sites and service sites. Thus, the project has switched to a supervised learning phase.

For a supervised learning the data had to be labelled. Moreover, a number of models had to be developed in order to achieve best F-score results. Thus, the models considered were Linear Support Vector classifier (Linear SVC), Gradient Boosting model and CatBoost model. The final model with which the prediction was made with CatBoost.

After the predictions have been received they have been merged with the .xlsx file for further analysis and interactive dashboard creation.

All in all, the plan of the project can be depicted as following (Figure 4):

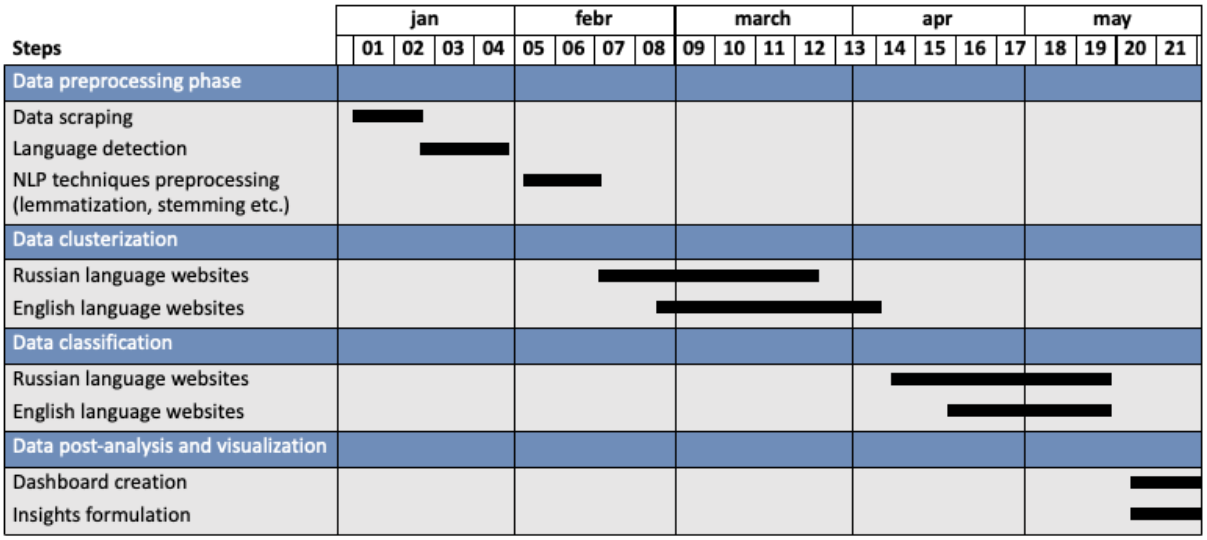


Figure 4 Plan of the empirical part fulfillment (Source: compiled by authors)

Thus, the empirical part of the thesis can be divided into three major parts: data clustering, data classification and results analytics.

2.2. Data clustering

The process of clusterization of the data provided by Aviasales can be broadly described as following (Figure 5):

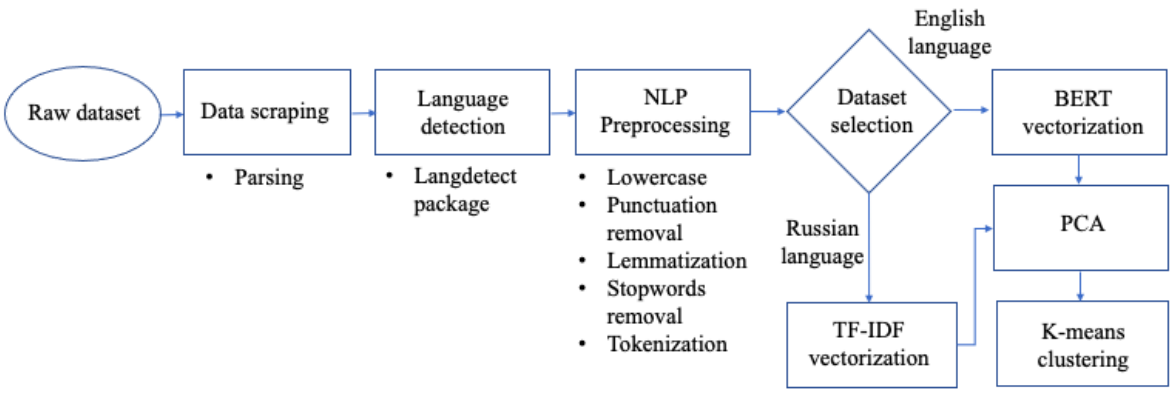


Figure 5 Data clustering process scheme (Source: compiled by authors)

2.2.1. Data preprocessing for clustering

Thus, the acquired raw data first has to be preprocessed. The data is basically non-numerical text data, which means that methods of natural language processing (NLP) are needed to get the insights.

NLP techniques are applied to a wide variety of tasks: content categorization, speech-to-text conversion, sentiment analysis, topic modelling, etc. In general terms, NLP tasks break down language into shorter, elemental pieces (i.e. tokens), try to understand relationships between the pieces and explore how the pieces work together to create meaning.

In terms of this thesis such aspect of NLP as data parsing plays a major role. The dataset representing all the affiliates firstly has to be parsed to be able to retrieve any sort of information later. By parsing the retrieval of the main keywords from the html is meant.

The process starts from the requesting data from the web page with the python module requests. Its primary function is to send HTTP requests with one command line. The need for this step arises from the nature of how http works. By itself http is a request-response protocol between a server and a client. This means that each time a client sends a request to a server to obtain information needed to access this site. That is the reason why despite this step seeming unimportant at the first glance, it however serves a vital function for the whole process.

The next step would be the information retrieval from the html itself using html2text package, which parses the string and directly gets the text information from the html code.

After those two steps, the dataset basically transformed from the table containing only the domain of a web site to the table containing also the main information that is published on that web source.

After this step the initial dataset would appear as following (Table 3):

Row	url	flag	text
5	bpponline.ru	direct advertiser	Access denied | bpponline.ru...
9	amondo.holiday	direct advertiser	Booking.com - Alles runf ums...
12	akvaplan.com	direct advertiser	Akvaplan-riva redirect Loading...Just a moment...
13	castrlnaurivierebasse.ft	direct advertiser	308 Permanent Redirect The...
14	calnboard.ru	direct advertiser	Создать бесплатный форум на MyBB..

Table 3 Dataset after parsing

(Source: compiled by authors based on Aviasales data)

The ‘text’ column contains the main message printed on the website, however, even judging by the first 5 sites, it can be seen that they all are in different languages. Thus, making it difficult to carry out the next step of the analysis.

The clusterization of all the datasets would be impossible due to the fact that the algorithm would just group datasets based on their language or at least language family they belong to, the division into smaller groups based on the language principle has to be carried out to be able to detect existing patterns.

To perform this the python langdetect library was used. The algorithm works on the basis of n-gram text categorisation.

N-grams simply means a parcing of a string, in this case a frase, in n characters. It usually implies any characters used in a word, however, this language detection especially across european languages is a more complicated subject. Thus, the n-gram approach takes pairs of characters found close to each other in the word.

The distribution of languages in the initial dataset is presented in Figure 6. Thus, the most popular affiliate sites languages are English, German and Russian.

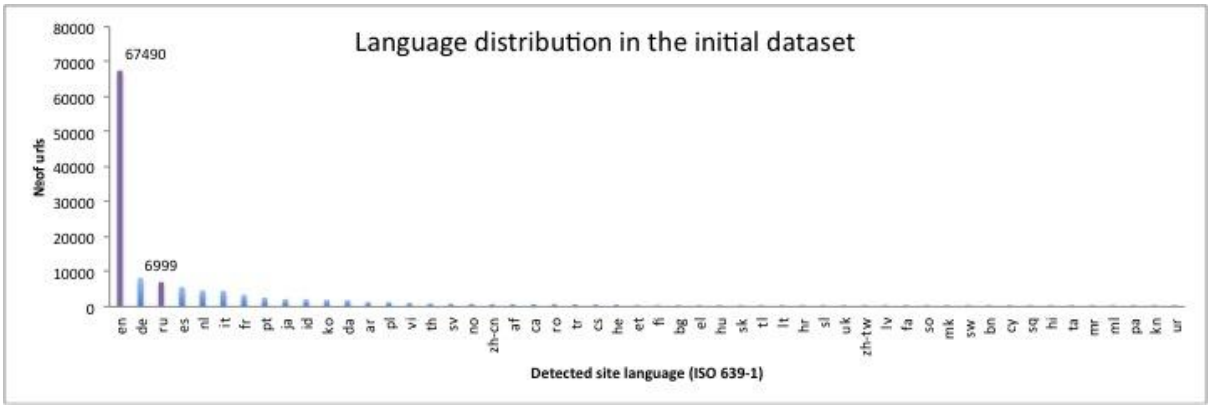


Figure 6 Language distribution in the initial dataset

(Source: compiled by authors based on Aviasales data)

After applying the langdetect package and creating a subsequent column to put a corresponding language to all the websites in the dataset, the result would look as following (Table 4):

Row	url	flag	text	language
5	bpponline.ru	direct advertiser	Access denied | bpponline.ru...	en
9	amondo.holiday	direct advertiser	Booking.com - Alles runf ums...	en
12	akvaplan.com	direct advertiser	Akvaplan-riva redirect Loading...Just a moment...	en
13	castrlnaurivierebasse.ft	direct advertiser	308 Permanent Redirect The...	en
14	calnboard.ru	direct advertiser	Создать бесплатный форум на MyBB..	ru

Table 4 Dataset after langdetect package application

(Source: compiled by authors based on Aviasales data)

To fulfill the business goals of the thesis English and Russian websites are selected for the further study.

After execution of these steps the following challenges were discovered:

- The data contained sites that were not available or deleted. This could present challenges for the unsupervised learning part of the research as the noise could interfere with the clustering process. Thus, for this purpose data cleansing process had to be done.
- Though the langdetect package represents one of the best solutions in the language detection process it does not give irreproachable results. Therefore, websites with domains that belong to the non-english-speaking countries were removed from the data.

Moreover, it is important to note that the text of the websites can not be used as a whole. It needs to be derived to a standard form and divided into separate words. For these purposes nltk package and re module are to be used.

nltk package is an open source collection of libraries that allow multiple text data manipulation as well as data visualisation. re module, namely, re.sub command allows to replace the data according to certain conditions. Therefore the initial text went to the following stages (Table 5):

Process	Process description	Initial text	Results
Data transformation into lowercase	Replacing the upper-case symbols with lowercase	Travelmart: заказ авиабилетов, доставка авиабилетов вам	travelmart: заказ авиабилетов, доставка авиабилетов вам
Punctuation removal	Removal of all punctuation signs: ‘,’, ‘:’, ‘!’ etc	travelmart: заказ авиабилетов, доставка авиабилетов вам	travelmart заказ авиабилетов доставка авиабилетов вам
Lemmatization	Carrying out morphological analysis and finding the lemma (base form) of the word	travelmart заказ авиабилетов доставка авиабилетов вам	travelmart заказ авиабилет доставка авиабилет вы
Stopwords removal	Removal of the most common words that do not contribute into the analysis. Russian language examples: ‘я’, ‘где’, ‘все’	travelmart заказ авиабилет доставка авиабилет вы	travelmart заказ авиабилет доставка авиабилет

Table 5 NLP Russian language websites data preprocessing

(Source: compiled by authors based on Aviasales data)

Also, important to note that among the peculiarities of the Russian language is the usage of the letter ‘ё’, which sometimes appear in text. In order to standardize the tokens studied ‘ё’ is replaced by ‘е’. Moreover, as Russian language websites were further vectorized with TF-IDF method custom tokenization was omitted.

As a result of the application of manipulations described Russian language dataset acquire a new column ‘proc’ (processed) with processed text (Table 6):

Raw	URL	Flag	Text	Language	Proc
0	flyticket.ru	direct advertiser	Travelmart: заказ авиабилетов, доставка авиабилетов	ru	travelmart заказ авиабилет доставка авиабилет
1	vizitkaplus.ru	direct advertiser	Срок регистрации домена закончился.	ru	срок регистрация домен закончиться купить домен
2	airport63.ru	direct advertiser	Выбрать маршрут: Москва-Бангкок, Москва-Пхукет	ru	выбрать маршрут москва бангкок москва пхукет
3	na-more.su	direct advertiser	Отдых на Черном море. Курорты, города и страны	ru	отдых черное море курорт город страна
4	krabiguiding.com	direct advertiser	Krabi SJ travel - главная	ru	krabi sj travel главный

Table 6 Russian language dataset after NLP preprocessing

(Source: compiled by authors based on Aviasales data)

In the case of English websites additionally BertTokenizer was applied. The idea of the tokenizer is to divide a sentence into separate words (tokens). The manipulation is usually done through detection of blank spaces.

Therefore, for English websites NLP preprocessing looked as follows (Table 7):

Process	Process description	Initial text	Results
Data transformation into lowercase	Replacing the upper-case symbols with lowercase	Religious Tourism Bergoglium - 2016 - 2017 languages	religious tourism bergoglium - 2016 - 2017 languages
Punctuation removal	Removal of all punctuation signs: ‘,’, ‘:’, ‘!’ etc	religious tourism bergoglium - 2016 - 2017 languages	religious tourism bergoglium 2016 2017 languages
Lemmatization	Carrying out morphological analysis and finding the lemma (base form) of the word	religious tourism bergoglium 2016 2017 languages	religious tourism bergoglium 2016 2017 language
Stopwords and noise removal	Removal of the most common words that do not contribute into the analysis. English language examples: ‘I’, ‘about’, ‘after’. Removal of in-text numbers	religious tourism bergoglium 2016 2017 language	religious tourism bergoglium language
Data tokenization	Tokenizer divides a string into a substring, thus the sentence is transformed into separate words	religious tourism bergoglium language	‘religious’ ‘tourism’ ‘bergoglium’ ‘language’

Table 7 NLP English language websites data preprocessing

(Source: compiled by authors based on Aviasales data)

Moreover, as the English language data contained almost ten times more sites than the Russian language dataset, the amount of noise in the form of broken links could significantly confuse the model. Therefore, the attempt to clear the dataset was also taken: broken links were identified based on the words commonly found on such sites: ‘access denied’, ‘redirect’, ‘error’, ‘no longer exists’ and removed from the data.

Nevertheless, the final dataset looked similar to the Table 6 for Russian datasets presented above and the column ‘proc’ was also added (Table 8):

Raw	URL	Flag	Text	Language	Proc
0	bergoglium.com	direct advertiser	Religious Tourism Bergoglium - 2016 - 2017 languages	en	‘religious’ ‘tourism’ ‘bergoglium’ ‘language’
1	irelandgolfer.com	direct advertiser	Ireland Golf - Ireland Golf Courses	en	‘ireland’ ‘golf’ ‘ireland’ ‘golf’ ‘courses’ ‘directory’
2	travelsis.net	direct advertiser	Traveling Sisters	en	‘traveling’ ‘sisters’ ‘skip’ ‘content’
3	compathy.net	direct advertiser	Compathy - Travel Collections around the world	en	‘compathy’ ‘travel’ ‘collections’ ‘around’ ‘world’
4	amalficoastonline.eu	direct advertiser	Amalfi Coast - A paradise on Earth - EXPERT GUIDE	en	‘amalfi’ ‘coast’ ‘paradise’ ‘earth’ ‘expert’ ‘guide’

Table 8 NLP English language dataset after NLP preprocessing

(Source: compiled by authors based on Aviasales data)

2.2.2. Data vectorization

The next important step is data vectorization. Vectorization is needed to solve the non-numerical nature of the data problem as natural language is basically a text. Vectorization algorithms assign value to certain words and, thus, allow to evaluate their importance.

The data for vectorization significantly varies in volume. Thus, websites containing information in Russian language constitute about 7 thousand rows, while the English part of the data set is equal to approximately 70 000 instances (67 490 rows). Therefore, different approaches were chosen: for the Russian language websites TF-IDF vectorization algorithm was applied, while for English language websites BERT was used.

Russian websites vectorization approach

TF-IDF vectorization algorithm is among the most popular and widely used vectorization approaches. It allows to transform text into numerical form through evaluation of the word's importance for the whole text. The results of TF-IDF transformation can be fed to algorithms such as Naive Bayes, Support Vector Machines, clustering. The advantage of TF-IDF is that in relation to basic methods like word counts the results are significantly improved (Monkeylearn, 2021).

The math behind the approach is multiplication of the term frequency (TF) and the inverse document frequency (IDF). Term frequency reflects how many times the word is met in the document. The inverse document frequency represents how many documents with this word are in the collection. In other words it is the importance of the word in relation to the whole corpus (Monkeylearn, 2021).

Mathematically the TF-IDF algorithm of a word ‘w’ in a text ‘t’ over all of the text from all websites ‘T’ is presented by the formula (Formula 1):

$$tf\,idf\,(w,\,t,\,T)\;=\;tf\,(w,\,t)\;*\;idf\,(w,\,T)$$

where:

w – a word,

t – a text from one website,

T – text from all websites,

tf – the word frequency in the text from one row:

$$tf\,(w,\,t)\;=\;log\,(1\,+\;freq\,(\,w,\,t)$$

idf – the inverse frequency of the word all across the whole text from all the row:

$$idf\,(\,w,\,T)\;=\;log\,(\frac{N}{count\,(\,w\;\in T{:}w\in t)})$$

Formula 1 TF-IDF Calculation

Thus, the algorithm has been applied to Aviasales data ‘proc’ column. Based on the unique meaning of the word and their weight assigned via TF-IDF method the following matrix is obtained (Table 7):

Raw	URL	Flag	Text	Language	Proc	акция	бизнес	билет
0	flyticket.ru	direct advertiser	Travelmart: заказ...	ru	‘travelmart’, ‘заказ’, ...	0	0.088393	0
1	vizitkaplus.ru	direct advertiser	Срок регистрации...	ru	‘срок’ ‘регистрация’...	0	0	0
2	airport63.ru	direct advertiser	Выбрать маршрут...	ru	‘выбрать’ ‘маршрут’ ...	0.028735	0.025992	0.047775

Table 9 Russian language dataset after TF-IDF vectorization (extract)

(Source: compiled by authors based on Aviasales data)

The dimensions of the dataset acquired are 6 rows and 29 026 columns meaning that the text was divided into 29 019 features

English websites vectorization approach

Although the data preprocessing of the English websites in essence appears to share the majority of the steps with that of the Russian language websites, there is one notable difference that has to be addressed. Since the number of the English websites is considerably higher than the number of the Russian websites, substantially more computation power and time is needed. Moreover, the attempt to apply TF-IDF vectorization used in the case of Russian websites did not bring any discernible results. Thus, the data demanded a different approach.

Among possible vectorization techniques word embedding or word2vec was considered. Word embedding methods learn a real-valued vector representation for a predefined fixed sized vocabulary from a corpus of text. The learning process is either joint with the neural network model on some task, such as document classification, or is an unsupervised process, using document statistics (Brownlee, 2019). word2vec considers the

word in the contact and, thus, presents a powerful tool, however prone to overlearning (Hazoom, 2018).

Taking into account the fact that vectorization techniques can significantly influence the result of further clustering and the complex nature of the data it was decided to appeal to the latest and most advanced tools, namely, Bidirectional encoder from transformers (BERT) model. According to OpenAI (2018) study a neural network trained on Transformer architecture shows higher final results than word2vec and other vectorization methods. Based on these findings BERT neural network was created by Google in 2018 (vc.ru, 2020). BERT model is pre-trained on a combination of sentences with around 15% of masked or hidden from network words. The network not only evaluates which words suit the proposed context but also takes into account the next sentence and decides whether they are connected or not. These processes are scientifically called masked language modelling (MLM) and next sentence prediction (NSP). (BERT documentation, 2021).

Thus, the specific pretraining algorithm allows BERT to achieve high results with relative method simplicity and hold the status of state-of-the-art-model.

As a result of BERT vectorization applied to English language websites 768 features were obtained and the English language was transformed into following table (Table 8):

Raw	URL	Flag	Text	Language	Proc	0	1	2
0	bergoglium.com	direct advertiser	Religious Tourism Bergoglium - 2016 - 2017 languages	en	‘religious’ ‘tourism’ ‘bergoglium’ ‘language’	0.229134	-0.212548	0.818740
1	irelandgolfer.com	direct advertiser	Ireland Golf - Ireland Golf Courses	en	‘ireland’ ‘golf’ ‘ireland’ ‘golf’ ‘courses’ ‘directory’	0.119551	0.104040	0.857746
2	travelsis.net	direct advertiser	Traveling Sisters	en	‘traveling’ ‘sisters’ ‘skip’ ‘content’	0.275895	-0.272740	0.826218

Table 10 English language dataset after BERT vectorization (extract)

(Source: compiled by authors based on Aviasales data)

Therefore, as a result of the steps described both Russian and English language datasets are now prepared for clusterization. However, the number of features is significant and, thus, dimensions need to be decreased. Thus, it was decided to apply principal component analysis (PCA). Before diving into the details it is necessary to present the theory of both these methods.

2.2.3. General introduction to PCA and clustering method

Principal component analysis (PCA)

Principal component analysis is the commonly used preprocessing step before clusterization, classification, segmentation etc. As PCA is a dimensionality reduction approach with the help of PCA a large dataset can be transformed into a smaller one, however, saving the most important information. Nevertheless, it is vital to note that some authors, for example Harrington (2012), point out that useful information can still be lost, however traded for simplicity.

More precisely, PCA provides decrease in dimensionality by projecting the data onto linear subspace so that the least squares approximation is maximizing the variance of the projection coordinates (Neumayer et. al, 2019). Therefore, the method searches for such planes and lines in the K-dimensional space that present the closest fit to the data explored (Jolliffe and Jackson, 1993).

The process of PCA can be partially explained with the help of the following representation (Figure 7):

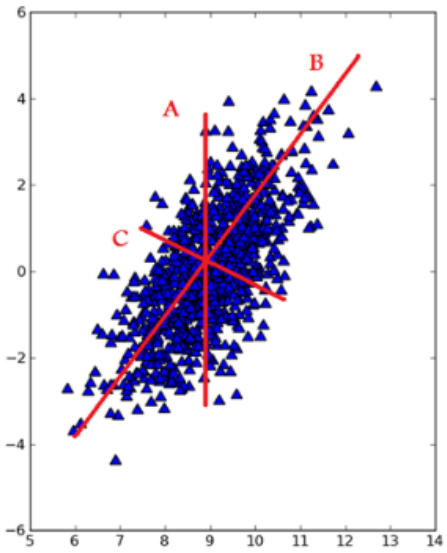


Figure 7 Axes rotation in PCA algorithm (Source: Harrington, 2012)

In fact, PCA rotates the axes of the data based on the nature of this data. The larger the line the more variation is explained and the more important this part of data is. Thus, in the figure line B is the axis covering the largest amount of variability. After that the orthogonal to the first line is found, thus, line C and so on.

Important step of the PCA method is data standardization: the range of the variables have to be derived to similar terms to equal their contribution (Jaadi, 2021). The next step is covariance matrix computation that reflects how the variables vary from the mean with respect to each other. After that eigenvectors are calculated as eigenvector analysis on covariance matrix allows to find the structure of the most important features i.e. determine principal components (Harrington, 2012).

PCA belongs to feature extraction methods, therefore, principal components are new variables constructed on the base of initial variables. Thus, as a result of PCA application input variables are combined in such a way that new independent variables are created (Brems, 2017). Principal components are uncorrelated and usually the first component explains the most variance as it represents the largest number of initial variables. Therefore, the significance of each next component decreases. Thus, a feature vector, a vector that combines all the most valuable components is defined (Jaadi, 2021).

All in all, PCA is a useful tool that allows the simplification of data analysis through reduction of initial data dimensions and initial variables transformations. Important to note that because of such manipulations principal component meaning can be quite hard to explain.

Clustering

Generally put, clustering is one of the many unsupervised learning algorithms used to define and unite groups with the same or similar patterns and to differentiate them from groups with different characteristics. Clustering is the assignment of putting the population or data points into various gatherings or groups to such an extent that data points in similar groups are more like other data points in the similar group and unlike the data in different gatherings. It is fundamentally a collection of instances based on closeness and divergence between them.

It is important to mention that the clustering approach is the form of an unsupervised algorithm. Unsupervised learning means the usage and analysis of the unlabeled data.

The general process of an unsupervised algorithm looks like (Figure 8):

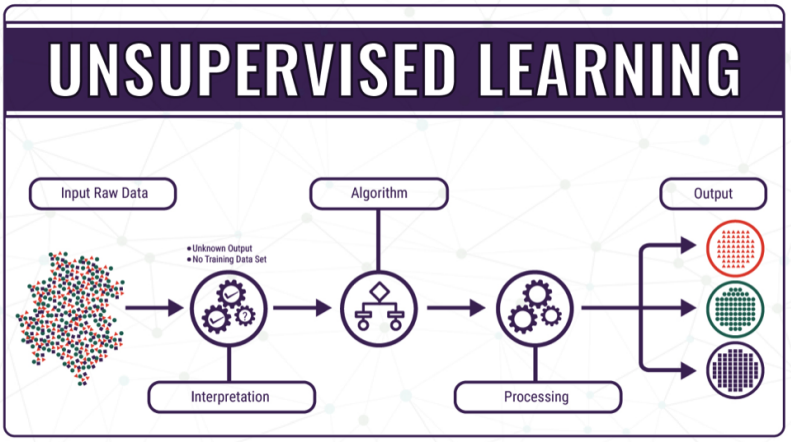


Figure 8 The process of unsupervised learning (Source: Calltouch, 2021)

The aim of such an approach is to let the computer find the underlying patterns without any intervention. Apart from clustering among unsupervised techniques are association and dimensionality reduction (Delua, 2021). Thus, unsupervised learning will allow to get important insight that will contribute to further general managerial conclusions.

The clusterization algorithm chosen for the analysis is K-means clustering. The reasons for choosing such an approach is its relative simplicity, however high resultivity. Moreover, the method is fast and included in a number of Python packages (Henrique, 2019).

K-means clustering divides n observations into k clusters where each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster.

Important note to take into account is that the number of clusters has to be introduced manually, and is fixed. Thus, the researcher defines a target number k, which refers to the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the center of the cluster. Each data instance is put into each of the clusters through reducing the in-cluster sum of squares (‘K-Means Clustering Explained’, 2020).

In other words, the K-means algorithm identifies the k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible (‘Predicting website categories’, 2021). The ‘means’ in the K-means refers to averaging of the data; that is, finding the centroid, and a data point is clustered into a certain cluster based on how close the features are to the centroid (‘K-means Clustering Algorithm’, 2021).

The difference between centroids and clusters themselves are calculated according to the Euclidean distances.

If p and q are points in R3, the Euclidean distance from p to q is the number, which can be presented mathematically as Formula 2:

$$d\,(p, q)\ = ||\,p\ -\ q||$$

Formula 2 Euclidean distance calculation in R3 (Source: sebastianraschka, n.d.)

In a field of Machine Learning Euclidean distances are just a measure of difference between two or more different groups of data, and represented by a following formula (Formula 3):

$$\sqrt{\sum_{i=1}^{n} (\,q^{i}\ -\ p^{i}\,)^{2}}$$

Formula 3 Euclidean distance in Machine Learning (Source: sebastianraschka, n.d.)

Geometrically in a two dimensional space this can be represented by a simple straight line.

2.2.4. Results of data clustering

Due to the initial affiliates' dataset being divided into two major subsets, – the one containing only Russian websites and the one containing only English websites, the results are also represented in two sets.

Results of the clusterization of Russian websites

As the transformed dataset included 29 019 features PCA was applied to reduce the dimensionality of the data. 16 components explaining 50% of variance were implemented.

The principal components' individual and cumulative variance are presented in the following graphs (Figure 9):

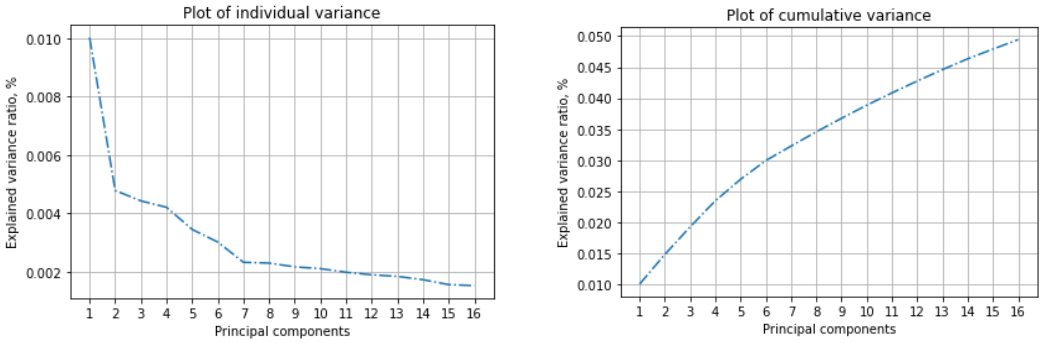


Figure 9 Plots of principal components' individual and cumulative variance in the Russian language dataset

(Source: compiled by authors based on Aviasales data)

Thus, now it is possible to proceed with the K-means clustering itself.

One peculiar aspect of K-means clustering that has to be addressed is that the number of clusters is introduced manually. In terms of this paper, the iteration was performed on the range from 0 to 40 possible clusters. The maximum number was selected based on common sense. Thus, too many clusters makes further analysis and description confusing. Moreover, a large number of clusters leads to emergence of cases with solely one or two websites.

The selection criterion conventionally used for the final number of clusters definition is the elbow curve. The method consists of illustrating distortions in one graph, which essentially is a sum of the squared differences of each datapoint from the centre. The distance is calculated according to the Euclidean principle. However, it is possible to choose from the wide range of available computation options such as Manhattan distance and Pearson, Spearman or Kendall correlation-based measurements.

Therefore, at each iteration the average distortions as well as inertia were computed for each number of clusters in a set range. The overall results look like following (Figure 10):

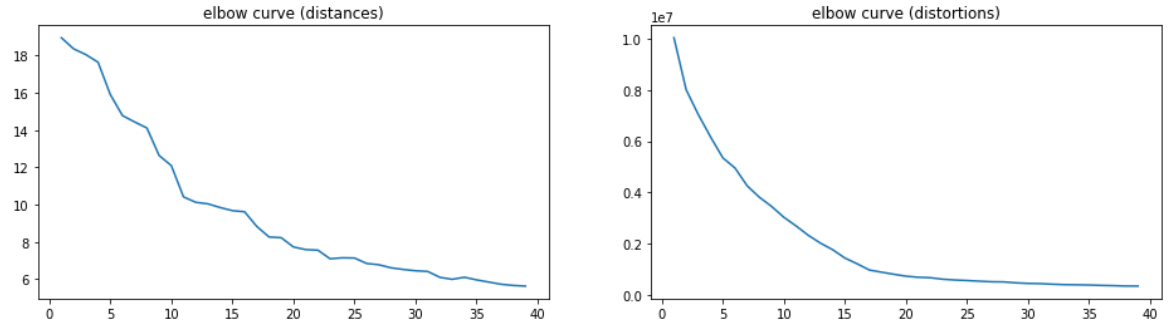


Figure 10 Elbow curve for Russian language websites

(Source: compiled by authors based on Aviasales data)

Thus, based on the data the decision was made to break the dataset into 4 clusters. They can be summarized as follows (Table 11):

Cluster №	Cluster name	General description	№ of websites	Key words examples	Website examples
1	Travel-themed websites	Various travel themed websites from hotels and resorts to travel agencies and airports	6762	авиабилет, москва, билет, тур, отель	cologne-tour.de (Cologne excursions), azimuth.aero (airline)
2	Affiliate and partnership programs	Various marketing and partnership programs	8	маркетинг, партнерский, оффер, рекламодатель	storader.com (partnership program)
3	Real estate	Buying and selling real estate abroad	7	аренда, вилла, недвижимость, франция	comodo.ru (real estate agency)
4	City portals	Cities sites with message boards and forums	213	россия, вакансия, новость, резюме	noyabrck.ru (Noyabrsk city site)

Table 11 Description of Russian language clusters

(Source: compiled by authors based on Aviasales data)

The projection of the clusters on 2-dimensional space can be presented as Figure 11 Important to note that as Python numeration starts from 0 the classes presented in Figure 11 are the following: 0 = class 1: travel-themed websites, 1 = class 2: affiliate and partnership programs, 2 = class 3: real estate, 3 = class 4: city portals.

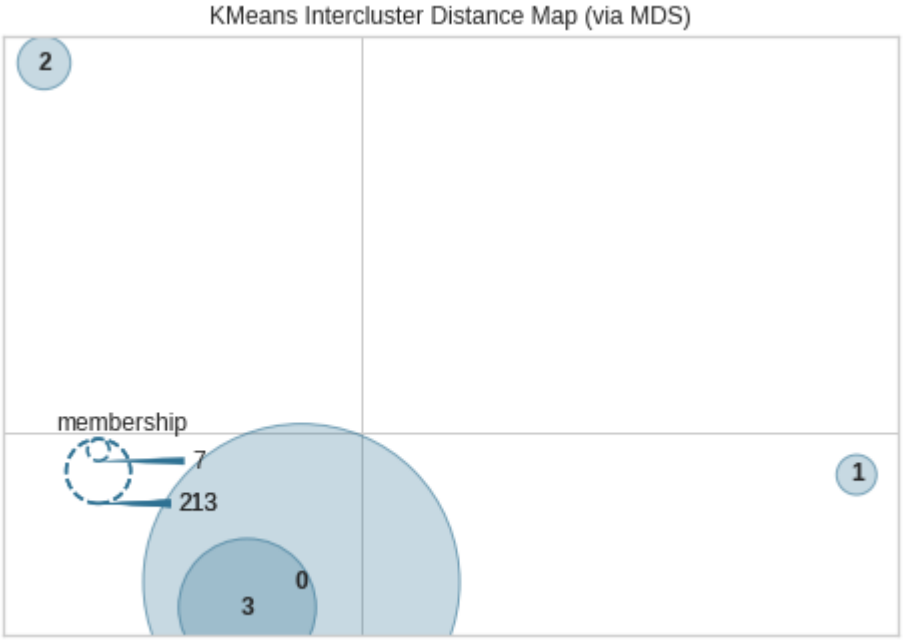


Figure 11 Russian websites intercluster distance map
(Source: compiled by authors based on Aviasales data)

The most common frequently found in each cluster words can be presented as following Wordclouds (Figure 12):

Figure 12 Russian website clusters Wordclouds

(Source: compiled by authors based on Aviasales data)

It can be seen from the general outlook of the received clusters that there is an obvious unevenness in the sizes of the clusters. Moreover, Wordclouds show that though the general message is known noise still interferes with the data.

The model mostly attributed all the observations to the cluster 1, namely, it consists of 6 762 websites out of 6 999. The problem can be perceived as the following: the words used across travel-themed sites are similar as words like ‘vacation’, for example, can appear both on travel agency websites, specific hotel websites as well as resorts websites or travel blogs. In terms of unsupervised learning the machine was not able to dive deeper and discern such thin differences.

Clusters 2 and 3 contain dangerously small amount of websites. Moreover, some of these websites are clones presented in different domains. Nevertheless, partnership marketing network and real estate cites were discerned.

The remaining cluster is a collection of somewhat old-fashioned city portals that include forums, newsboard, as well as parts of the site, where locals can buy or sell various stuff. It partially overlaps with cluster 1.

All in all, the main problem for further research mostly lies in the inability of the model to further analyse cluster 1 and divide it into several more sub-clusters with more concrete meaning.

Results of the clusterization of the English websites

For the English dataset the starting number of components was also considered to be 16. However, in the case of English websites the individually explained variance of the component drops to 0.021 at 6th component, while the total variance explained by the first 5 components is almost 75%. These can be seen on the following graphs (Figure 13):

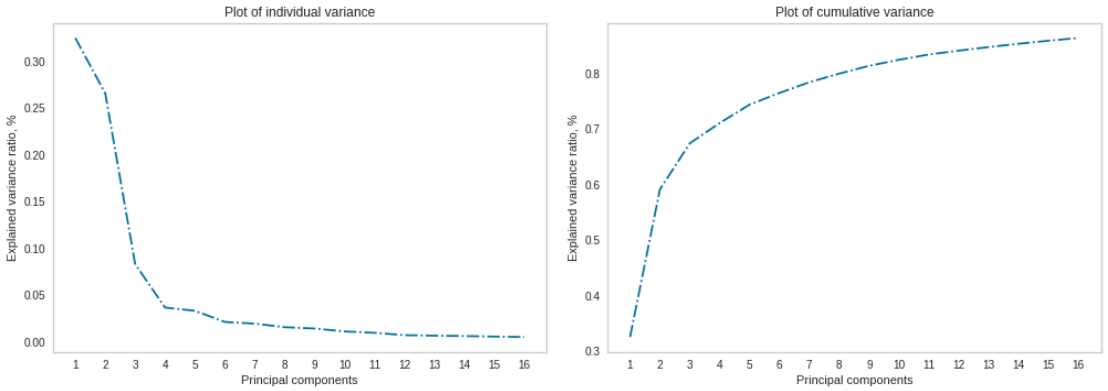


Figure 13 Plots of principal components' individual and cumulative variance in the English language dataset

(Source: compiled by authors based on Aviasales data)

Thus, the decision was made to decrease the number of components to 5.

In the process of the clustering itself similar to Russian websites the iteration is made on the range of 0 to 40 clusters. The elbow curve based on the dataset is presented on Figure 14:

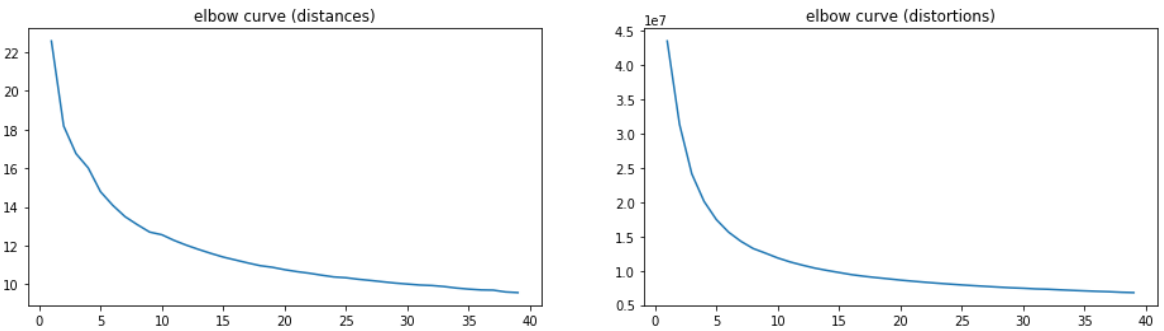


Figure 14 Elbow curve for English language websites

(Source: compiled by authors based on Aviasales data)

Though the elbow is hard to discern it appears to be between the 8th and 10th cluster. Taking into account the large number of variables in the data the appropriate number of clusters was considered to be 10.

In 2-dimensional space the clusters can be presented as following (Figure 15):

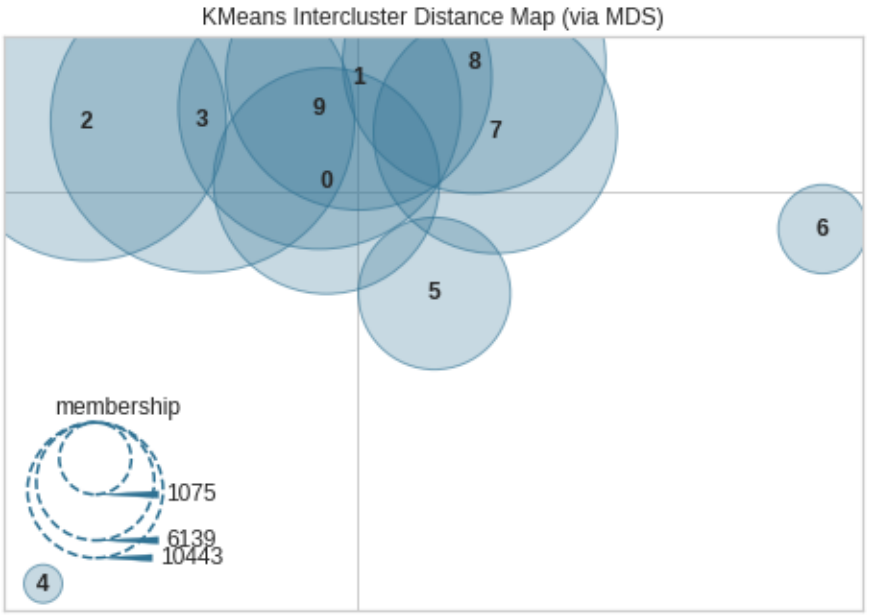


Figure 15 English websites intercluster distance map

(Source: compiled by authors based on Aviasales data)

On the graph it is seen that the main amount of clusters intersect and, thus, they are hard to discern. Their general description is presented in the following table (Table 12):

Cluster №	№ of websites	Key words examples	Website examples
1	4595	travel, hotels, home, world	homeyhostel.com (hostel), earpdivisionexpo.com (exposition)
2	8715	travel, hotels, world, city, book	eternal-guild.com on-line game), geographerwilltravel.blogspot.com (travel blog)
3	10443	travel, home, world, city, hotels, blog	qatarhandball2015.com (handball tournament), dutchlifesciences.com (conference)

Table 12 (continuation)

Cluster №	№ of websites	Key words examples	Website examples
4	14783	travel, hotels, home, world, blog, news	scam-rescue.ca (investment scams information), munichimprovtheater.com (improvisation theater)
5	247	media, public,destination, posts, comments	todayallcoupons.com (coupons), bullybuddyzone.com (boutique)
6	1075	travel, booking, search, best, price	astigmuseum.blogspot.com (personal blog), innovabaltictour.eu (golf tournament)
7	320	facebook, twitter, pinterest	sister.travel (travel in Asia)
8	6139	travel, hotel, best, free	cappadociaonline.com (Cappadocia travel), farscapeworld.com (TV series fan site)
9	8321	travel, hotels, dating, city	magicweek.co.uk (magazine about magic), bomadg.in (personal travel blog)
10	11067	travel, new, best, hotels, home	hotelroomdirectory.com (hotel booking site), phillycollector.blogspot.com (personal blog)

Table 12 Description of the English language clusters

(Source: compiled by authors based on Aviasales data)

Thus, it can be seen that although the clusterization results of the English websites do not share the same unevenness problem as the Russian ones, the distribution of the websites is unclear. The most used words intersect. For example, the words ‘travel’, ‘hotels’, ‘home’

are attributed to almost all clusters. This leads to various-themed sites being united as one cluster, though such approach does not make sense in the viewpoint of a human mind. As a result, it is impossible to clearly define clusters. Overall, no distinct hidden patterns were found from the data.

Furthermore, another problem was discovered. Though the attempt to clean the data from wrongfully detected languages by domain search was made, it turned out that mixed language sites remained in the dataset. For example, sister.travel website from cluster 7 represents an Asian website, which uses a lot of English words like ‘home’, ‘world’, ‘life’ and etc. Due to the large number of websites in the dataset manual check of each website is mot a feasible idea. Thus, a certain amount of noise remains in the dataset and possibly interferes with the clustering process.

Overall results of the clusterization

Analyzing the results of applying algorithms to both datasets it can be seen that although some clusters are formed, they possess lack of consistency, as well as general integrity. Therefore, achieved outcomes are not suitable for deriving valuable and applicable conclusions. Clusterization of both Russian and English language clusters only showed the existence of a large bulk of travel-themed sites with similar keywords that are hard for the machine to discern. Moreover, English language websites preprocessing revealed a high amount of noise in the data, which implies bad quality of the affiliate themselves. Namely, many of them either no longer exist or are forgotten by their creators. Thus, it is natural to review a part of both languages websites manually and switch to supervise learning.

Moreover, Aviasales expressed the wish to switch to the supervised Machine Learning (ML) approaches and identify particular classes of the websites for further analysis. The company’s desired classification was the following: ‘service’, ‘content’ and ‘promo codes or cashback services’.

2.3. Data classification

The next step of the affiliate analysis was a switch for supervised learning methods.This approach allows carry out the analysis in accordance with Avisales’ interest and strategic vision.

Supervised learning is essentially a more generally used type of ML algorithms due to its perceived capability of producing more reliable results. Among supervised learning techniques are regression algorithms and classification tools can be named (Delua, 2021).

The main difference between supervised and unsupervised learning is that the latter uses the data that has previously been labelled by a human. Thus, basically, a supervisor (i.e. a human) manually sets a number of ground rules of finding a solution to a problem and then the algorithm trained on these rules learns how to perform without the help of the supervisor.

Putting this in mathematical terms, supervised learning is a type of algorithms that uses an input X and an output Y following the function (Formula 4):

$$Y = f(X)$$

Formula 4 Function used by the supervised algorithm (Source: Delua, 2021)

The main goal is to approximate a function f in such a way that by introducing a new set of Xs, the function would predict the output or a set of outputs Y.

The classification process in this paper is based on the Gradient Boosting family of algorithms. This choice can be explained by the fact that the Gradient Boosting algorithms by its ability to process large quantities of different types of data (images, words, and quantitative data) with relatively large speed and high accuracy.

2.3.1. Data preprocessing for classification

As Aviasales predefined the classes needed for achievement of business goals, multiclass classification was considered to be the suitable method for further analysis.

Thus, initially three classes were defined based on the insights from the company's advisor and the general understanding of the industry:

1. Content sites – sites that do not sell any goods or services, however contain information, description and narrative in the form of posts or plain text. Examples of such sites include travel guides or news sites.
2. Service sites – sites selling goods or services, on which description or additional narrative is minimized and the main emphasis is made on the offer. Examples of this category are travel agencies.
3. Cashbacks and promo codes – sites containing offers on discounts and cashbacks

There were sites that did not fall into any of the defined categories. Thus, an additional class 'other' was introduced. Moreover, due to the fact that on both datasets,

Russian and English, the langdetect package did not manage to perfectly determine the language another category was added ‘error’. The class would represent a website in a language that is not either Russian or English. It is also important to note that for the supervised part it was decided to leave broken links in the data to train the machine to discern them as a separate group.

Thus, finally, the following five classes were created: content sites, cashback or promo codes sites, service sites, error and not available.

For the Russian language dataset 1100 sites (16% of the dataset) were labelled manually, while for the English sites the number was 2119 (3%). Together with the company the decision was made to limit manually classified sites to the numbers above as the models were not improving significantly after the increase of the number of labelled sites.

At this stage the following problems were faced:

The data was parsed in the same way as in the clustering process described above. However, as the data contained not available and wrongfully detected language sites, it led to the appearance of cases, when the parsing algorithm was not able to extract any information from the affiliate sites. The problem was observed in English language dataset, nevertheless, the Russian dataset was also checked for missing values. Thus, for example, the English language dataset looked in the following way (Table 13):

Row	Site url	Class	Site text
0	https://hktravelers.blogspot.com	content	Backpackers Forum Pakistan
1	https://datingrelashionshipsandmarriage.blogspot.com	content	Dating Relationships Marriage Dating
2	http://dramanauskaite.blogspot.com	content	ESP for Hotel and Catering Industry
3	https://italyvacationpackages.it	error	NaN
4	https://pinkiesandparadise.com	not available	NaN

Table 13 English classification dataset after parsing

(Source: compiled by authors based on Aviasales data)

Missing values were dropped from the dataset. Thus, it was transformed in the following way (Table 14):

Row	Site url	Class	Site text
0	https://hktravelers.blogspot.com	content	Backpackers Forum Pakistan
1	https://datingrelashionshipsandmarriage.blogspot.com	content	Dating Relationships Marriage Dating
2	http://dramanauskaite.blogspot.com	content	ESP for Hotel and Catering Industry
6	https://beautifulreortszone.blogspot.com	content	Beautiful Resorts Zone
9	https://cheap-pleane-tickets-studentsblogspot.com	service	Cheap Plane Tickets Students skip to main

Table 14 English classification dataset after missing values drop

(Source: compiled by authors based on Aviasales data)

Another faced difficulty was that a number of affiliates went down during the working process, interfering with the data labelling step. For example, if the site was working at the time of the manual labelling process it was not attributed to the group ‘not available’. However, during the execution of the model itself it was parsed as the empty site, confusing the model. To solve the problem the decision was made to use the previously parsed data provided by Aviasales.

To perform the model the data was split into training and test set, the distribution of classes among dataframes was the following (Table 15):

Class name	Initial distribution of classes across the dataframe	Distribution of classes across the dataframe after NaN drop	Distribution of classes in the train set	Distribution of classes in the test set
Russian language websites				
Content	572	572	474	98
Service	272	272	202	70
Cashback/promo codes	6	6	4	2
Not available	114	114	91	23
Other	136	136	109	27
Error	0	0	0	0
English language websites				
Content	854	674	535	139
Service	473	264	212	52
Cashback/promo codes	73	51	44	7
Not available	298	31	27	4
Error	332	233	181	52
Other	89	61	52	9

Table 15 Distribution of classes throughout the dataset, test set and train set

(Source: compiled by authors based on Aviasales data)

The biggest proportion in the datasets belonged to the class ‘content/not content’ while ‘service/not service’ and ‘error’ classes were runner-ups.

Important to note that for classification purposes a new class ‘service 1’ was introduced to the models. These were the sites that redirected straight to Aviasales and were considered by the company itself as service sites. Due to their different nature they were attributed to their own class in order not to confuse the model.

2.3.2. Classification models and results of their application

To achieve more consistent and reliable results several models were constructed and applied.

The chosen family of methods was gradient boosting classifier. The term 'gradient' refers to taking a derivative of the same function several times. In turn that reflects the process of how gradient descent algorithm works: the algorithm iteratively takes a certain loss function to its minimum.

Typically, all gradient-boosting algorithms include 3 elements (Gandhi, 2018):

1. Loss function – a function that represents some sort of 'punishment' or 'sanctions' for the incorrect classifications made by the model.
2. Weak learner – a 'weak' classification algorithm that will be used by the model as a starting point. Basically, each subsequent model that is added to an existing one will 'correct' the residuals in the predictions and make the whole model stronger.
3. Additive model – To each new model be able to 'correct' the existing one, it should be able to reduce the loss. It is done by parameterizing the model, then modifying those parameters so they would improve the model. Then the output is added to previous outputs.

The process stops at either a fixed number of added models or when the loss reaches its minimum value.

Model 1: Linear Support Vector classifier (SVC)

The starting model for the analysis of the websites was Linear Support Vector classifier. The main idea of this algorithm is to construct a hyperlane (i.e. a lane existing in 4 and more dimensions) in N-dimensional space to classify the text data (Gandhi, 2018).

A hyperplane is a lane that separates one class of data from another. Thus, the classification process resembles that of logistic regression, the only difference is that in this case if the output layer result is larger than 1, it belongs to a class 1, and in case the result is less than 1, it belongs to a class 0.

As in any Gradient Boosting algorithm model, the cost function is constructed and the final results are achieved by sequential changes in weights after taking partial derivatives in respect to those weights.

The SVC model was applied with the default parameters (Table 16):

Parameter name	Parameter meaning	Chosen value
C	Penalty for the error term	1
Loss	Los function	Squared hinge loss
Penalty	Penalization norm	Standard
Dual	Solving dual or primal optimization problem	True
Tol	Tolerance for stopping	1e-4
Multi_class	Determination of multiclass strategy	One-vs-rest
Fit_intercept	Calculation of the intercept	True
Intercept_scaling	Regularization parameter	1
Class_weight	Parameter that allows to attribute more weight to a particular class	1
scale_C	Parameter that makes C independent of the number of samples	True

Table 16 Parameters of the applied SVC model

(Source: scikit-learn documentation, 2011)

Thus, the model was applied to both Russian and English datasets and the following classification reports were obtained (Figure 16):

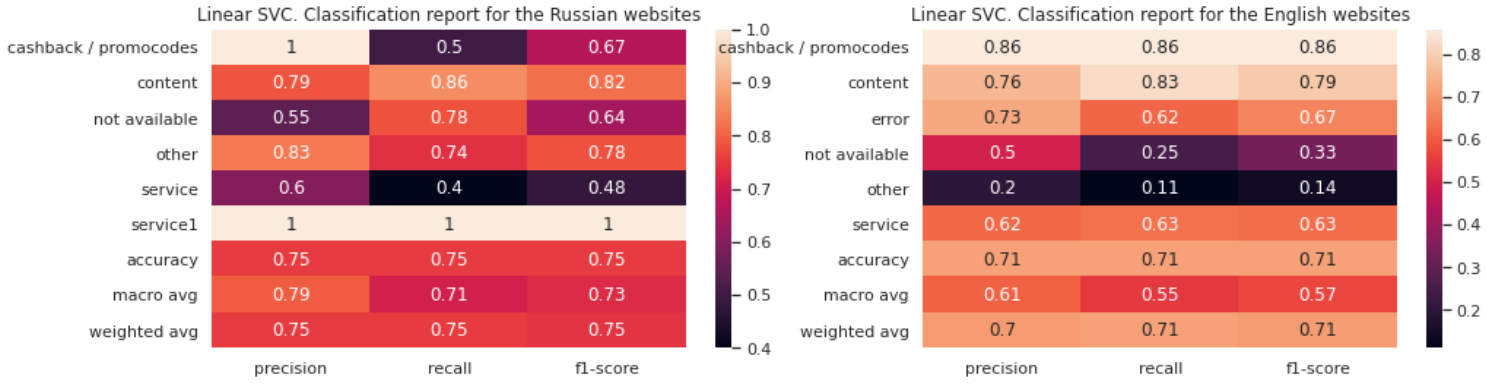


Figure 16 Classification reports for Russian and English language websites after application of Linear SVC model

(Source: compiled by authors based on Aviasales data)

It can be seen that the model predicted Russian language sites better than the English ones. In the case of Russian language sites though the special service class ‘service 1’ was predicted perfectly, the usual service obtained an F-score of only 48%. In the case of Russian language websites the best-predicted classes were ‘content’ and ‘other’. Conversely, in the case of English language dataset the other class was poorly predicted receiving an overall F-score of 0.14%. At the same time ‘cashback and promo codes’ and ‘content’ site scores were acceptable. Nevertheless, it was important to try the other models in an attempt to increase the scores.

Model 2: Gradient boosting model

The second model to try was the general Gradient boosting model. The model was applied together with randomized search. Randomized search allows to tune the parameters as it tries out each of them in terms of the model and the dataset. Randomized search allows to tune the parameters as it tries out each of them in terms of the model and the dataset. The fixed number of parameters is usually applied. For Gradient boosting model the following parameters were applied and tested (Table 17):

Parameter name	Parameter meaning	Tested values	Finally chosen value Russian language dataset	Finally chosen value English language dataset
n_estimators	the number of gradient stages to perform by the algorithm	200, 800	200	200
max_features	the number of features that the model takes into account while learning	auto, sqrt	auto	auto

Table 17 (continuation)

Parameter name	Parameter meaning	Tested values	Finally chosen value Russian language dataset	Finally chosen value English language dataset
max_depth	a parameter that sets a maximum value for nodes in each individual tree (weak model)	10, 40, None	None	None
min_samples_split	the minimum number of samples needed to split an internal node	10, 30, 50	30	50
min_samples_leaf	the minimum number of samples needed to be at a leaf node	1, 4	4	4
learning rate	rate at which a model learns	0.1, 0.5	0.5	0.5
subsample	size of a subsample	0.5, 1	0.5	0.5

Table 17 Parameters of the applied Gradient Boosting model

(Source: compiled by authors based on Aviasales data)

The Gradient Boosting model produced the following classification reports (Figure 17):

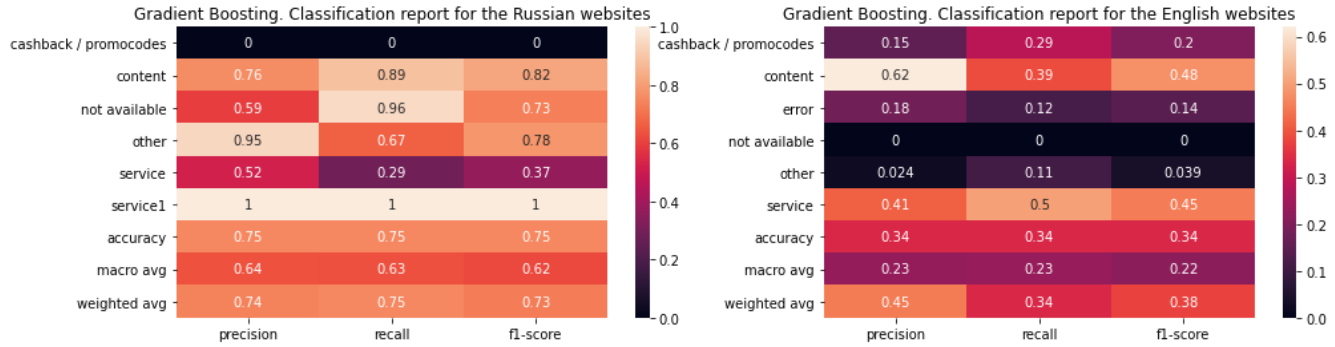


Figure 17 Classification reports for Russian and English language websites after application of Gradient Boosting model

(Source: compiled by authors based on Aviasales data)

The model received results comparable to the linear SVC in terms of Russian language websites, however performed significantly worse in the case of the English language websites. In the case of the Russian language 'content' and 'other' classes achieved relatively high F-1 score, while 'service' showed only 36% and 'cashbacks and promo codes' were not predicted.

In the case of the English language websites all classes were predicted quite poorly despite the drop of the missing values. Only the 'content' class showed F-1 score of 48%, however it is obvious that the model did not lead to higher scores as was initially expected. Thus, another attempt had to be made.

Model 3: CatBoost Classifier

The last model applied was CatBoost Classifier, which is an open-sourced library of gradient boosting algorithms. Gradient Boosting models solve classification or regression problems by uniting together several 'weak' prediction models.

The main difference of CatBoost from all other gradient boosting algorithms is its ability to perform well with categorical features (the ones that have a discrete set of values that cannot be compared with each other) (Garkavenko, 2020).

To increase the chances of obtaining the best model, randomized search was also applied. In term of this thesis the following parameters were introduced in the random search and finally chosen (Table 18):

Parameter name	Parameter meaning	Tested values	Finally chosen value Russian language dataset	Finally chosen value English language dataset
Iterations	Max number of trees created	100, 200, 300	300	300
Learning rate	Rate of the learning process	0.03; 0.1	0.1	0.1
Depth	Depth of the tree	2, 4, 6, 8	6	6
l2_leaf_reg	Regularization parameter	1, 2, 3, 4 ,5, 7, 9	7	4

Table 18 Parameters of CatBoost model obtained with randomized search

(Source: compiled by authors based on Aviasales data)

For all other parameters default values were used. Thus, the following classification reports were achieved (Figure 18):

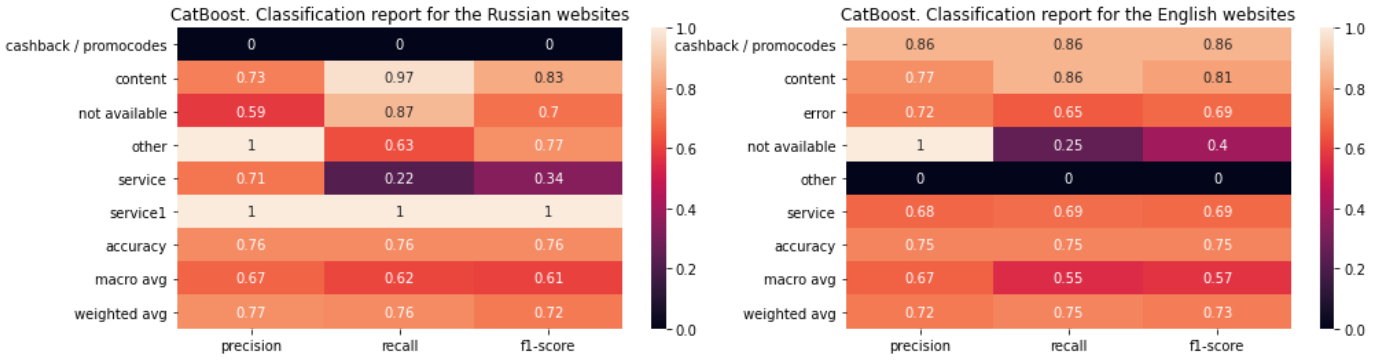


Figure 18 Classification reports or Russian and English language websites after application of CatBoost model

(Source: compiled by authors based on Aviasales data)

In the case of Russian language websites ‘content’, ‘not available’ and ‘other’ classes achieved quite high scores, while the model failed to predict ‘cashback and promo codes’. Interesting to note that ‘other’ class achieved quite a high F-1 score, though it initially contained very diverse variables. Nevertheless, ‘service’ class was rather poorly predicted mainly due to low recall meaning that a low amount of relevant items was selected.

Conversely, for English language websites CatBoost classifier could predict rather decently such classes as ‘cashback and promo codes’ and ‘content’, while the ‘other’ class was not identified. That is probably due to the fact that in case of the English language a majority of the websites belonging both to ‘content’ and ‘cashback and promo codes’ classes have similar structure, meaning that, for instance, the general structure of travel blogs which belong to the content class is often done in the same manner. ‘Service’ and ‘error’ classes showed F-1 score of 0,69 most probably due to higher amount of variability and content dispersion. In other words, websites within those two groups are not having similar structure or similar set of keywords. In fact, since the class ‘error’ was specifically introduced to contain all the websites that are not in either English or Russian language. Therefore, it is easy to imagine how the languages can vary within this class. The same can be said about ‘service’ class because this category contained not only services dedicated to the travel theme like selling plane tickets or bookings but also included absolutely different websites with highly diverse content. ‘Not available’ and ‘other’ classes showed the lowest quality among all. The category ‘other’ once again can be explained by a high diversity among its content, since the introduction of this class has been specifically carried out to include websites whose content did not belong in the above-mentioned classes.

Therefore, the comparison of F-1 and accuracy scores of the executed models looks the following way (Table 19):

Russian language websites			
Class / Accuracy	Linear SVC	Gradient Boosting	CatBoost
Cashback and promo codes	0.67	0	0
Content	0.82	0.82	0.83
Not available	0.64	0.73	0.7
Other	0.78	0.78	0.77
Service	0.48	0.37	0.34
Accuracy	0.75	0.75	0.76

Table 19 (continuation)

English language websites			
Class / Accuracy	Linear SVC	Gradient Boosting	CatBoost
Cashback and promo codes	0.86	0.2	0.86
Content	0.79	0.48	0.81
Error	0.67	0.14	0.69
Not available	0.33	0	0.4
Other	0.14	0.039	0
Service	0.63	0.45	0.69
Accuracy	0.71	0.34	0.75

Table 19 Comparison of classification models
(Source: compiled by authors based on Aviasales data)

Despite CatBoost model inability to discern Russian language cashback and promo code sites, it showed overall higher results than the other two models in terms of English language websites. Thus, it was decided to choose the CatBoost model for further predictions and data analysis.

2.4. Further analysis and data visualization

Data obtained from the classification has been exported for further analysis and data visualization. Excel was chosen as one of the tools for interactive dashboard creation as it is one of the most widespread and familiar for managers programs. Additional dashboard was created in Tableau as the program represents the most advanced analytical tools. It has been merged with the .xlsx dataset containing 303 223 rows that included information on the vertical, advertiser and affiliates that promote a certain advertiser.

The dataset is presented in the following table (Table 20):

Vertical	Advertiser	Affiliate
Car Rentals	101lugaresincreibles.com	noticiasidetodo.blogspot.com
Information	10best.com	rhodel.com
Aggregator	123millhas.com	oneworld-7.blogspot.com
Aggregator	123millhas.com	cupomdagalera.com.br

Table 20. Excel dataset provided by Aviasales (Source: Aviasales)

Here vertical basically means the type or a niche of an advertiser. Advertiser is a merchandiser and an affiliate is publisher. As the data is presented in an advertisers' viewpoint the advertiser is repeated in the dataset as many times as many affiliates it is promoted by. Moreover, the number of unique affiliates do not correspond with the initial Python dataset.

The merge of this dataset with the affiliate languages and classes allows to get more insights from the data. From the Python dataset affiliate language and class were added via VLOOKUP function to allow detailed analytics based both on peculiarities between Russian and English websites and their determined class. The interactive Excel dashboard was formed to derive managerial insights (Figure 19):

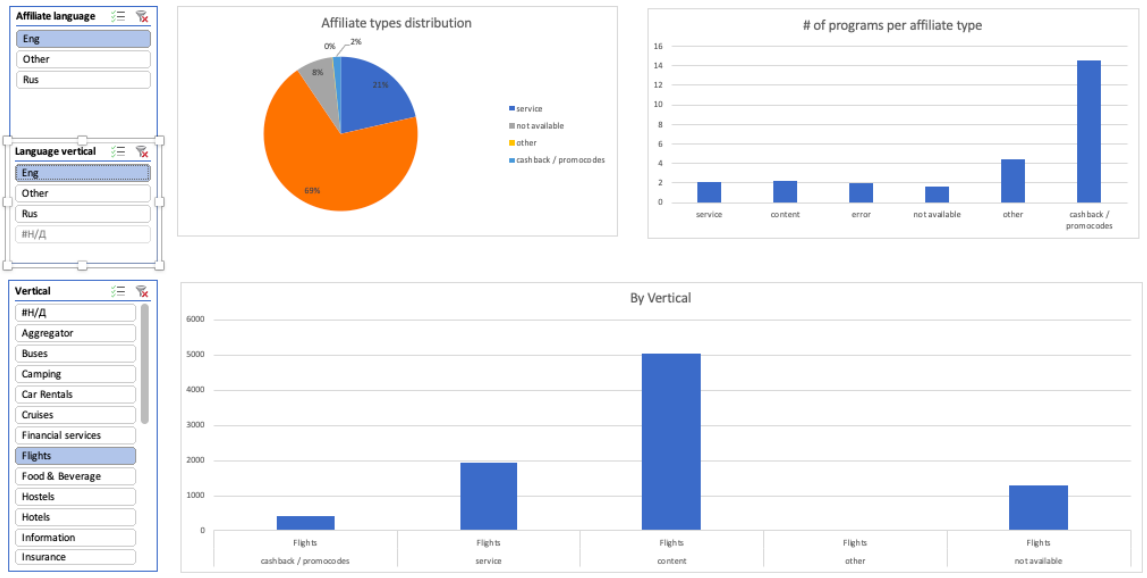


Figure 19 Interactive Excel dashboard

(Source: compiled by authors based on Aviasales data)

The outlook of the additional Tableau dashboard appears as following (Figure 20):

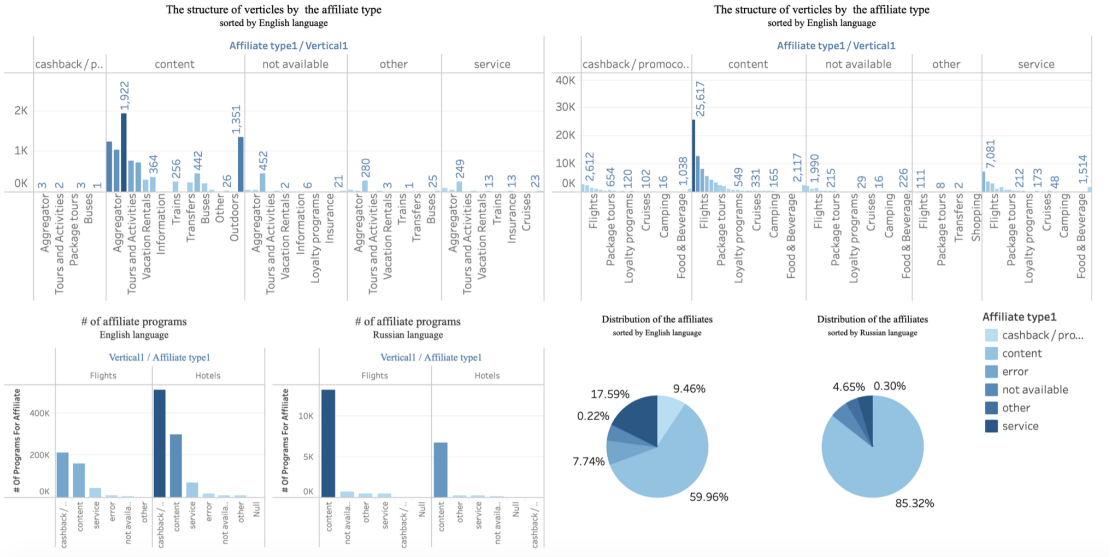


Figure 20 Tableau dashboard

(Source: compiled by authors based on Aviasales data)

However, even though Tableau possesses the necessary computing power and variety of illustrating capabilities, there are some serious obstacles due to which it was decided to abstain from including the detailed graphs in this paper:

1. The most important reason is that there is no necessity in doubling each detailed graph since they have been also made in excel. Since the excel tool is more popularised in both practical and scholar communities, as well as the fact that it can be easily understood even by people without previous interaction with the tool, the excel graphs were chosen as primary ones.
2. Secondly, Tableau graphs are impossible to download directly and in order to prevent the quality reduction of the picture, it is more wise to show individual graphs in the excel format. Moreover, tableau graphs are also incapable of changing their size in a manner as flexible as the excel one.
3. Last but not least, the tableau tool does not provide its user with the option of data modification, and therefore any changes in the data should be made directly in the data set priorly to uploading them to the Tableau. This specific feature does not allow to reduce the duplicates of some data instances that happen to be present in these particular datasets.

Therefore, the overall dashboards have been made in both formats, however, for more detailed pictures the excel tool was the primary one.

Russian language affiliates

Among approximately 3000 affiliates presented in the Russian language the content class is the most widespread one. It accounts for 70% of the whole dataset, while the second largest group — service represents only 9% of the data.

Thus, the division of Russian datasets by classes is presented in the following Figure (Figure 20):

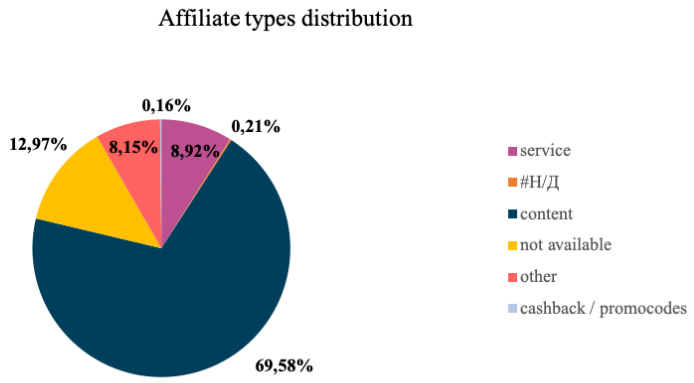


Figure 20. Distribution of Russian language affiliates according to their class (Source: compiled by authors based on Aviasales data)

Important to note that 13% of the data are broken links. This partly reflects the quality of the participating affiliates: many of them fastly transform into sites that no longer present interest to the creator and get shut down due to unpaid hosting fees. Moreover, a lot of Russian language affiliates use freemium website constructors like uCoz or Weebly that also revise the users activity. Broken links basically present noise and confuse the managerial analysis. That is why it is important to detect and remove such sites from the system.

Content sites are especially widespread in the “Flights” and “Hotels” verticals, which can be explained by the existence of a large number of blogs devoted to hotel reviews, best flights information sharing as well as the own websites of hotels and aviation companies. Service sites represent a similar picture. Thus, it can be deduced that the reason of such popularity is that flights and hotels are of the main consumers’ interest in the travel industry. Thus, these services are much more demanded than, for example, insurance or transfers. Cashback/promo code sites are modestly presented and the main verticals connected to them

is ‘Aggregator’. This reveals the alignment between affiliates and the advertiser. It is quite natural that the aggregator will be promoted on the site with cashbacks and promo codes as the target audience of such types of sites are people looking for cost saving or discount. An aggregator might be of their interest as it can help find cheapest offers as well as get additional discounts, for example, for buying a hotel room and renting a car at the same time.

Therefore, the structure of the verticals by the affiliate is the following (Figure 21):

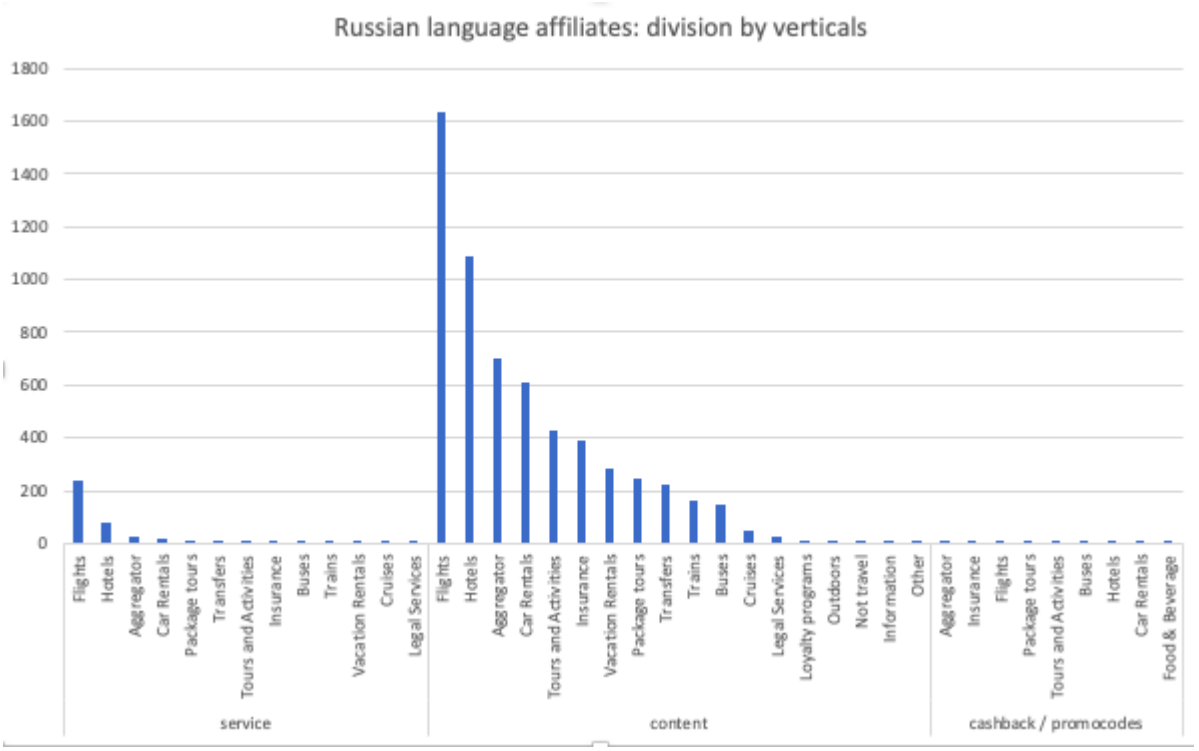


Figure 21 The structure of verticals by the Russian language site affiliate type.

(Source: compiled by authors based on Aviasales data)

It is also interesting to consider the analysis of the main verticals: “Flights” and “Hotels” to look at the data from different perspective (Figure 22):

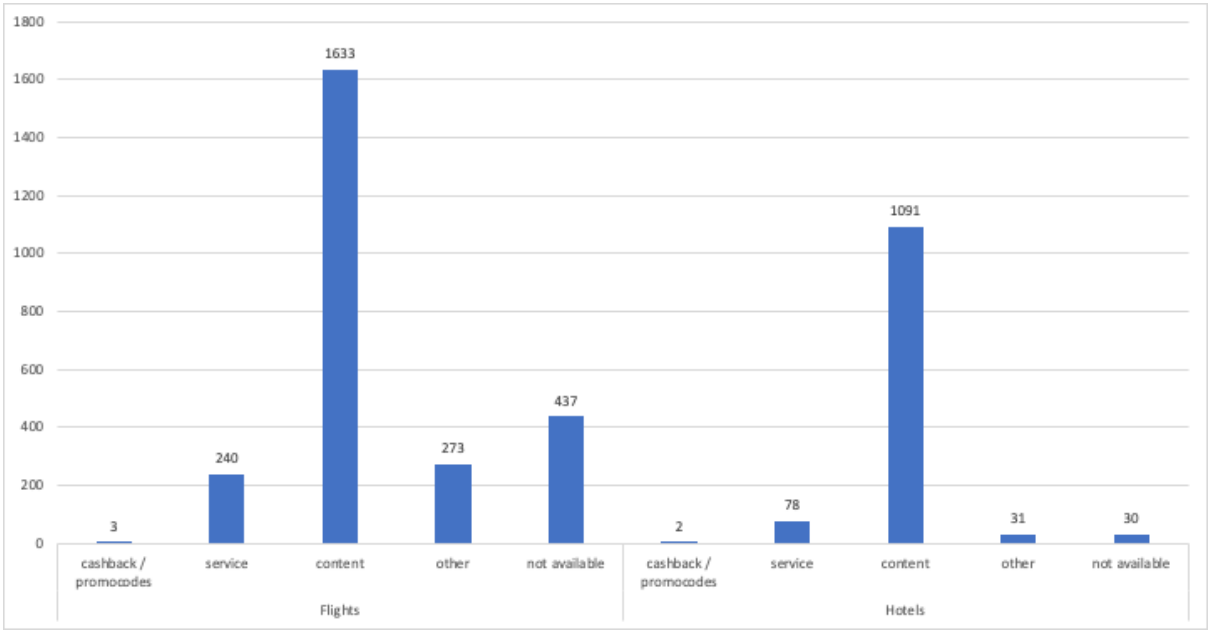


Figure 22 The structure of Flights and Hotels verticals

(Source: compiled by authors based on Aviasales data)

Flights have a significant number of not available links, which means that this category is the most susceptible to noise.

Another interesting information to look into is the number of affiliate programs a certain type of affiliate participates into. Here cashback/promo code sites are obvious leaders with participation in approximately 5 affiliate programs. The full data on the matter is presented in the Figure below (Figure 23):

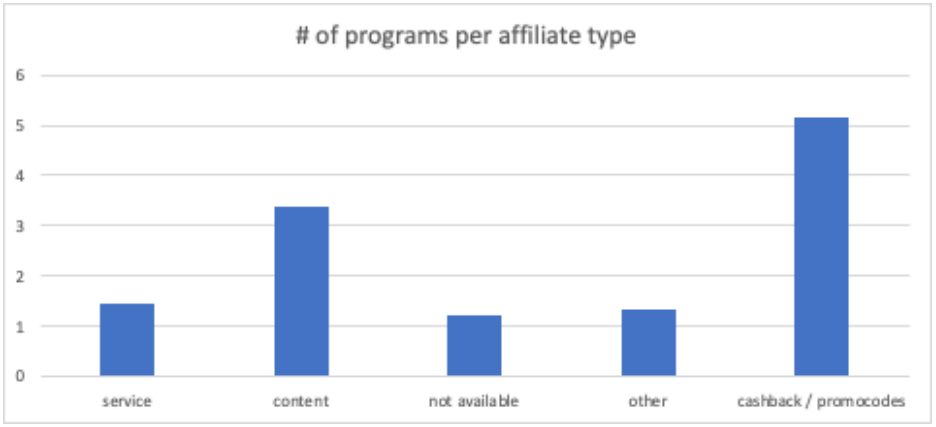


Figure 23 The number of programs per affiliate type (Russian websites)

(Source: compiled by authors based on Aviasales data)

Here the discussion returns to the issue of trust presented in the literature review. Thus, the presence of many affiliate programs in terms of one site can irritate potential consumers as well as scare them off due to similarity with fraudulent sites.

All in all, main Russian affiliates are of a content type, prevailingly presented in 'Flights' and 'Hotels' verticals. The affiliate type within the most affiliate programs are cashbacks/promo codes.

English language affiliates

The Excel data contained 50 815 English language affiliates. Here, similar to Russian language case, content sites are in the lead. Moreover, they are also followed by service sites while cashback and promo codes account for modest 2%. The full classes distribution is the following (Figure 24):

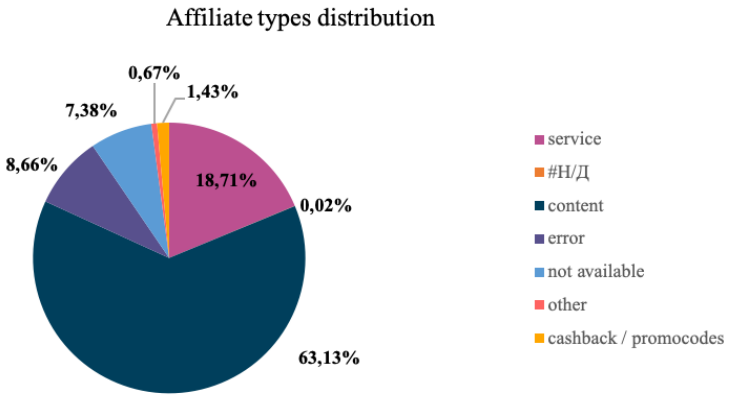


Figure 24 Distribution of English language affiliates according to their class
(Source: compiled by authors based on Aviasales data)

Interesting to note that the % share of not available sites is substantially less than in the case of Russian sites, however, the conclusions in this case are hard to make. The Russian language sites selection is smaller than the English sites, thus, the actual number of broken sites in the English language case is almost 7 times higher. Nevertheless, in relative terms it can be assumed that English language affiliate network consists of more quality made sites in comparison to the whole English language dataset than Russian language database

Similar to Russian affiliates 'Flights' and 'Hotels' are the most popular verticals among content and service sites. Thus, the structure is the following (Figure 25):

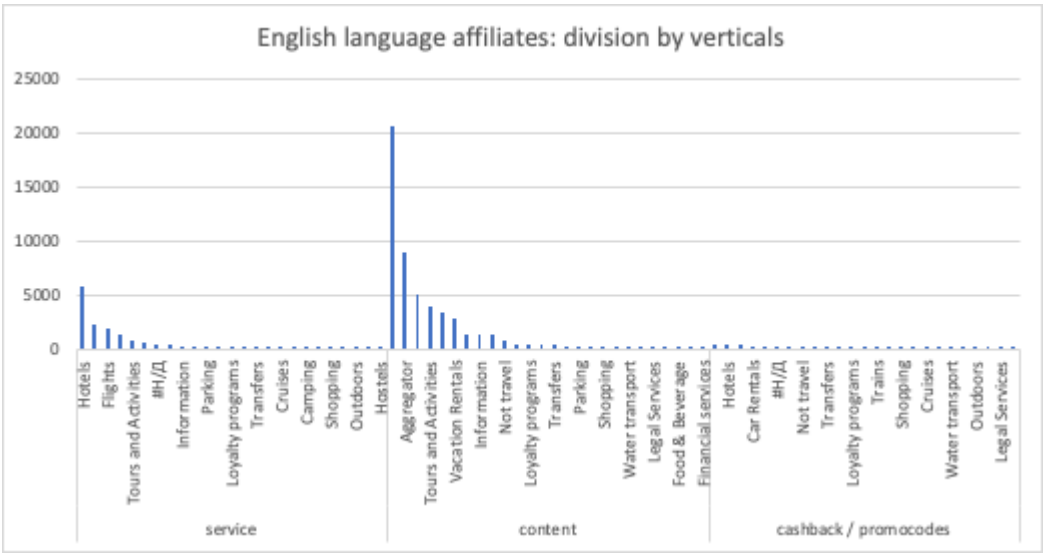


Figure 25 The structure of verticals by the English language site affiliate type.

(Source: compiled by authors based on Aviasales data)

By looking at the verticals from another viewpoint, it is again seen that the content sites are majorly involved across all the verticals. The Top-5 verticals are the following (Figure 26):

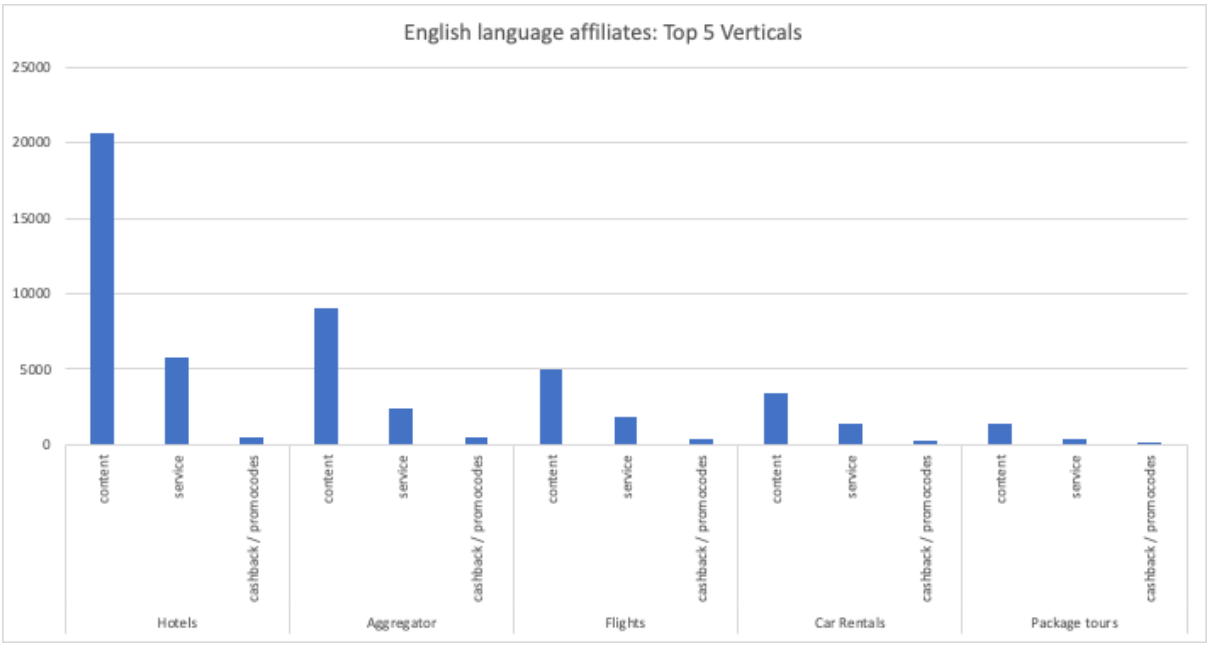


Figure 26 The structure of classes across Top-5 verticals

(Source: compiled by authors based on Aviasales data)

As in the case of the Russian websites Flights and Hotels verticals are among the most involved in the affiliate programs. Once again cashbacks and promo codes are the type of affiliates with the largest number of affiliate programs presented on the website (Figure 27):

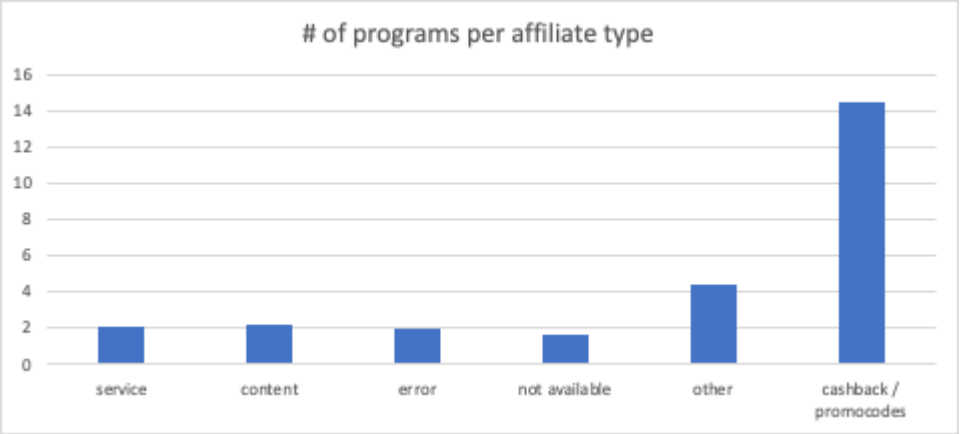


Figure 27 The number of programs per affiliate type (English websites)

Interesting to note that in the case of English language the number of affiliate programs is almost 3 times higher than in the Russian language websites. This is explained by the existence in the data of the coupon sites like thinkup.com, which solely exist based on coupons and accounts for 1015 affiliate programs. Moreover, English-speaking countries are well-known for their affection towards coupons and promo codes. Thus, in the context of the USA the term ‘a coupon nation’ or ‘a nation of coupon addicts’ is widely used in the press, for example Forbes (Thau, 2013). This country has a whole culture of coupons, which is reflected in TV reality shows and mass media. Thus, the Wall Street Journal first introduced the term ‘extreme couponing’ (Martin, 2010), which was later used as the title of the TLC Channel show devoted to the matter. Moreover, RetailMeNot Research (2013) showed that shoppers in the UK, Australia and Canada are also extremely involved in bargaining, especially in finding online deals. Such a mindset influences the behavior of affiliates that try to respond to consumer demands.

Thus, the main findings from both datasets are:

1. Content sites is the most widespread type of the affiliates among both languages studied
2. The dominance of content sites is spread across all the verticals
3. Cashback and promo codes sites are the affiliates that show the most interest in participation in affiliate programs. On average they participate in 5 and 15 affiliate

programs with the max number of 12 and programs 1015 on one site for Russian and English websites respectively.

2.5. Managerial application and further directions of research

Clusterization of the websites revealed the underlying differences between Russian and English language datasets. While the Russian language websites had certain groups that were standing out like real estate or city portals, the English language ones were deeply interconnected. At the same time Russian language websites contained a large travel themed sites group that included all sorts of affiliates from hotels to restaurants and travel agencies. The research showed that the linguistics of travel themed sites is quite similar and, thus, the clusterization approach is not feasible for travel affiliate program analysis. Thus, for managers this means that before making strategic decisions it is important to divide affiliates in accordance with the company's business objectives and based on own experience. The clusterization analysis showed that the differences among affiliate exists and affiliates must be divided into subgroups, however the machine was not sophisticated enough to replace the human mind.

The CatBoost classification model presented in terms of this paper represents a tool that can be used by businesses in modern affiliate marketing analysis. Classification of affiliates is extremely important as it allows to build a multifaceted strategy that takes into account peculiarities of each of the defined groups of affiliates. Thus, after manual division of affiliates into strategic groups a classification model can be applied to ease the further analysis.

In terms of this thesis the main classes of affiliates defined by Aviasales were 'content', 'service' and 'cashback and promo codes'. The in-depth analysis of these classes revealed that 'content' sites present the largest class involved into affiliate programs. Moreover, Flights and Hotels verticals are connected to the main part of 'content' class sites. Thus, for Aviasales this means that travel blogs and content sites that aggregate information about flights and hotels attract the attention of the most consumers. At the same time the abundance of these types of affiliates leads to loss in quality of their content, which, in turn, may negatively influence the advertiser's brand. Moreover, the analysis revealed that cashback and promo code sites are particularly susceptible to the affiliate programs and have a lot of links. This can create an impression of fraudulent activity, which once again affects the brand. It is important to take into consideration whether the company wants to join

hundreds of other programs already presented on such types of sites as the value of such cooperation can be diminished.

From the managerial point of view it is also important to take into account the abundance of not available or straightly redirecting to Aviasales affiliates. Large number of not available sites once again points out their quality. Such partnership does not imply long-term relationships and indicates that the owner of the affiliate website just wants to obtain easy and fast money. Straight redirect websites especially negatively influence the advertiser, which is Aviasales in this case. When the user expects to read some content, but instead moves right to Aviasales website, this creates an impression of a mirror site. Moreover, as the user was redirected, the action was formally made and thus Aviasales had to pay to this affiliate. However, in this situation the user had no intention to open the advertisers website. Thus, the affiliate simply gets money for nothing. Thus, it is extremely important for Aviasales to monitor the quality of the affiliates. Moreover, another suggestion is to develop a quick guide for affiliates on how to maintain the website and how the affiliate link must look like.

All in all, recommendations for Aviasales are the following:

- Focus on affiliates involved in flights and hotel verticals
- Check the network for fraudulent sites: both those sites that are dangerous for the users and those who deceive affiliate program by straight redirection of the user to the advertisers website
- Implement a system to check the quality of the affiliates content and derive a quick guide on affiliate sites management for these websites owners

From the academic point of view the further research questions can be the following:

- How does the quality of affiliate links influence an advertiser's brand?
- What attributes of affiliate links influence the user clicks on the affiliate link the most?
- What type of payment mechanism (CPC, CPA) is better for advertisers?
- Which affiliates present the most opportunities for monetisation?

Moreover, an important direction of the research is the study of the overall affiliate programs, namely, advertisers themselves. The authors of this thesis have carried out first steps in this direction. Thus, the advertisers were broken down into travel and non-travel ones. The solution of the initial data-problem required 'hand' labeling of the affiliates into two categories. This step was a necessary part of the supervised learning performed on the dataset. To be more precise and technical about the completed work, the labeling process was

basically the division of the dataset into two main groups: travel and non-travel, which in terms of Big Data analysis can be interpreted as 1/0 labelling. Since this work is out of the scope of this thesis, the details provided are limited, however, they can be the basis of the further research in the separate paper.

Conclusion

All in all, the paper presented a comprehensive approach towards the investigation of affiliate marketing in the travel sector. It contributed to the shrinkage of the existing literature gap, namely by describing the peculiarities of affiliates in the travel industry, and also providing the theoretical investigation of the specificity of the sector itself. Moreover, it implemented a real-life case study from one of the biggest travel aggregators – Aviasales, and analyzed its approach towards affiliate marketing, as well as provided managerial recommendations.

Despite the current slack in the growth of the global tourism industry (mainly due to the COVID-19 situation), the travel sector has a wide range of opportunities. As it was pointed out in the theoretical part of the paper some countries have already started to show signs of the increase in travel activities. Moreover, taking into account the overall size of the sector, travelling remains one of the most lucrious and prosperous industries for the implementation of affiliate marketing. That is why it was especially important to carry out the analysis of the affiliates in this industry.

Despite the growing popularity of affiliate marketing and increase in available data, from the technical perspective Big Data approaches are currently implemented in the field only to a limited degree. Thus, the thesis focused on the implementation of the modern Machine Learning algorithms to the analysis of the affiliate marketing program in Aviasales. Since affiliates are basically websites with text the paper described implementation of Natural language processing algorithms including language detection.

In terms of this thesis the affiliates were divided into two categories – English and Russian language websites. This approach allowed to determine the affiliates' structure within the program in relation to a particular market and, thus, analyse the situation more efficiently. Moreover, the paper considers various types of the affiliates and their relations with advertising verticals.

To achieve the goals presented in the introduction the paper introduced clusterization and classification algorithms that allowed both to determine hidden patterns of the data and

comply with Aviasales business requirements. The final part of the paper focused more on the practical recommendations for Avisales based on the existing approach and also provided ideas for future research.

The managerial conclusions were mostly dedicated to the importance of identification of possibly fraudulent web pages and checking the general quality of the affiliate since the original dataset includes a large amount of 'not available' content. Poor affiliate management can significantly damage the advertiser's brand and, thus, it is important to check the status of the affiliate network regularly. Moreover, it was also advised to focus on the two major verticals which are flights and hotels since they represent the most popular consumer queries in terms of travelling. Moreover, content websites were defined as one of the most important affiliate types. A win-win solution for an affiliate and an advertiser is the development of general content guidelines in order to be able to attract consumers and provide stimuli for them to make actions. Cashback and promo code sites were identified as the most involved in the affiliate programs type of affiliates. On such a type of sites up to 1000 affiliate programs can be presented, which diminishes their value and, thus, before including such type of sites into the affiliate network the managers must additionally evaluate risks and benefits and make sure that the long-term value can be achieved from such partnership.

The ideas of future research were formed taking into consideration the questions of how the characteristics of the affiliates content can contribute to monetisation and what is the influence of affiliate networks on perception of the Aviasales brand by customers.

In the end, despite the paper being one of the few dedicated to this vast and complex topic of affiliates in the travel industry, the stated research goals were completed. However, due to fast development of modern marketing approaches and a constantly changing environment, the opportunity for future research remains available.

References

1. 20 Online Review Stats to Know in 2019. (2021). Retrieved 9 April 2021, from https://www.qualtrics.com/blog/online-review-stats/

2. A Coupon Nation: Americans proudly use Coupons more than shoppers in Great Britain, India and China, among others. (2018, June 29). Retrieved May 17, 2021, from https://www.prnewswire.com/news-releases/a-coupon-nation-americans-proudly-use-coupons-more-than-shoppers-in-great-britain-india-and-china-among-others-220485721.html

3. A practical explanation of a Naive Bayes classifier. MonkeyLearn. Retrieved 21 May 2021 from: https://monkeylearn.com/blog/practical-explanation-naive-bayes-classifier

4. Agyei, J., Sun, S., Abrokwah, E., Penney, E. K., & Ofori-Boafo, R. (2020). Influence of Trust on Customer Engagement: Empirical Evidence From the Insurance Industry in Ghana. SAGE Open, 10(1), 2158244019899104

5. AI in affiliate marketing: How Do Marketers benefit from it? (2020, April 23). Retrieved May 29, 2021, from https://affise.com/blog/ai-in-affiliate-marketing-how-do-marketers-benefit-from-it

6. Akcura, M.T. (2010). Affiliated marketing. Information Systems and e-Business

7. Automation in Affiliate Marketing. (n.d.). Retrieved March 09, 2021, from https://affiliatevalley.com/guides/automation-in-affiliate-marketing

8. A. F., S., Lakshmi, I., Prashant, P., & Rahul, S. (2019). Effect of Trust, Quality of Products and Quality Services on Purchase Decisions on E-Commerce Shopee in Palembang City. Regular Issue, 3(12), 1-6. doi: 10.35940/ijmh.l0313.0831219

9. Bauman, A., & Bachmann, R. (2017). Online Consumer Trust: Trends in Research. Journal Of Technology Management & Innovation, 12(2), 68-79. doi: 10.4067/s0718-27242017000200008

10. Bert. (n.d.). Retrieved May 10, 2021, from https://huggingface.co/transformers/model_doc/bert.html

11. Bhatnagar, A., & Papatla, P. (2001). Identifying locations for targeted advertising on the Internet. International Journal of Electronic Commerce, 5(3), 23-44. doi: 10.1080/10864415.2001.11044210

12. Brems, M. (2019, June 10). A one-stop shop for principal component analysis. Retrieved May 19, 2021, from https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c

13. Brownlee, J. (2019, August 7). What Are Word Embeddings for Text? Machine Learning Mastery. https://machinelearningmastery.com/what-are-word-embeddings/.

14. Brownlee, J. (2020, August 14). A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning. Machine Learning Mastery. https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/.

15. Daniele, R., Frew, A. J., Varini, K., & Magakian, A. (2009). Affiliate marketing in travel and tourism. Information and Communication Technologies in Tourism 2009, 343-354.

16. Deloitte. (2020). 2019 Travel and Hospitality Industry Outlook. Retrieved March 5, 2021, from https://www2.deloitte.com/us/en/pages/consumer-business/articles/travel-hospitality-industry-outlook.html

17. Delua, J. (2021). Supervised vs. Unsupervised Learning: What's the Difference?. Ibm.com. Retrieved 15 April, 2021, from https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning.

18. Digital transformation in travel industry | Adamo Digital. Adamo Digital. (2020). Retrieved 8 January, 2021, from https://adamodigital.com/blog/how-digital-transformation-affect-the-travel-and-tourism-industry/.

19. Drugău-Constantin, A. L. (2018). Emotional and cognitive reactions to marketing stimuli: Mechanisms underlying judgments and decision making in behavioral and consumer neuroscience. Economics, Management, and Financial Markets, 13(4), 46-52.

20. Duffy, D. L. (2005). Affiliate marketing and its impact on e-commerce. Journal of Consumer Marketing.

21. Dwivedi, R. (2017). Analyzing Impact of Affiliate Marketing on Consumer Behavior withM-Commerce Perspective. SMS Journal Of Entrepreneurship And Innovation, 3(02). doi: 10.21844/smsjei.v3i02.9733

22. Edelman, B., & Brandi, W. (2015). Risk, Information, and Incentives in Online Affiliate Marketing. Journal of Marketing Research, 52(1), 1-12. doi: 10.1509/jmr.13.0472

23. Exploring the impacts of COVID-19 on travel behavior and mode preferences. (2020, November) https://www.sciencedirect.com/science/article/pii/S2590198220301664

24. Gandhi, R. (2018, July 5). Support Vector Machine - Introduction to Machine Learning Algorithms. Medium. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.

25. Garkavenko, M. (2020, July 10). Categorical features parameters in CatBoost. Medium. https://towardsdatascience.com/categorical-features-parameters-in-catboost-4ebd1326bee5.

26. Goldschmidt, S., Junghagen, S., & Harris, U. (2003). Strategic affiliate marketing. Edward Elgar Publishing.

27. Gregori, N., Daniele, R., & Altinay, L. (2014). Affiliate marketing in tourism: determinants of consumer trust. Journal of Travel Research, 53(2), 196-210.

28. Haq, Zia. (2012). Affiliate marketing programs: A study of consumer attitude towards affiliate marketing programs among Indian users Affiliate marketing programs: A study of consumer attitude towards affiliate marketing programs among Indian users. International Journal of Research Studies in Management. 1. 10.5861/ijrsm.2012.v1i1.84.

29. Harrington, P. (2012). Machine learning in action. Shelter Island: Manning.

30. Hazoom, M. (2018, December 26). Word2Vec for phrases - Learning embeddings for more than one word. Retrieved May 22, 2021, from https://towardsdatascience.com/word2vec-for-phrases-learning-embeddings-for-more-than-one-word-727b6cf723cf

31. Henrique, A. (2021, May 25). Clustering with k-means: Simple yet powerful. Retrieved April 19, 2021, from https://medium.com/@alexandre.hsd/everything-you-need-to-know-about-clustering-with-k-means-722f743ef1c4

32. Hooghe, M. (2017). Handbook on Political Trust, by S. Zmerli and T. van der Meer, Cheltenham, UK, Edward Elgar, 2017. Journal Of Trust Research, 7(2), 220-225. doi: 10.1080/21515581.2017.1364481

33. IAB. (2016). IAB Affiliate Marketing Handbook. Internet Advertising Bureau. https://www.iab.com/wp-content/uploads/2016/11/IAB-Affiliate-Marketing-Handbook_2016.pdf

34. Iva, S. (2008). Tourist affiliate program while using online booking system with possibility of entering B2B code. Turizam, 12, pp. 46-52.

35. Ivkovic, M., & Milanov, D. (2010, November). Affiliate internet marketing: Concept and application analysis. International Conference on Education and Management Technology, Cairo, pp. 319-323. doi: 10.1109/ICEMT.2010.5657647

36. Jaadi, Z. (n.d.). A step-by-step explanation of principal component Analysis (PCA). Retrieved May 6, 2021, from https://builtin.com/data-science/step-step-explanation-principal-component-analysis

37. Jolliffe, I.T., & Jackson, J. (1993). A User's Guide to Principal Components. The Statistician, 42, 76-77.

38. K-Means Clustering Algorithm - Sarath SL. Medium. Retrieved 10 May 2021 from: https://medium.com/@slsarath2/k-means-clustering-algorithm-5fa9d5d64326

39. K-Means Clustering Explained Visually In 5 Minutes. (2020). Retrieved 10 May 2021 from: https://medium.com/dataseries/k-means-clustering-explained-visually-in-5-minutes-b900cc69d175

40. Libai, B., Biyalogorsky, E., & Gerstner, E. (2003). Setting referral fees in affiliate marketing. Journal of Service Research, 5(4), 303-315. doi: 10.1177/1094670503005004003

41. Lock, S. (2020, December 9). Global tourism industry - statistics & facts. Statista https://www.statista.com/topics/962/global-tourism/

42. Mackey, T., Kalyanam, J., Klugman, J., Kuzmenko, E., & Gupta, R. (n.d.). Solution to Detect, classify, and Report illicit online marketing and sales of controlled substances via Twitter: Using machine learning and Web forensics to Combat Digital Opioid Access. Retrieved May 28, 2021, from https://www.jmir.org/2018/4/e10029/

43. Markets, R. A. (2020, July 8). 2020 Market Study on the Worldwide Car Rental Industry to 2027 - Adoption of Car Rental Management Software Presents Opportunities. Cision. https://www.prnewswire.com/news-releases/2020-market-study-on-the-worldwide-car-rental-industry-to-2027---adoption-of-car-rental-management-software-presents-opportunities-301089925.html

44. Martin, T. (2010, March 08). Hard times turn coupon clipping into the newest extreme sport. Retrieved April 14, 2021, from https://www.wsj.com/articles/SB10001424052748703615904575053413229901660

45. Mican, D. (2008). Optimized advertising content delivery in affiliate networks (Technical Report). Babes-Bolyai University, Romania.

46. Networks help drive affiliate marketing into the mainstream. (2016). Retrieved 2 February 2021, from

https://go.rakutenmarketing.com/hubfs/Networks_Help_Drive_Affiliate_Marketing_Into_The_Mainstream.pdf.

47. Neumayer, Sebastian; Nimmer, Max; Setzer, Simon; Steidl, Gabriele. Applied Mathematics & Optimization. Dec2020, Vol. 82 Issue 3, p1017-1048. 32p. DOI: 10.1007/s00245-019-09566-1.

48. Novoselov, A. (2017). The Current State of the Travel Affiliate Market. Travelpayouts Blog – Travel affiliate network. Retrieved 6 January 2021, from https://blog.travelpayouts.com/en/travel-affiliate-market-trends/.

49. Oktadiana, H., & Kurnia, A. (2011). How customers choose hotels. Binus Business Review, 2(1), 510-517.

50. Olbrich, R., Bormann, P. M., & Hundt, M. (2019). Analyzing the Click Path Of Affiliate-Marketing Campaigns: Interacting Effects of Affiliates' Design Parameters With Merchants' Search-Engine Advertising. Journal of Advertising Research, 59(3), 342-356.

51. Over 70+ Online Travel Booking Statistics (2020-2021). (2021). Retrieved 9 April 2021, from https://www.condorferries.co.uk/online-travel-booking-statistics

52. Papatla, P., & Bhatnagar, A. (2002). Choosing the right mix of on-line affiliates: How do you select the best?. Journal of Advertising, 31(3), 69-81.

53. Parker, S. (2020, January 27). AI and BI are vibrantly sparking new trends in affiliate marketing. Retrieved March 30, 2021, from https://www.smartdatacollective.com/ai-bi-are-vibrantly-sparking-new-trends-in-affiliate-marketing/

54. Porter, J. (2020). Grounded airline planes turned into pop-up restaurants sell out in 30 minutes. The Verge. Retrieved 15 January 2021, from https://www.theverge.com/2020/10/12/21512515/singapore-airlines-airbus-a380-pop-up-restaurant-changi-airport-coronavirus-pandemic.

55. Predicting website categories using Supervised Learning. Medium. Retrieved 10 May 2021 from: https://medium.com/@amithnmbr/predicting-website-categories-using-supervised-learning-4a4b3349bfcb

56. Racherla, P., Mandviwalla, M., & Connolly, D. (2012). Factors affecting consumers' trust in online product reviews. Journal Of Consumer Behaviour, 11(2), 94-104. doi: 10.1002/cb.385

57. Radford, A. (2020, March 02). Improving language understanding with unsupervised learning. Retrieved May 11, 2021, from https://openai.com/blog/language-unsupervised/

58. Roderick M., K. (2021). Rethinking Trust. Retrieved 9 April 2021, from https://hbr.org/2009/06/rethinking-trust.

59. Seo, D. (2020, May 09). BERT – крупнейшее обновление Google. Как оптимизировать ваш сайт под BERT. Retrieved May 5, 2021, from https://vc.ru/seo/101834-bert-krupneyshee-obnovlenie-google-kak-optimizirovat-vash-sayt-pod-bert

60. Sherwood, H. (2019). How Airbnb took over the world. the Guardian. Retrieved 4 February 2021, from https://www.theguardian.com/technology/2019/may/05/airbnb-homelessness-renting-housing-accommodation-social-policy-cities-travel-leisure.

61. Suchada, J., Watanapa, B., Charoenkitkarn, N., & Chirapornchai, T. (2018). Hotels and Resorts Rent Intention via Online Affiliate Marketing. Kne Social Sciences, 3(1), 132. doi: 10.18502/kss.v3i1.1402.

62. Suresh, V., VetriSelvi, M., Maran, K., & Shanmuga, P. (2018). A Study On Impact Of An Affiliate Marketing In E-Business For Consumer's Perspective. International Journal of Engineering and Technology (IJET), 10(2), 470-473.

63. Thaichon, P., Liyanaarachchi, G., Quach, S., Weaven, S., & Bu, Y. (2019). Online relationship marketing: evolution and theoretical insights into online relationship marketing. Marketing Intelligence & Planning, 38(6), 676-698. doi: 10.1108/mip-04-2019-0232

64. Thau, B. (2013, September 09). Americans are Big Couponers, while The Chinese are more inclined to shop online for deals. Retrieved May 03, 2021, from https://www.forbes.com/sites/barbarathau/2013/09/09/americans-are-big-couponers-while-the-chinese-are-more-inclined-to-shop-online-for-deals/?sh=57b87cc74227

65. The future of affiliate marketing in the age of machine learning: Bizacuity. (2020, July 15). Retrieved May 10, 2021, from https://bizacuity.com/future-affiliate-marketing-age-analytics-machine-learning/

66. Traveler Trends Tracker. (2020). Adara. https://adara.com/traveler-trends-tracker/

67. Travelling. (n.d.). Retrieved January 3, 2021, from https://dictionary.cambridge.org/dictionary/english/travelling

68. Tsvetkova, N. (2021). How to track affiliate links – the best tools to use. Retrieved 10 April 2021, from https://blog.travelpayouts.com/en/best-tools-for-tracking-affiliate-links-on-the-website/

69. Useful affiliate marketing statistics in 2021. (2021, January 27). Supermetrics. https://supermetrics.com/blog/affiliate-marketing-statistics

70. Wang, L., Law, R., Hung, K., & Guillet, B. (2014). Consumer trust in tourism and hospitality: A review of the literature. Journal Of Hospitality And Tourism Management, 21, 1-9. doi: 10.1016/j.jhtm.2014.01.001

71. Wilson, A. (2019, October 1). A Brief Introduction to Supervised Learning. Medium. Retrieved 24 May of 2021 from: https://towardsdatascience.com/a-brief-introduction-to-supervised-learning-54a3e3932590.

72. Авиасейлс. (2018, December 11). История Travelpayouts: от идеи до выплаты первого миллиарда. Retrieved April 11, 2021, from https://vc.ru/aviasales/52867-travelpayouts

73. Бизнес-класс: Как скандалист создал крупнейший поисковик авиабилетов в России. (n.d.). Retrieved March 01, 2021, from https://www.the-village.ru/business/stories/172379-aviasales-kalinova

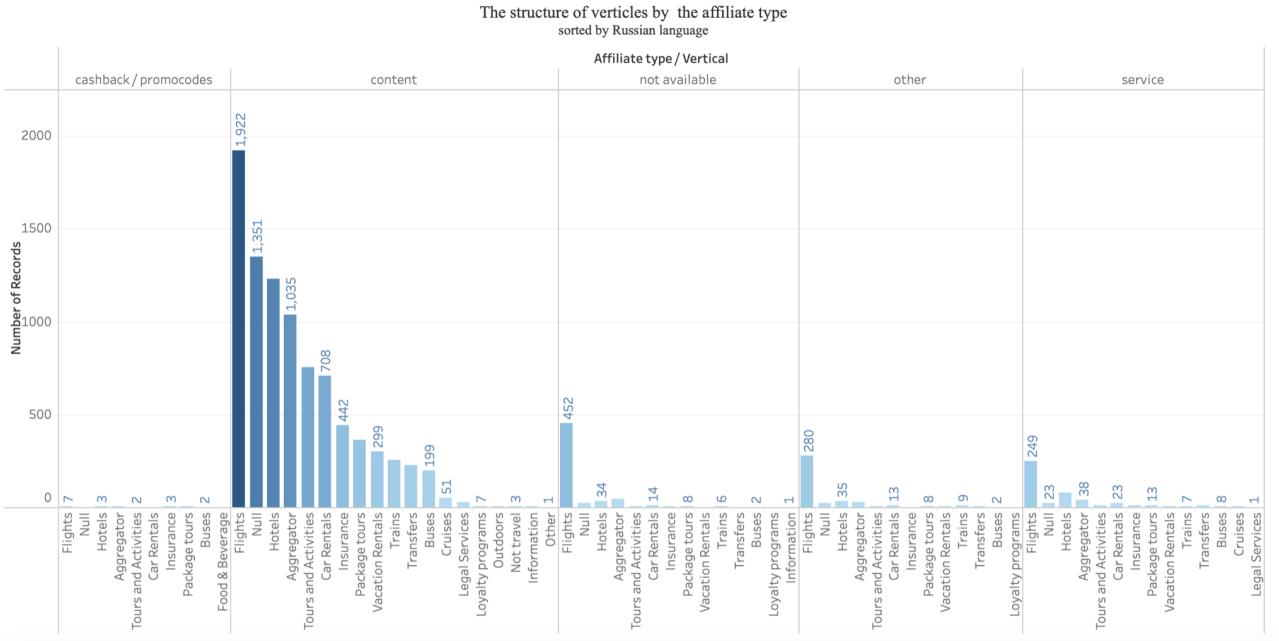
74. Казьмина, И., Жукова, К., Юзбекова, И., Петухова, Л., Бородина, В., Титова, Ю, . . . Яковенко, Д. (2020, February 20). 20 самых дорогих компаний Рунета. Рейтинг Forbes. Retrieved March 06, 2021, from https://www.forbes.ru/biznes-photogallery/393345-20-samyh-dorogih-kompaniy-runeta-reyting-forbes?photo=9

75. Машинное обучение. (n.d.). Retrieved April 25, 2021, from https://www.calltouch.ru/glossary/mashinnoe-obuchenie/

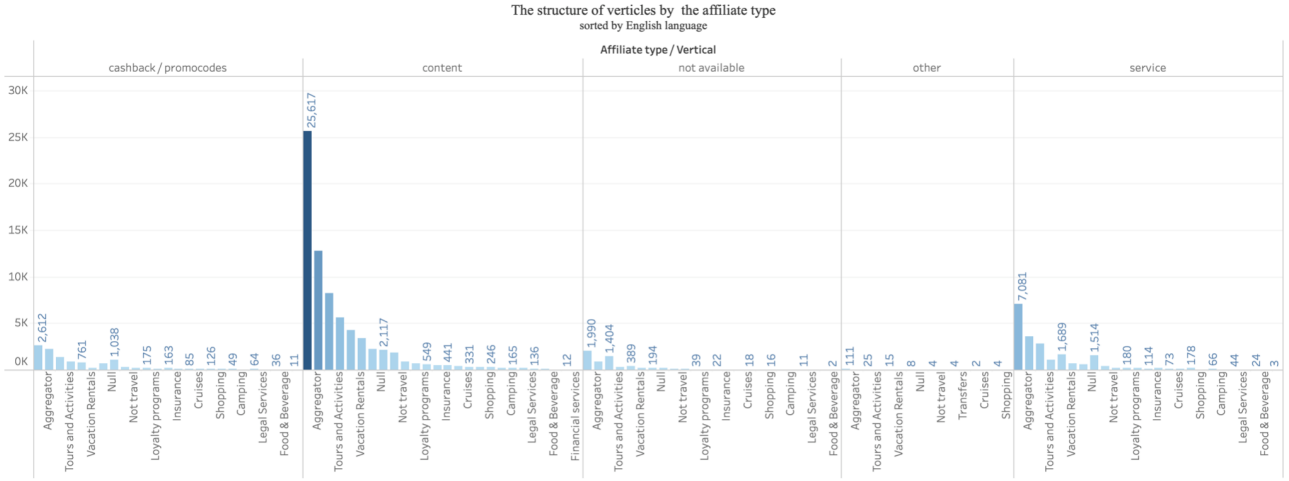
76. Немного про word2vec: полезная теория. NLPx. (2015, November 3). http://nlpx.net/archives/179.

APPENDICES

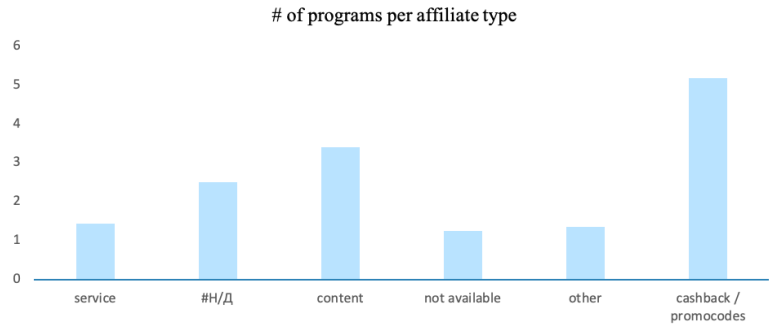
Appendix 1: Tableau dashboard. The structure of verticals by the Russian language site affiliate type.



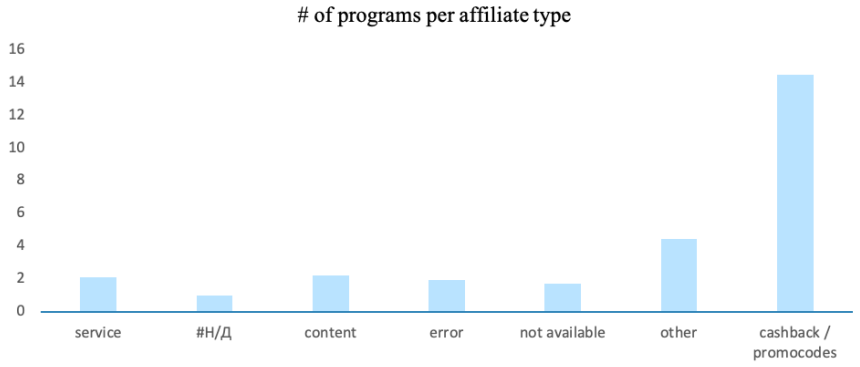
Appendix 2: Tableau dashboard. The structure of verticals by the English language site affiliate type.



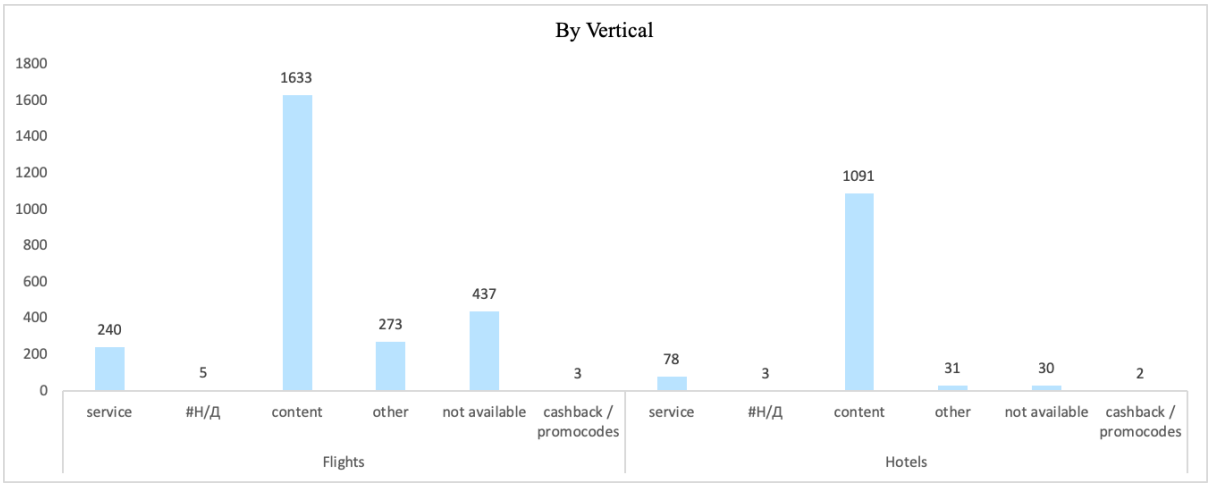
Appendix 3: Tableau dashboard. The number of programs per affiliate type (Russian websites)



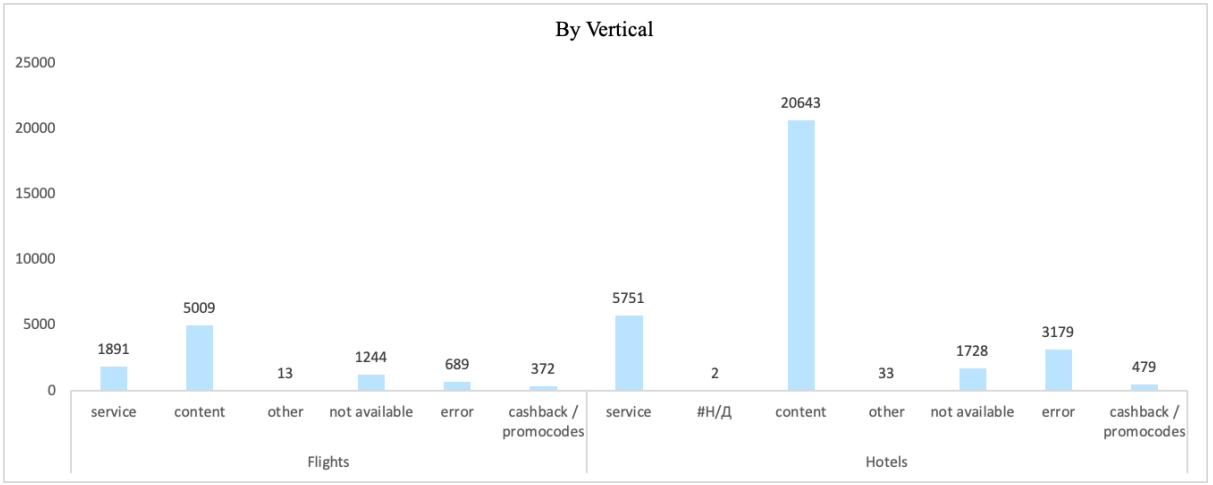
Appendix 4: Tableau dashboard. The number of programs per affiliate type (English websites)



Appendix 5: Tableau dashboard. The number of affiliates by Flights and Hotels vertical (Russian websites)



Appendix 6: Tableau dashboard. The number of affiliates by Flights and Hotels vertical (English websites)



Appendix 7: Python Notebook. Language detection function with pool

```
# language detection function with pool

def apply_parallel(texts, func, n_cores=2):
    pool = Pool(n_cores)
    split = np.array_split(texts, n_cores)
    res = [item for sub in pool.map(func, split) for item in sub]
    pool.close()
    pool.join()
    return res

def language_detect_text(text, lenght=100):
    try:
        res = detect(text[:lenght])
        return res
    except:
        return 'Not Detected'

def language_detect_texts(texts):
    return [language_detect_text(text) for text in texts]
```

Appendix 8. Python Notebook. NLP preprocessing function

```
def preprocessing(sentence, as_list=False):
    s = re.sub('[^а-яА-Яa-zA-Z]+', ' ', 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)
```

Appendix 9: Python Notebook. K-means clusterization with PCA

```
distances, distortions = [], []
n_clusters = range(1, 40)
scaler = StandardScaler()
pca = PCA(N_COMPS)
X = df[features].values
X = scaler.fit_transform(X)
if PCA_FLAG: X = pca.fit_transform(X)

for n_clu in tqdm(n_clusters):
    kmeans = KMeans(n_clusters=n_clu)
    kmeans.fit(X)
    distances.append(
        np.average(
            np.min(
                cdist(X, kmeans.cluster_centers_, 'euclidean'),
                axis=1
            )
        )
    )
    distortions.append(kmeans.inertia_)
```

Appendix 10: Python Notebook. CatBoost model function

```
model = CatBoostClassifier(
    task_type='CPU',
    verbose=0,
    loss_function='MultiClass',
    eval_metric='AUC'
)
grid = {
    'iterations': [100, 200, 300],
    'learning_rate': [.03, .1],
    'depth': [2, 4, 6, 8],
    'l2_leaf_reg': [1, 2, 3, 4, 5, 7, 9]
}
randomized_search_result = model.randomized_search(
    grid,
    X=X_tr,
    y=y_train,
    cv=3,
    plot=True,
    n_iter=40,
    verbose=True,
    refit=True
)
```