LAPPEENRANTA-LAHTI UNIVERSITY OF TECHNOLOGY LUT
School of Business and Management
Master in Business Analytics

Irina Makarkina

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

Examiners: Postdoctoral Researcher Christoph Lohrmann
Professor Pasi Luukka

ABSTRACT

University:	Lappeenranta-Lahti University of Technology LUT
Faculty:	School of Business and Management
Major:	Degree in Business Analytics
Author:	Irina Makarkina
Title:	In-Depth Analysis of Publishers in Travel Affiliate Marketing Based on Aviasales Data
Year:	2021
Master's Thesis:	83 pages, 22 figures, 22 tables and 0 appendices
Examiners:	Postdoctoral Researcher Christoph Lohrmann and Professor Pasi Luukka
Keywords:	Affiliate marketing, Machine Learning (ML), Natural Language processing (NLP), clustering, classification, Aviasales, travel industry

The understanding of the underlying interconnections between the parties involved in the affiliate marketing i.e. advertisers and affiliates (publishers) gradually becomes a requirement for managers, who aim to implement and develop successful affiliate marketing strategies. Nevertheless, the studies and business cases devoted to this phenomenon are limited. This Master's Thesis provides an in-depth analysis of the affiliates (publishers) part of the affiliate marketing based on the dataset of the Aviasales company.

In the literature review the notion and mechanism of affiliate marketing were discussed. Moreover, the overview of the affiliate marketing studies was presented together with the description of the previous attempts to categorize affiliates.

In the empirical part of the thesis methods of Machine Learning were applied. Thus, Natural Language Preprocessing was used to prepare affiliate websites data for further analysis. Clustering models (K-Means with PCA, DBScan) were applied to reveal the underlying data patterns. In terms of classification models (Gradient Boosting, CatBoost) the main types of affiliates were studied: content sites, service sites and cashback and promo code sites.

In terms of Aviasales data it was found that the content sites is the most widespread type of the affiliates. Moreover, cashback and promo code sites were defined as showing the most interest in participation in affiliate programs. Based on the findings, managerial implications and theoretical contribution of the work are given.

ACKNOWLEDGMENTS

I would like to thank all the staff from Saint Petersburg State University and LUT University, who helped me during my studies on a Double-Degree Program, especially Olesya S. Ustimenko and Kaija Huotari. Despite the pandemics, I was very excited to study in Finland this last year.

Also, I would like to say thank you to my scientific supervisor Postdoctoral Researcher Christoph Lohrmann for valuable and reasonable comments. I feel like you truly helped to improve the quality of this Thesis.

Finally, I am saying thanks with all my heart to my mother Nataliia Makarkina, my grandmother Margarita Petrovna Makarkina and all my dear friends, who constantly supported me on my study journey.

Best regards,
Irina Makarkina

LIST OF ABBREVIATIONS

HP – Hewlett-Packard

NLP – Natural Language Processing

DBScan – Density-based spatial clustering of application with noise

PCA – Principal components analysis

CatBoost – Categorical Boosting

CPC – Cost Per Click

CPA – Cost Per Action

UTM – Urchin Tracking Module

SEO – Search Engine Optimization

IAB – Internet Advertisin Bureau

IoT – Internet of Things

BERT – Bidirecional encoder from transformers

MLM – Masked Language Modelling

NSP – Next Sentence Prediction

NLTK – Natural Language toolkit

HTML – HyperText Markup Language

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. Introduction

1.1. Background

With the rapid development of modern technologies comes the rapid change in consumer behavior and importance of digital marketing increases. Thus, digitalization leads to more knowledgeable and demanding consumers, while abundance of information diminishes the span of consumers' attention (Nimmermann, 2020). This means that advertisers are looking for more ways to stand out in the digital environment. At the same time, consumers' trust related to opinions, expressed digitally in the form of personal blogs, reviews or forums, rises (Nimmermann, 2020). These factors contribute to the growing popularity of affiliate programs.

Affiliate marketing is the commission-based type of online marketing, which implies collaboration between an advertiser and an affiliate. An advertiser is the owner of the product or service that looks for a way to attract new customers and increase sales. An affiliate is a mediator between the advertiser and a customer that places the link to an advertiser's website on its own website in order to receive commission from the affiliate (Olbrich et al., 2019). A commission is usually paid when a customer takes some sort of an action: either clicks on the promoted link or buys an advertiser's product. For an advertiser, partnership with affiliates helps to generate traffic, increase sales and brand recognition (Mican, 2008).

According to a Forrester Consulting report (2016) 81% of brands around the world used affiliate marketing as a promotion tool in 2016. Among them were companies like e-commerce giant Amazon, computer software and hardware company Hewlett-Packard (HP) and sports outfit manufacturer UnderArmor (Forrester Consulting, 2016). Moreover, in 2016 Amazon's revenue from affiliate marketing was evaluated as 10 billion dollars (Prussakov, 2016). In 2015 the four most involved in affiliate marketing industries were fashion, health and beauty, sports and travel (Prussakov, 2015).

1.2. Research questions

The aim of this paper is to carry out an in-depth study of which characteristics the affiliates possess and how this knowledge can be used by advertisers based on Aviasales metasearch engine example. The term metasearch engine means that the company does not sell the tickets

directly but helps customers to find the best offers. Moreover Aviasales has its own affiliate marketing platform called TravelPayouts (Aviasales, 2021).

The main focus of the thesis will be on the following research questions:

1. In the context of Aviasales, which types of websites most often participate in the affiliate programs?
2. In terms of Aviasales data is there any specific pattern between the type of an affiliate and an industry of an advertiser?

1.3. Machine Learning application to the case

The usage of Machine Learning algorithms will allow to find answers to the proposed research questions. Thus, as the affiliates are the websites with textual information the application of Natural Language Processing (NLP) algorithms is required. Methods of NLP will combine the content from all the affiliates, help to extract keywords and transform text data in numerical form i.e. prepare the data for the analysis via supervised and unsupervised Machine Learning models.

Unsupervised models, for example, clustering, use unlabeled data in order to find the hidden patterns of the dataset. The application of this type of models allows to get the general outlook of the data and determine the clusters of the websites within the dataset. Identification of possible clusters is presumed to be useful for further managerial application analysis as different types of affiliates may require different incentives, for example, different payment mechanisms or different managerial approaches. Moreover, different clusters may present different levels of importance for Aviasales. Nevertheless, clustering techniques are expected to contribute to the answer to the first research question on the types of affiliates participating in Aviasales affiliate network.

Supervised learning models, for example, classification work with already labelled data. Application of these types of models will be useful to analyze Aviasales' already existing assumptions on how affiliates need to be grouped. Thus, supervised learning methods will give a different angle of the problem. Supervised learning methods will prevailingly contribute to the answer of the second research question on interrelations between affiliates and advertisers.

1.4. Structure of the thesis

In terms of this thesis, first, the notion and mechanism of affiliate marketing will be introduced and explained. Second, the literature review will be presented. Literature review section will discuss the main directions of affiliate marketing research as well as give the overview of previous attempts of affiliates' categorization. Third, the context of the research, namely, the peculiarities of affiliate marketing in the travel industry and brief overview of Aviasales company will be given.

Fourth, the methodology of the research will be presented. Thus, in terms of this thesis Machine Learning methods will be used. Therefore, the Natural language processing tools will be explained. After that, unsupervised learning clustering methods like K-Means clustering and Density-based spatial clustering of application with noise (DBScan) will be introduced. Moreover, Principal component analysis (PCA) as dimensionality reduction tool will be discussed. The thesis will further present supervised learning classification tools like Gradient Boosting and Categorical Boosting (CatBoost). As all the mentioned models will be created and executed via Python programming language its brief overview will also be presented.

Fifth, the overview of the dataset provided by Aviasales and the results of the application of the mentioned models will be given. Moreover, the results of the affiliates' analysis will be combined with Aviasales data on advertisers for further discussion of managerial applications of the research. Finally, the general conclusion will be given.

2. The notion and mechanism of affiliate marketing

Before considering 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 is paid for every visitor that comes to an advertiser website from hyperlinks published by this third party.

Thus, according to Dwivedi (2017) the three main participants of the process are:

1. Advertiser – a party that sells its products or services online;
2. Affiliate (a publisher) – an intermediary that uses its website or application to publish a hyperlink that leads to the advertiser's website.
3. Customer – an individual or a company that buys the product or service and, thus, generates revenue streams;

Therefore, schematically affiliate marketing can be presented in the following way:

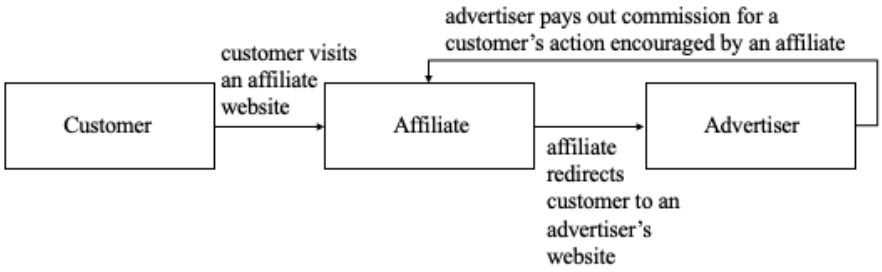


Figure 1. Affiliate Marketing Framework

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) (Olbrich et al., 2019). This means that commission is paid by an advertiser to an affiliate for a certain customer's activity: a click, a form fulfillment or an acquisition of a product. 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 an advertiser belongs. Different business domains have different customer acquisition costs that influence the amount of the affiliates' commission. According to Haq (2012), average commissions vary from 1 to 15% percent of each sale an affiliate helps to make.

Thus, in terms of such a scheme, the income of an affiliate largely depends on the traffic and leads its site generates to an advertiser's website. To track the traffic both advertisers and affiliates usually use instruments like cookies or Urchin Tracking Module (UTM) tags allowing to follow consumer actions (Gomer et al., 2013). Cookies can be defined as small fragments of textual data sent by a web server to a user's computer and kept there. They facilitate the process of users' profiles definition by saving the information about their personal preferences and,

thus, enabling targeted advertising (Smit et al., 2014). A UTM tag is a piece of code added to the end of a website link. This tool permits to identify a user's viewing session parameters i.e. a source a user came from to a website or a campaign that caught a user's attention (Semeradova and Weinlich, 2020). In general, tracking allows the advertiser to partially control the quality of the channel and manage commissions, while for an affiliate it presents an opportunity to personalize offers. Moreover, the process of retargeting, meaning advertising to users that showed an interest in a product, however have not bought it yet, is based on the collection of cookies (Semeradova and Weinlich, 2020). Thus, with the help of tracking tools affiliates can also return customers to their websites to further encourage clicks and purchases.

In general, it can be said that affiliate marketing in recent years has truly become one of the most widespread digital instruments. According to a Forrester Consulting report (2016), 81% of advertisers used affiliate marketing in 2016.

3. Literature review

3.1. Main directions of affiliate marketing research

Overall, affiliate marketing includes 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 and Brandi (2015) 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, may result in a loss of customers for the advertiser in the long term. Thus, the author states that when a customer buys through an affiliate the loyalty is formed for an affiliate and not for an advertiser and in case of a repurchase a customer returns to an affiliate website.

Attempts to give recommendations on affiliate marketing management were given by Ivkovic and Milanov (2010), who discussed general requirements for affiliate programs: modern software, constant technical support availability, adequate pricing and 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 is the most popular one in the study of Croatian hotels affiliate programs.

Moreover, Bhatnagar and Papatla (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 and Bhatnagar (2002) stated that affiliate programs bring the most results when the businesses of an advertiser and an affiliate align.

Part of the 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. Thus, Daniele et al. (2009) pointed out that an 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 an advertiser's brand representatives (Downs et al., 2008). 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 and Bhatnagar's (2002) study 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.

Therefore, though the affiliate marketing research has been quite modest it covered different directions. Nevertheless, it is important to note that most researchers relied either on case studies (e.g. Duffy(2005), Mican(2008)), questionnaires (e.g. Bhatnagar and Papatla (2001), Hossan and Ahammad (2013)) or interviews (e.g. Gregory et al., 2014). However, Machine Learning analysis that was not actively applied to the matter may result in a discovery of hidden affiliate marketing data patterns that will lead to further managerial conclusions beneficial for marketing professionals.

3.2. Approaches to affiliates categorization

Despite affiliates being one of the core parts of the affiliate marketing system, attempts to study, classify or group them are limited. Academic papers (e.g. Papatla and Bhatnagar (2002), Gregory et al. (2014)) only briefly mention possible ways to divide affiliates into manageable groups.

Thus, Goldschmidt et al. (2003) proposed to divide affiliates based on traffic generated by them. The following categories were described (Table 1):

Category name	Category description	Number of visitors per month
Hobby sites	Sites devoted to topics that usually represent an author's hobby, for example, travelling blogs. They contain a mix of relevant to the topic information as well as author's personal data and notes or posts.	less than 10 000

Category name	Category description	Number of visitors per month
Vertical sites	Sites devoted to one particular topic that does not represent an author's hobby, for example, a dating portal. They provide in-depth information on the subject and usually have a focused audience.	between 10 000 and 50 000
Super-affiliates	Relatively unfocused in terms of chosen topic sites, for example, a newspaper page. These sites try to appeal to a wider audience.	more than 50 000

Table 1. Goldschmidt et al. (2003) categorization of affiliates

Another classification developed by the Internet Advertising Bureau (IAB) (2016) also revolved around the issue of traffic, nevertheless, emphasized the different aspects. 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 in the following way (Table 2):

Category name	Category description
Reward sites	Affiliates that offer a reward or a bonus for a consumer that buys through its link
Content sites and blogs	Affiliates that provide unique content on the topics of the audience's interests

Category name	Category description
E-mail sites	Affiliates that actively use own databases and newsletters in order to attract consumers
Comparison sites	Affiliates that present mechanisms allowing users to compare offers on certain products or services
Retargeting sites	Affiliates that track visitors interests and actively use digital instruments to re-engage them
Pay-per-click sites	Affiliates that use custom landing pages and keywords to stimulate purchase
Voucher and deal sites	Affiliates that offer coupons or various discounts as a compliment for the purchase
Social sites	Affiliates that actively use social networks to generate traffic

Table 2. IAB (2016) categorization of affiliates

Both categorization approaches are limited and do not provide affiliate managers with the in-depth information. Thus, Goldschmidt's approach can be considered as quite general, while IAB's approach 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 these authors' own experience and knowledge of the field and are not supported by empirical studies. Moreover, both Goldschmidt and IAB do not take into account peculiarities of travel-themed websites.

4. Context of the research

4.1. Travel industry specifics of affiliate marketing usage

The concept of affiliate marketing is widespread among various industries. For example, the retail industry contributes 43% of the global affiliate promoting market income followed by telecom and media, travel and recreation areas, which contribute 24% and 16% respectively (Wang et. al, 2014). In terms of the present research affiliate marketing will be considered in the context of the travel industry. Thus, the following section will provide a brief overview of this domain.

The travel industry offers all types of assistance connected to movement of people from one place to another. According to MarketLine Industry Profile (2020) the following categories form the travel industry: hotels and motels, airlines, travel intermediaries, casino and gaming, passenger rail and foodservice. In detail, the hotels and motels segment implies all types of accommodation provided, while the airlines segment consists only of passenger air transportation and excludes air freight. Moreover, travel intermediaries are defined as companies that assist in selling accompanying travel products or services. An example of a travel intermediary is a car rental service or a travel agency. Casino and gaming segment includes all forms of gaming and betting, for example card games or roulette, with the exception of online services. Passenger rail is made up of all rail services including international and intercity trains. Foodservice covers food and beverages sold through restaurants, bakeries, pubs, clubs, bars as well as leisure venues, hotels and motels (MarketLine Industry Profile, 2020).

Travel industry is considered to be favorable for affiliate marketing programs development. Thus, according to Prussakov (2015), the travel industry is in the Top-4 most attractive and numerous affiliate domains. This can be explained by the industry's predisposition to affiliate program development.

Thus, the travel industry can be evaluated as extremely competitive as a lot of market players compete for the attention of the limited customer base. Moreover, the emergence of aggregators like Aviasales or Booking made information and comparison easily available for consumers and, therefore, additionally increased the power of buyers. Thus, competitors are constantly

looking for ways to attract customers, especially in a digital space and are prone to try affiliate marketing strategies.

Moreover, the technological factor has always been one of the most influential ones in terms of the travel industry. The way the customers interact with the companies in the travel industry is constantly changing along with the spread of digitalization. More and more travel services sales happen online. According to Bremner and Popova (2020), 73,6% of consumers use mobile devices or tablets for travel search purposes and 47% of consumers use a computer to purchase travel services. Moreover, 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 demand for players to be peculiar and visible online (Deloitte, 2020). Therefore, the business rapidly moves to digital and, thus, calls for online promotion strategies like affiliate marketing.

Travel industry is a service industry tightly connected with consumers' needs for pieces of advice and relevant information. As a result, various travel blogs as well as reviews and recommendations sites exist and continue to appear. As consumers tend to rely on them (Oktadiana and Kurnia, 2011) advertisers are more prone to include these sites in their own affiliate network.

All in all, the profitability, social aspects and the high level of digitalization makes the travel industry an attractive market for affiliates and advertisers.

4.2. Company overview

Aviasales is the largest Russian flight tickets metasearch engine (Aviasales, 2021) founded in 2007 by Konstantin Kalinov. The term 'metasearch engine' can be defined as a website that does not carry out its own indexing but rather combines and reorganizes the results of other search engines and provides a unified access to them (Meng et al., 2002). Therefore, the company does not sell any tickets itself but finds and compares the best offers. A user then can be redirected to an advertiser's website, where he/she can buy the tickets. Thus, a user receives

the information free of charge as the company profits from the advertisers' commissions (Chernikova, 2014).

Headquartered in Phuket the company has offices both in Moscow and Saint Petersburg (Chernikova, 2014). Aviasales actively operates on Kazakhstan's, Uzbekistan's, Belarusian, Ukrainian and Tanzanian markets (Kazmina et. al, 2020). Among its main competitors Skyscanner, Momondo, Kayak and Yandex.Avia can be named.

Moreover, in 2011 Aviasales launched its own travel affiliate marketing program called TravelPayouts (Baidin, 2018). In terms of this program Aviasales is in the role of the advertiser. Any website that is interested in receiving commissions from flight tickets and hotel booking is an affiliate. Moreover, since it is not obligatory to own a website to become a part of the program a broader term 'partner' is used. Partners include both affiliates and other participants that do not own a website, for example, Facebook groups owners. (Travelpayouts, 2021). Travelpayouts is a pay-per action affiliate scheme with tracking mechanisms based on cookies. Thus, when a consumer clicks on a partner's link, the affiliate's identifier gets written in a cookie file stored on a consumer's computer for a certain period of time i.e. lifetime of a cookie. Any purchases that consumer makes during the lifetime of a cookie, which in the travel industry usually amounts to 30 days, are attributed to the affiliate (Travelpayouts, 2021).

On average the affiliate can receive 1.1-1.5% commission from the value of the flight ticket sold and 4-5% from the value of a hotel booking. To the moment of this thesis being written Aviasales has already paid out 1 640 588 909 rubles (approximately equal to 18 831 369 euros or 22 801 792 US dollars) of commissions. (Travelpayouts, 2021)

Interesting to note that, starting as a founder's hobby the company now has entered the list of top 10 most expensive companies in the Russian Internet (Runet). Thus, in 2020 it was evaluated by Forbes as a 180 million dollars company (Kazmina et al., 2020).

All in all, Aviasales is an established travel industry player that has extensive knowledge and data related to travel affiliate marketing.

5. Methodology

5.1. Machine Learning

Machine Learning can be defined as the capacity of a computer to solve a particular problem based on the previous cases (Jo, 2021). Therefore, Machine Learning algorithms allow a machine to generalize, meaning to find solutions of real-life problems through the experience received from training examples. Machine Learning is considered to be a part of the Artificial Intelligence (AI) field, nevertheless it involves finding hidden patterns in the data and using those patterns to execute tasks like prediction or classification (Alpaydin, 2014).

Machine Learning includes supervised and unsupervised learning techniques. The main difference between these two domains lies in the presence or absence of labels in data (Kotsiantis, 2007). Thus, in the case of supervised learning the outcome is expected because the data is already labelled by a human. Moreover, supervised learning algorithms like classification are chosen according to the acceptable level of performance predefined by a researcher. Thus, in terms of supervised learning the data is usually split into train and test sets that allows a researcher to tune the model and teach it to achieve the required results. The learning stops when the machine reaches a certain proportion of correctly defined data attributes (Kotsiantis, 2007). Conversely, in unsupervised learning the data is unlabelled as the method is aimed at detection of hidden data patterns. Thus, unsupervised learning like clustering can help to find labels for further supervised learning (Hofmann, 2001).

Therefore, taking into account that affiliate marketing implies large amounts of data on affiliates and advertisers characteristics and generally presents a complicated marketing instrument, it can be concluded that Machine Learning tools can help find insights for building affiliate marketing networks and forming affiliate marketing strategy. Unsupervised learning methods will allow to find hidden patterns in the data and help to reveal possible ways of affiliates grouping for the supervised learning part. Supervised learning models will predict classes for all the affiliates in the dataset, which will facilitate further analysis and help formulate conclusions and managerial insights.

5.2. Natural Language Processing (NLP)

Natural Language Processing (NLP) is a field of Machine Learning. The term 'natural language' reflects the way a person expresses thoughts through a language. Thus, everything that is spoken, written, read or listened to is presented in the form of natural language (Thanaki, 2017). According to Manning and Schutze (1999), NLP can be defined as a computer method devoted to automatic recognition, comprehension and analysis of human language. This means that NLP tools allow to feed text data to a computer so that a machine could further process it in a manner similar to human thinking. Thus, through the application of NLP techniques the computer begins to find patterns and, therefore, 'to understand' the words and sentences.

NLP tools can be applied to a large number of tasks: speech recognition, translation, sentiment analysis, text classification etc. Thus, it is no wonder that in recent years NLP has transformed into a popular method of recognizing insights from textual data both in academic and business environments (Bansal et al, 2019). For example, Gabel et al. (2019) as well as Lee and Bradlow (2011) actively used NLP tools to analyze the market structure, while Das and Chen (2007) applied them to stock evaluation. In the field of marketing NLP is mostly used in a form of sentiment analysis as a form of feedback evaluation, search engine queries analysis as a key to consumer behavior understanding or descriptions and features analysis as a tool of new product development (Kang et al., 2020).

Vast field of NLP can be divided into two major directions: natural language understanding (NLU), meaning deciphering of the documents and their further processing, and natural language generation (NLG), implying the production of new textual data (Kang et al., 2020). In terms of this thesis NLU presents the main interest as the underlying idea is to find insights from already existing text data collected from affiliates websites.

NLP techniques allow to pre-process textual data for further analysis through tokenization, stemming or lemmatization and stopwords removal. Tokenization can be defined as division of text into separate sentences and division of sentences into separate words or minimal units that a computer can understand i.e. tokens (Bhavsar et al., 2017). This is usually done by finding white spaces (Hardeniya et al., 2016). Thus, for example, after tokenization a sentence 'I want cookies' is transformed into tokens 'I', 'want' and 'cookies'. Both stemming and lemmatization processes allow to shrink tokens to their stems. However, while stemming is

transformation of the word to its root via suffix removal, lemmatization is a more complex methodology that considers context and the nature of the word itself and applies different rules for different parts of speech (Hardeniya et al., 2016). Thus, lemmatization is able to detect the connection between, for example, words ‘good’ and ‘best’, while stemming cannot do so. Stopwords removal is also an important step of NLP preprocessing as it allows to reduce noise in the data. Stopwords present the most commonly used words that do not bear important meaning. Thus, an example of a stopword can be an article ‘the’ (Kang et al., 2020).

Moreover, after text preprocessing is done an important step for further manipulations is vectorization. Vectorization is the process that allows to represent textual data in a numeric way. The process can be performed with the usage of various techniques, however one of the most popular is term frequency-inverse document frequency (TF-IDF) methodology (Borad, 2020).

TF-IDF approach shows how many times a word is repeated in a document i.e. the relative importance of the word in a particular document. It can be calculated in several ways according to a weighting scheme: binary, raw count, term frequency etc.

$$\text{TF} = \left(\frac{f_{t,d}}{\sum_{t' \in d} f t', d}\right) \quad (1)$$

where t – a term (a word), d – a document or a website text, $f_{t,d}$– frequency of a term in a document,$\sum_{t' \in d} f_{t',d}$ – number of words in a document (Hamdaoui, 2019)

Inverse document frequency (IDF) shows whether a word is common for all the collection of documents and assigns lower weight for the words that are frequently used:

$$IDF = log\left(\frac{N}{count\ (\ d\ \in D : t \in d)}\right) \quad (2)$$

where t – a term (a word), d – a document or a website text, D – collection of all documents, websites.

Both TF and IDF represent fractions, thus, multiplication is needed to evaluate each word against every document (Bhavsar et al., 2019). Overall, the TF-IDF is calculated in the following way:

$$TF\ IDF\ =\ TF\ (t,d)\ *\ IDF\ (t,D)\ (3)$$

where t – a term (a word), d – a document or a website text, D – collection of all documents, websites.

Another vectorization method is Bidirectional encoder from transformers (BERT) vectorization, which is a part of BERT Neural Network created by Google (BERT documentation, 2021). BERT neural network is a model pre-trained on a combination of sentences that include 15% of masked or hidden from network words (MASK). This means that the network is fed by the sentences like 'I went to the [MASK] and bought [MASK]'. Thus, the network has to evaluate which words suit the proposed context, for example, 'mall' and 'clothes'. In more complex examples BERT 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). In terms of the BERT model the sentences are first tokenized i.e. broken down into separate words and parts of words or tokens via a pre-trained BERT tokenizer that is based on the WordPiece algorithm. Such an algorithm can divide a word into several subwords i.e. parts of a word. For example, in case of the word 'functionality' BERT tokenizer identifies the subwords 'function' and '##lity'. ## before a subword signifies that the algorithm considers this token to be a suffix (Yeung, 2020). BERT vectorizer allows to transform text data in numerical form (BERT documentation, 2021).

5.3. Cluster analysis

Cluster analysis belongs to the field of unsupervised Machine Learning techniques. It gives a general understanding of the dataset, searches for underlying data patterns and forms groups of related data objects i.e. clusters (Gnanadesikan, 1988). It is important to note that Jain (2010) defines the task of clustering algorithms as to find natural groupings of a set of data points. The actual definition of the notion 'cluster' causes controversy among the researchers, however, it

is prevailingly referred to as a collection of observations more similar to each other than to the rest of the data, meaning isolation and compactness (Jain, 2010).

Clustering algorithms are based on measures of similarity. Thus, as a measure of similarity the distance criterion has to be introduced. Though a number of options like Mahalanobis or Itakura-Saito distances exist, one of the most widespread methods of such calculation is Euclidean distance (Wu, 2012). Euclidean distance calculation in the N-dimensional space is the following:

$$d^2(u, v) = \sum_{k=1}^{N} (u_k - v_k) \quad (4)$$

where u and v are data vectors.

In the case of affiliates' analysis, clustering will allow to sort highly dispersed data in groups, and, therefore, help to detect the structure of the affiliate network. This will facilitate further research, help to detect unifying affiliates' characteristics and lead to a formulation of managerial insights.

As data partition can be considered a complex task several clustering approaches will be applied and compared. Namely, K-Means clustering and density-based spatial clustering of applications with noise (DBScan) are to be used.

5.3.1. K-Means clustering

K-Means clustering is considered to be one of the oldest and most popular clustering algorithms. Thus, the method was proposed by MacQueen (1967). According to Abbas (2008), the K-Means algorithm performs well on large datasets and shows better results than other clustering models. Moreover, due to its advantages such as ability to work with different types of data, robustness and simplicity it has been included by Wu et al. (2008) in Top-10 data mining algorithms.

K-means belongs to the group of so-called centroid cluster models (Wu, 2012). A centroid is a point that represents the center of a cluster (Pourahmad, 2020). The idea of the algorithm is that it attributes data points to clusters so that the squared error between the centroids and the

data points assigned to each centroid is minimized (Jain, 2010). Thus, mathematically K-Means algorithm can be presented by the following objective function:

$$J = min \sum_{k=1}^{K} \sum_{x \in C_k} \pi_x dist(x_t, m_k) \quad (5)$$

Where K – total number of clusters, $1 \leq k \leq K$, x_t – a data point from the dataset, $\{x_1, \ldots, x_n\}, \pi_x$ – the weight of x, m_k– the centroid of cluster C_k.

The assignment of data point x_t to the i-th cluster can be expressed through a membership function:

$$I(x_t, i) = \begin{cases} 1, \ if \ i = argmin\left(\left|x_t - m_j\right|\right)^2 \ j = 1, \ldots., k \\ 0 \ , otherwise \end{cases} \quad (6)$$

K-Means clustering is an iterative algorithm that includes a number of steps. First, the number of points K called centers or seeds has to be chosen by a user. One of the tools helping to define this parameter is the Elbow method. This method is based on the evaluation of the square of the error cost function. Thus, it is expected that as the number of clusters increases the degree of aggregation of each cluster also increases leading to decrease in the cost function. Therefore, a case when a researcher continues to rise the number of predefined clusters but the decrease in cost function stagnates serves as a stop indicator (Liu and Deng, 2021).

Second important step is the initialization of centroids for a selected number of clusters and initial data points division. As K-Means converges only to a local minima, initialization of centroids technique plays an important role in terms of the final results (Jain, 2010). Nevertheless, in most cases initial centroids are chosen randomly (Wu, 2012).

Thus, the first partition is created in which closest to centroids data points are attributed to certain clusters. According to MacQueen (1967) partition division is also based on within class variance. In case of a k-tuple x = $(x_1, x_2, \ldots, x_k)$, with x_i belonging to a random sequence of points E_N a minimum distance partition S(x) = $\{S_1(x), S_2(x), \ldots, S_k(x)\}$ of E_N is:

$$S_1(x) = T_1(x),\ S_2(x) = T_2(x)S_1^{'}, \ldots., S_k(x) = T_k(x)S_1^{'}(x)\ S_2^{'}(x)\ldots S_{k-1}^{'}(x)\ (7)$$

where:

$$T_i(x)\ =\ \{\xi: \xi \in\ E_N, |\ \xi - x_i| \leq |\ \xi - x_j|, j\ =\ 1,2,\ldots,k\}\ (8)$$

After this the centroids are updated continuously according to the following equation:

$$m_k = \sum_{x \in C_k} \frac{\pi_x x_i}{n_k}\ (9)$$

until the convergence criterion is met (Jain and Dubes, 1988).

Therefore, though the K-Means clustering algorithm is sensitive to outliers it is a widespread algorithm that works with large amounts of data. In the case of Aviasales the size of the data is large and since it is initially textual the problem with outliers is minimized, thus, the K-Means algorithm would be a suitable model for the analysis.

5.3.2. DBScan clustering

DBScan belongs to density-based clustering methods (Wang and Yu, 2001). These algorithms define clusters as dense regions of objects separated by regions of low density. In the context of clustering the term density is defined as the ratio between the number of data points contained within the ε- neighborhood and the volume of the resulting shape of the ε-neighborhood. ε- neighbourhood of a data point $x_i$can be described with the following formula:

$$N_\varepsilon(x_i)\ = \{x \in D\ |\ d(x_i, x) \leq \varepsilon\}\ (10)$$

where x_i, x – data points, d – distance.

Though the approach of K-Means and DBScan clustering differs, the latter method can also be applied to large datasets. Moreover, density-based clustering methods can detect arbitrarily formed clusters i.e. non-convex clusters and are less sensitive to outliers (Chadjipadelis et al., 2021). However, among the method's disadvantages is increase in required computational power compared to K-Means algorithm (Wang and Yu, 2001).

In the mechanism of DBScan the user does not need to predefine the number of clusters, however in the case of varying densities DBScan can return too few or too many clusters (Chadjipadelis, 2021). In addition to ε-neighbourhood parameter described earlier minPts is also defined. minPoints is a parameter that defines the minimum number of core points within each cluster. Therefore, if the amount of points in a datapoint ε- neighbourhood is no less than the number of minPoints, then this datapoint is called 'core'. There is only one core point in each cluster. Set of core points can be expressed in the following equation:

$$P = \{x \in D \mid ||N_\varepsilon(x_i)|| \geq minPoints\} \quad (11)$$

where P – set of core points, $N_\varepsilon(x_i)$ – ε- neighbourhood.

Moreover, further division of clusters is based on concepts of reachability and connectivity. Reachability evaluates whether the data point can be reached from another point. Connectivity shows whether data points belong to the same cluster (Celebi, 2015). Points reachable from a core point are called directly density-reachable points. If there is a chain of points reachable from a core point they are considered to be density reachable. Two points reachable from another point are density-connected. Points that are not density connected to any other points are considered to be noise points (Zheng et al, 2019). Thus, a cluster is formed if all points within it are reachable and mutually density-connected.

Thus, the DBScan algorithm includes the following steps. First, the algorithm initiates at a random data point x_i and its ε-neighbourhood is calculated. If the requirement of minPoints is satisfied in terms of this ε-neighbourhood, then the cluster C is formed. Otherwise, if the number of nearest points is insufficient, x_i is perceived as noise. At the same time the point perceived as noise can later become a member of another cluster if it is in the ε- neighbourhood of another point x_j.

If the point $x_i$is considered as core meaning it has more than specified number of minPoints in its ε-neighbourhood it means that the interior of a cluster is found. Thus, all density reachable points in x_i ε-neighbourhood are attributed to a cluster C together with their own ε-

neighbourhoods if they were also previously identified as core points. The process iterates until the whole density-connected cluster is found. After that the process restarts with a new data point $x_j$that is not yet attributed to any cluster or identified as an outlier (Wang and Yu, 2001).

In terms of affiliate websites' clustering DBScan technique will present another angle of the possible data division as it is useful in determining clusters of various shapes and sizes.

5.4. Principal component analysis (PCA)

Principal component analysis (PCA) is a dimensionality reduction approach, which allows to solve problem of redundant variables in complex datasets. With the help of PCA the most important variables in the data are identified and transformed into a set of new orthogonal vectors (Abdi and Williams, 2010).

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).

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. Next, eigenvector analysis on covariance matrix allows to find the structure of the most important features i.e. determine principal components (Harrington, 2012).

If A is a matrix of size n then v is a nonzero eigenvector of matrix A if there is the following relation:

$$Av = \lambda v \quad (12)$$

λ represents the values for which matrix A transforms the vector v into a collinear vector i.e. eigenvalues. Eigenvalues are scalar values (Vlase et. al, 2019).

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 (Vlase et. al, 2019). It means that in terms of PCA a set of p features of n units is converted into $r \leq p$ uncorrelated features i.e. principal components. Principal components are uncorrelated and the first component explains the most variance while the variance explained by each next component decreases (Jaadi, 2021).

Mathematically, if matrix X is a matrix of n observations on p features ($n \times p$) in terms of PCA a user aims to transform X to a set of uncorrelated features that can explain the maximum possible variance:

$$Z = X \times B \quad (13)$$

$$B = p \times p \quad (14)$$

Therefore, first, if $b_1 = p \times 1$ is the first column of B so that $z_1 = Xb_1$ a user wants to achieve $\max\{z_1^T z_1 = b_1^T X^T X b_1\}$, such that $b_1^T b_1 = 1$. This is solved through the following equation:

$$S = b_1^T X^T X b_1 - \lambda_1 (b_1^T b_1 - 1) \quad (15)$$

resulting in:

$$X^T X b_1 = \lambda_1 b_1 \quad (16)$$

where λ_1 is the eigenvalue of $X^T X$ and b_1 is the eigenvector. The same procedure is repeated for other columns of B. As a result a user gets a set of eigenvectors of $X^T X$ $B = [b_1, \dots b_p]$ in decreasing λ_i order (Abdi and Williams, 2010).

Overall, in terms of this thesis, PCA is considered to be a useful tool that will allow to simplify the data analysis through reduction of initial data dimensions and initial variables transformations.

5.5. Classification models

5.5.1. Gradient boosting classifier

Gradient boosting classifier belongs to the ensemble learning techniques. Ensemble learning is a combination of multiple Machine Learning algorithms in terms of one new model. It can be

done through mixing train data, mixing combinations or mixing models. Thus, boosting models are based on mixing combinations techniques as in terms of this approach more emphasis is given to improvement of models that show poor classification results i.e. weak learners. It means that new models are sequentially added to correct the errors of already existing models (Kumar and Jain, 2020).

In terms of boosting methods cost function is defined to measure the performance. This function includes two parts: training loss and regularization. Training loss function L(θ) shows how predictive the model is on training data while regularization term $\Omega(\theta)$ controls the level of the model's complexity i.e. helps to avoid overfitting (Bartlett, 1998).

Most commonly used loss functions are mean squared error and logistic regression (Bowd et al., 2020). Moreover, there are several types of regularization: Lasso (L1) regularization, Ridge (L2) regularization and elastic net, which is the combination of these methods:

$$L1 = \alpha \times \sum_j |\theta_j| \quad (17)$$

$$L2 = \lambda \times \sum_j \theta_j^2 \quad (18)$$

$$\text{Elastic net} = \alpha \times \sum_j |\theta_j| + \lambda \times \sum_j \theta_j^2 \quad (19)$$

The difference between Lasso and Ridge regularization is that Lasso shrinks coefficients of the less important features to 0 (Bowd et al., 2020).

Gradient boosting classifier is aimed at continuous increase of weak learners' performance through calculation of residual errors. Thus, residual error of each prior classifier is used to train the next model in the ensemble (Bowd et al., 2020). Moreover, gradient boosting is based on gradient descent algorithm. The pseudo code of the gradient descent method is the following:

1. Parameters are initialized randomly
2. The gradients G of the loss function are calculated in accordance with the parameters
3. The parameters are updated by a chosen learning rate, which determines the size of the steps needed to reach minimum
4. The algorithm is repeated until the loss function stops reducing or termination criteria is met (Bowd et al., 2020).

According to Hastie et al. (2009) gradient boosting models are usually made up from decision trees. As decision trees are used both for regression and classification, decision trees used for classification purposes are referred to as classification trees. As all decision trees, classification trees are formed by nodes and leaves. A node represents a certain characteristic and, thus, splits the data into two or more subsets, while each leaf represents a class (Maimon and Rokach, 2014). Classification trees' approach to tackling Machine Learning tasks includes creation of rules by finding out underlying statistical patterns and relationships within the data. In general, classification trees take into account the information about data distribution and split the data into subsets with each subset being more homogeneous than the previous one. This iterative process is called recursive partitioning. As a result of recursive partitioning the sequence of nodes and thresholds of variables are obtained (Maimon and Rokach, 2014).

The model can be further improved through randomized search — a technique that allows to configure an optimal set of parameters. Randomized search tests random combinations of possible model parameters and selects the best options (Bartlett, 1998).

5.5.2. Categorical Boosting (CatBoost)

Categorical Boosting (CatBoost) is a Machine Learning algorithm for gradient boosting based on decision trees (CatBoost documentation, 2021). It was initially developed by engineers of Russian Information Technology (IT) company Yandex to improve the quality of the Yandex search engine (Dorogush et al., 2018).

The model executes a unique algorithm representing variation of gradient boosting technique. Thus, the trees that the model is made of are binary and symmetrical. It means that at each level the data are compared in the same way to the same feature with the same values (Razrabotka, 2017). An arbitrary tree used in CatBoost looks as following (Figure 2):

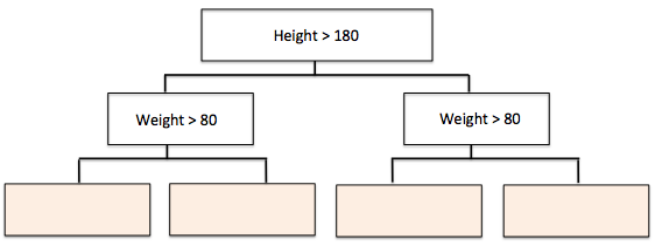


Figure 2. An arbitrary tree used in CatBoost Algorithm

The main peculiarity of CatBoost is that it is able to work with categorical variables as it includes a one-hot encoding technique (Yandex, 2017). One-hot encoding transforms categorical variables with multiple values into features, where each value is represented by a column of all 0 except one 1 (Li et al., 2018). Moreover, CatBoost algorithm is less prone to overfitting due to a specific formula of leaf value calculation:

$$\text{leafValue (doc)} = \sum_{i=1}^{doc} \frac{g(approx(i), target(i))}{docs\ in\ the\ past} \quad (20)$$

For each object the leafValue is calculated as the average gradient of all objects in the list that were in the leaf before a certain object (Razrabotka, 2017).

5.5.3. Classification table and classification metrics

The output of any classification model is the list of labels predicted, which can be both correct or incorrect. Thus, in order to evaluate the quality of the prediction the labels predicted by a model are compared with the actual labels of the dataset in the classification table. In a binary case the classification table looks as following (Table 3):

Classification table		TRUE	
		Condition positive	Condition negative
Predicted	Predicted positive	True positive	False positive
	Predicted negative	False negative	True negative

Table 3. Binary classification table

Thus, if the class is predicted as positive and is actually positive the prediction is called True Positive (TP). If the class is predicted as positive and is actually negative the prediction is called False Positive (FP). If the class is predicted as negative but actually is positive then the prediction is called False negative (FN) (Herrera et al., 2016).

Accuracy or R^2score is the ratio between the number of correctly predicted labels and the total number of observations (Herrera et al., 2016):

$$Accuracy = \frac{number\ of\ correct\ predictions}{total\ number\ of\ observations} = \frac{TP+TN}{TP+TN+FP+FN} (21)$$

Precision shows how many selected observations are relevant (Herrera et al., 2016):

$$Precision = \frac{TP}{TP + FP} (22)$$

Recall shows how many relevant observations are selected (Herrera et al., 2016):

$$Recall = \frac{TP}{TP + FN} (23)$$

F-score metric considers both precision and recall and calculated as a harmonic mean of these metrics (Herrera et al., 2016):

$$F_\beta = (1 + \beta^2) \times \frac{Precision \times Recall}{(\beta^2 \times Precision) + Recall} (24)$$

where $\beta^2 -$ weight of the importance of precision.

In the case of multiclass problem the classification table looks as following (Table 4):

		TRUE		
		Class A	Class B	Class C
Predicted	Class A	TPa	Eba	Eca
	Class B	Eab	TPb	Ecb
	Class C	Eac	Ebc	TPc

Table 4. Multiclass classification table

TPa is the true prediction of class A, Eba is the error of predicting class B as class A. In this case accuracy remains the ratio between the correctly predicted labels and the total number of observations:

$$Accuracy = \frac{number\ of\ correct\ predictions}{total\ number\ of\ observations} = \frac{TPa+TPb+TPc}{TPa+Eba+Eca+Eab+Ecb+Eac+Ebc} (25)$$

Recall and precision are calculated with respect to each class:

$$Precision = \frac{TPa}{TPa + Eba + Eca} \quad (26)$$

$$Recall = \frac{TPa}{TPa + Eab + Eac} \quad (27)$$

Additionally, macro and weighted average metrics are introduced. Macro-average is the averaging of the unweighted mean calculated for each separate class, while weighted average is the support-weighted mean calculated for each separate class (Herrera et al., 2016):

$$Weighted\ average = w \times class\ A + w \times class\ B + w \times class\ C \quad (28)$$

$$Macro\ average = 0.33 \times class\ A + 0.33 \times class\ B + 0.33 \times class\ C \quad (29)$$

5.6 Python programming language

The Machine Learning concepts described will be programmed in Python programming language with the usage of Jupyter Notebook.

Thus, though NLP tools can be used via different programs and languages, the Python environment is considered to be one of the best options of their implementation (Thanaki, 2017). Thus, Python represents an easy-to-use and intuitively understandable platform that allows fast development and testing (Thanaki, 2017). Moreover, it contains a large number of open source packages, including popular natural language toolkit (NLTK) and BeautifulSoup libraries (Hardeniya et al., 2016).

BeautifulSoup allows users to perform web scraping and get data from the websites through HyperText Markup Language (HTML) parsing. Content of websites are loaded into a BeautifulSoup object and an HTML parser is applied to it. As a result a soup object containing the text and HTML tags that need to be removed is created. After cleansing only the stripped content of the website remains and is ready to use (Bhavsar et al., 2017).

Langdetect Python package based on Google language detection library allows to identify websites' languages and select only those that present interest in terms of the research (Pypi.org, 2021).

Natural Language Toolkit (NLTK) package allows to carry out all the steps of text preprocessing, which includes tasks like tokenization (nltk.word_tokenize(sentence)), stemming (PorterStemmer.NLTK_EXTENSIONS) and stopwords removal (nltk.corpus.stopwords) (nltk.org, 2021).

In Python TF-IDF algorithm is implemented via sklearn library (TfidfVectorizer), while BERT model is included in transformers package (BertTokenizer, BertModel).

Other concepts described will be implemented with the help of the following Python libraries (Table 5):

Machine Learning model	Python package
K-Means clustering	sklearn.cluster.KMeans
DBScan clustering	sklearn.cluster.DBSCAN
PCA	sklearn.cluster.PCA
Gradient Boosting classifier	sklearn.ensemble.GradienBoostingClassifier
CatBoost	CatBoost open library

Table 5. Python libraries used for Machine Learning models implementation
(Source: Scikit-learn.org, 2021; CatBoost.ai, 2021)

6. Application of Machine Learning concepts to Aviasales data

6.1. Aviasales dataset

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

The main raw dataset includes solely the information on the affiliate websites (Table 6):

Row	url	flag
1	bpponline.ru	direct advertiser
2	amondo.holiday	direct advertiser
3	akvaplan.com	direct advertiser
4	castrlnaurivierebasse.ft	direct advertiser
5	calnboard.ru	direct advertiser

Table 6 Example of the data in the initial dataset

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 provided data the company uses its internal classification from the viewpoint of the affiliate network owner that does not imply the same meaning of term advertiser that was described in the literature review.

The auxiliary dataset includes information on the vertical (industry), advertiser and affiliates that promote a certain advertiser (Table 7):

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 7. Excel dataset provided by Aviasales

The vertical means the type or a niche of an advertiser. Both terms ‘advertiser’ and ‘affiliate’ are used in accordance with definitions presented in the literature review. Nevertheless, 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 main dataset.

After parsing and langdetect were applied it was found that the main dataset includes 52 unique languages with English and German being the most popular ones. In terms of this thesis English and Russian datasets present the main interest, therefore they were selected for the further research. The number of the English and Russian sites in the dataset are 67 490 and 6999 respectively (Figure 3):

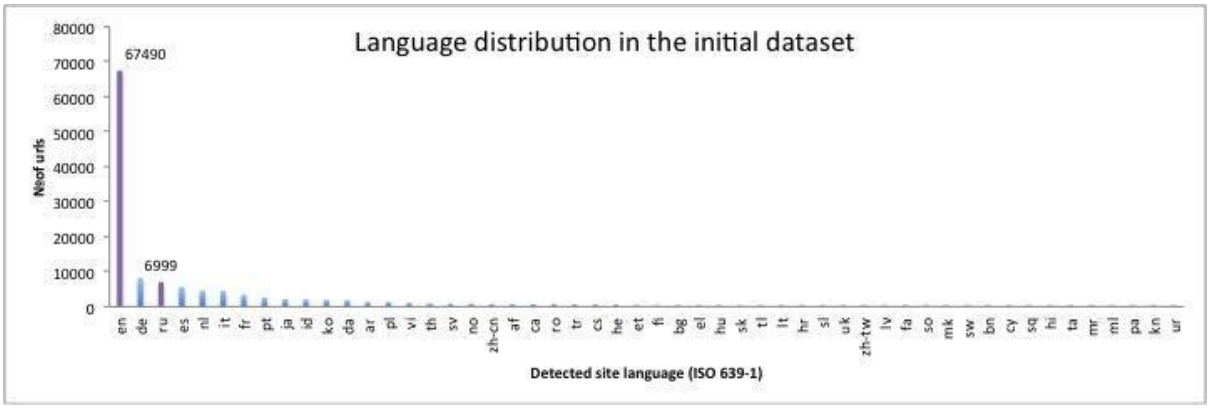


Figure 3 Language distribution in the initial dataset

6.2. Data preprocessing and vectorization for unsupervised learning

After parsing, application of the langdetect package and selection of Russian and English websites the dataset looked as following (Table 8):

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.. (Create free forum on MyBB)	ru

Table 8 Dataset after langdetect package application

On this stage several challenges were discovered. A number of sites were not available or deleted. This could have presented challenges for the unsupervised learning part of the research as the noise could interfere in the clustering process. Therefore sites containing words like ‘access denied’, ‘redirect’, ‘error’, ‘no longer exists’, ‘temporarily unavailable’ were searched for and dropped.

Moreover, 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, for example .nl or .fr , were dropped from the data.

NLP preprocessing for Russian websites included the following stages (Table 9):

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 9 NLP Russian language websites data preprocessing

Also, important to note that among the peculiarities of the Russian language is the usage of the letter ‘ё’, which sometimes appears in texts. According to the norms of the Russian language in the Internet environment, its usage is optional (Pakhomov, 2010). Therefore, several authors can write the same word in a different way, for example, ‘лёд’ or ‘лед’, which in both cases

means ‘ice’. However, in the case of NLP analysis these two variations of one word can be mistakenly treated as two separate tokens. Therefore, in order to standardize the tokens ‘ё’ is replaced by ‘e’ in all cases. Moreover, as Russian language websites were further vectorized with TF-IDF method custom tokenization was omitted.

Based on the unique meaning of the words and their weight assigned via TF-IDF method. The total text was divided into 29 018 features and the following matrix is obtained (Table 10):

Raw	URL	Flag	Text	Language	Proc	акция (share)	бизнес (business)	билет (ticket)
0	flyticket.ru	direct advertiser	Travelmart: заказ (order)...	ru	‘travelmart’, ‘заказ’, ...	0	0.088393	0
1	vizitkaplus.ru	direct advertiser	Срок регистрации… (Period of registration...)	ru	‘срок’ ‘регистрация’...	0	0	0
2	airport63.ru	direct advertiser	Выбрать маршрут… (Choose an itinerary)	ru	‘выбрать’ ‘маршрут’ ...	0.028735	0.025992	0.047775

Table 10 Russian language dataset after TF-IDF vectorization (extract)

English language websites were pre-processed in a similar manner, however additionally BertTokenizer was applied. Therefore, for English websites NLP preprocessing looked as follows (Table 11):

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

Process	Process description	Initial text	Results
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 11 NLP English language websites data preprocessing

BERT vectorization was applied to English language websites. 768 features were obtained and the English language dataset was transformed into following table (Table 12):

Raw	URL	Flag	Text	Language	Proc	feature 0	feature 1	feature 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 advertised	Traveling Sisters	en	‘traveling’ ‘sisters’ ‘skip’ ‘content’	0.275895	-0.272740	0.826218

Table 12 English language dataset after BERT vectorization (extract)

6.3. Results of data clustering

6.3.1. Russian language websites clustering

As the transformed dataset included 29 018 features PCA was applied to reduce the dimensionality of the data. 58 components explaining 90% of variance were implemented. The principal components' individual and cumulative variance are presented in the following graphs (Figure 4):

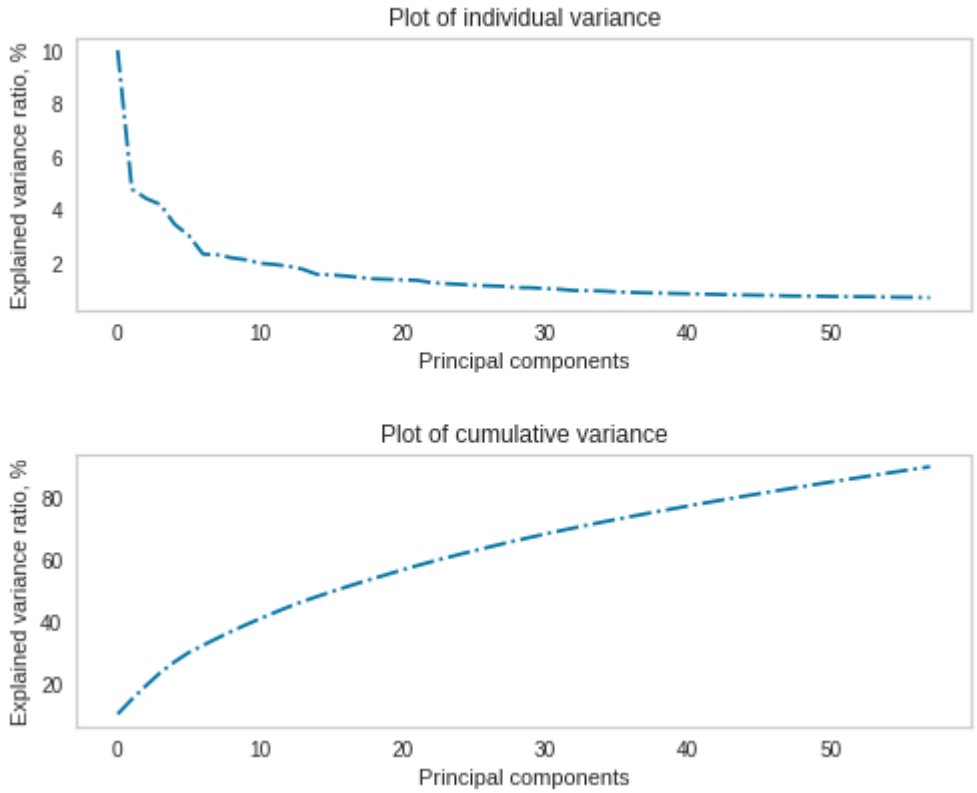


Figure 4 Plots of principal components' individual and cumulative variance in the Russian language dataset

For K-Means clustering the iteration was performed on the range from 0 to 40 possible clusters. Too many clusters makes further analysis and description confusing. Moreover, it leads to the emergence of clusters with solely one or two websites. The final choice of number of clusters selected was based on the elbow curve, which showed number of clusters k = 18 as the adequate quantity (Figure 5):

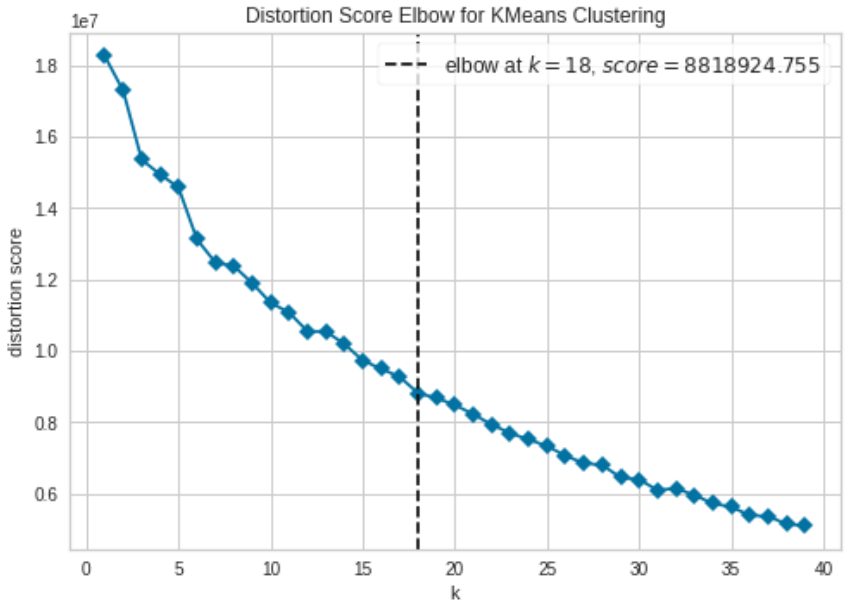


Figure 5 Elbow curve for Russian language websites

The 18 clusters are summarized in Table 13. The count starts from 0 as in the Python programming language in order to facilitate the perception of the further graphs.

Cluster №	Cluster name	General description	№ of websites	Key words examples	Website examples
0	Miscellaneous sites	Sites that were not attributed to any smaller cluster. Include both travel-related sites like airports and variously-themed sites like video games or roller-skating clubs	4694	тур (tour), отель (hotel), отдых (rest), город (city), путешествие (journey)	http://indiya.india-goa.today/ (India tourist portal) aeroport-ulyanovsk.ru (Ulyanovsk airport) roller.ru (roller skating club)
1	Real estate	Buying and selling real estate abroad	7	аренда (rent), вилла (villa), недвижимость (real estate), франция (France)	comodo.ru (real estate agency)

Cluster №	Cluster name	General description	№ of websites	Key words examples	Website examples
2	Travel blogs	Blogs, where users share their personal traveling experience	6	понравиться (to like), машина (car), путешествие (journey), блог (blog)	flowerkoi.ru beavoyager.com
3	Aviabilet travel agency	Aviabilet's travel agency multiple domains	11	авиабилет (flight ticket), дешевый (cheap), авиалиния (airline), воздух (air)	aviabilet.gr aviabilet.ae
4	Affiliate and partnership programs	Various marketing and partnership programs	8	маркетинг (marketing), партнерский (partner), оффер (offer), рекламодатель (advertiser)	storader.com (partnership program)
5	Message boards for Russians living abroad	Message boards that allow to buy or sell various goods from people of Russian-speaking community living abroad	7	регион (region), услуга (service), помощь (help), работа (work)	doska-de.ru doska-cz.ru
6	Coupons / promo codes sites	Sites offering discount via coupons	3	купить (buy), медицинский центр (medical center), лазерный (laser)	kupikupon.com.ua
7	Airports imitating sites	Sites usually containing word 'airport' in the domain, however just advertising Aviasales	25	авиабилет (flight ticket), рейс (flight), авиакомпания (airline), время (time)	irkutsk-airport.ru (real Irkutsk airport: https://iktport.ru/ru/)

Cluster №	Cluster name	General description	№ of websites	Key words examples	Website examples
8	Travel portals	Websites devoted to travelling to a particular place with information on tourist attractions, hotels, transport and allowing to find a travelling buddy etc	22	остров (island), страна (country), карта (map), америка (America)	planetolog.ru netpoputchika.ru
9	City portals	Sites aimed at citizens and not tourists. Usually with message boards and forums	213	россия (Russia), вакансия (vacancy), новость (news), резюме (CV)	kaliningradlife.ru (Kaliningrad city portal)
10	Abandoned blogs I	Blogs that instead of posts contain a large number of advertisers' links	4	купить (buy), крем (cream),скидка (discount), похудение (weight loss), доставка (delivery)	seliger-2008.blogspot.com
11	Wikipedia	Wikipedia library and its domains	3	избранный (favorite), статья (article), медиа (media), премия (prize)	ru.wikipedia.org wikiredia.ru
12	Aviapoisk metasearch engine	Aviapoisk and its domains	6	авиабилет (flight ticket), авиакомпания (airline), предложение (offer)	aviapoisk.uz aviapoisk.kg
13	Blogs and personal websites	Blogs and personal websites devoted to various topics including travelling	1506	сайт, читать, новость, свой	chechundra.ru (blog about technologies) danilsnitko.ru (personal blog)

Cluster №	Cluster name	General description	№ of websites	Key words examples	Website examples
14	Aviasales mirrors	Sites that mimic the names of Aviasales competitors' websites as well as domains that include misprints of the word 'Aviasales'	101	авиабилет (flight tickets), москва (Moscow), санкт, петербург (Saint Petersburg), крым (Crimea)	scyskaner.ru (imitation of Skyscanner.ru) afiasales.ru (aviasales.ru misprint)
15	Local city news sites	Sites with the latest information on what happens in various Russian cities	11	россия (Russia), человек (human), женщина (woman), covid	https://ngs.ru/ (Novosibirsk online)
16	Potentially fraudulent sites	Sites offering to buy domains that imitate the names of the real housing projects	135	скидка (discount), бесплатный (free), официальный (official), банк (bank), защита (protection)	заречный-квартал.рф (uses the name of Zarechniy Kvartal complex, which official website is https://za-kvartal.ru/)
17	Abandoned blogs II	Empty sites, presumably, based on the hosting name, abandoned travel blogs that now only contain a link to Aviasales	228	авиабилет (flight ticket), москва (Moscow), билет (ticket), электронный (electronic)	cbultoz.blogspot.com floclood.blogspot.com

Table 13 K-means clustering. Description of Russian language clusters

The projection of these clusters on 2-dimensional space can be presented as Figure 6:

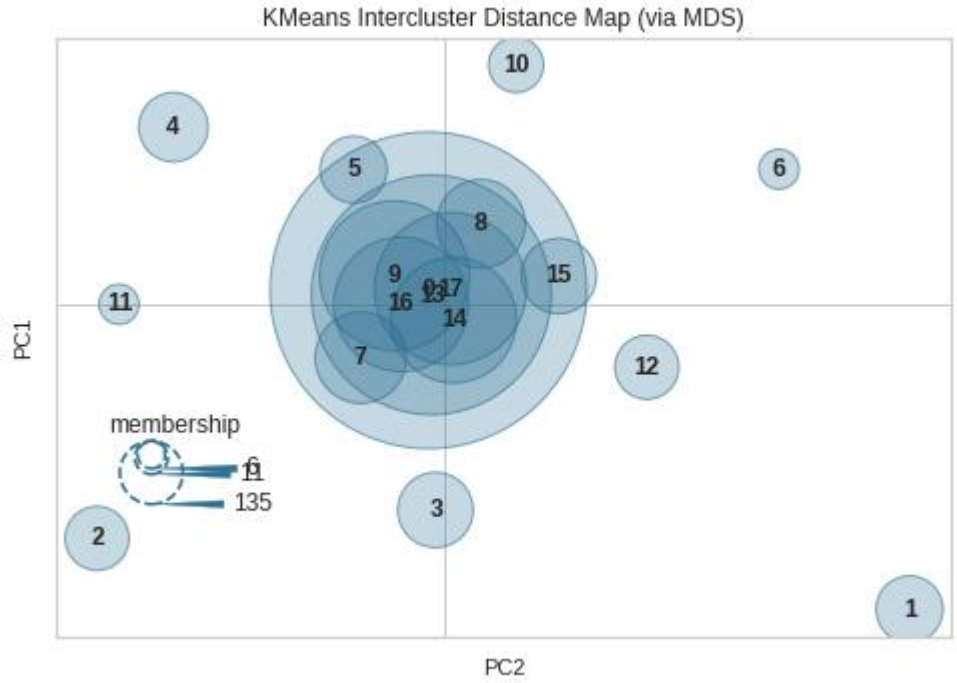


Figure 6 Russian websites intercluster distance map

It can be seen that cluster 0 (Miscellaneous sites) is the largest one and it overlaps with several other clusters. This is explained by the fact that cluster 0 contains various types of websites that the model could not attribute to any other cluster due to similarity in key words. Conversely, clusters like 11 (Wikipedia), 3 (Aviabilet travel agency) and 12 (Aviapoisk metasearch engine) can be considered as too narrow as they basically include various domains of the same site. Clusters 1(real estate), 4 (affiliate and partnership programs), 6 (coupons and promo codes sites), 10 (abandoned blogs II) also stand out, however they as well consist of a very limited number of sites. The model was not able to fully divide class 13 (Blogs and personal website). Thus, this cluster includes travel blogs as well. The model is also confused by the presence of the sites imitating other sites: cluster 7 (Airports imitating sites) and cluster 16 (Potentially fraudulent sites).

DBScan approach is presumably able to show another angle of the data. To be able to compare the results with K-Means the minPoints is chosen to be equal with the minimum number of sites in K-Means clusters, which is 3 (in K-Means cluster 6 and 11).

Next ε- neighbourhood is identified through the distances between clusters and minPoints value of 3 presented in Figure (7):

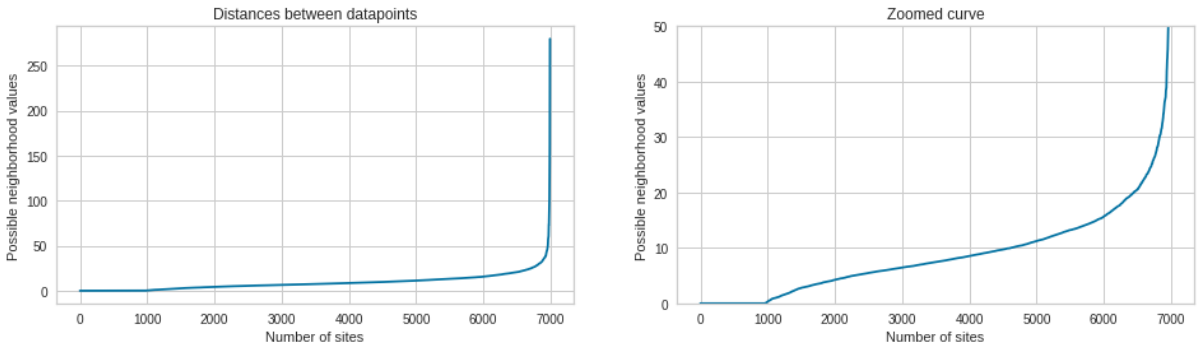


Figure 7. ε- neighbourhood identification through distance graph

The suitable value of ε- neighbourhood is on the slope of the curve, which is between 10 and 20. Thus, judging by zoomed curve ε- neighbourhood value is 16. Thus, with minPoints = 3 and ε- neighbourhood = 16 DBScan as well as K-Means identified 18 clusters (Table 14).

Cluster №	Cluster name	General description	№ of websites	Key words examples	Website examples
0	Miscellaneous sites	Sites that were not attributed to any smaller cluster. Include both travel-related sites like airports and variously-themed sites like video games or roller-skating clubs	4934	тур (tour), отель (hotel), отдых (rest), город (city), путешествие (journey)	http://indiya.india-goa.today/ (India tourist portal) aeroport-ulyanovsk.ru (Ulyanovsk airport) roller.ru (roller skating club)
1	MyZafira autoclub	MyZafira autoclub domains	4	форум (forum), zafira, opel	myzafira.com myzafira.ru

Cluster №	Cluster name	General description	№ of websites	Key words examples	Website examples
2	Abandoned blogs II	Empty sites, presumably, based on the hosting name, abandoned travel blogs that now only contain a link to Aviasales	157	авиабилет (flight ticket), москва (Moscow), билет (ticket), электронный (electronic)	cbultoz.blogspot.com floclood.blogspot.com
3	Sanatoriums	Sanatoriums in Russia, where people can rest and treat their health as well as visit mud bath or pool	7	санаторий (sanatorium), лечение (medical treatment), номер (room), кисловодск (Kislovodsk)	sanatoriy-rus-essentuki.ru sanatoriy-rodnik.ru
4	Hello South website	Hello South website and its domains	3	дом (house), анапа (Anapa), гостевой (guest), алушта (Alushta)	privetyug.ru privetyug.com
5	Potentially fraudulent sites	Sites offering to buy domains that imitate the names of the real housing projects	135	скидка (discount), бесплатный (free), официальный (official), банк (bank), защита (protection)	заречный-квартал.рф (uses the name of Zarechniy Kvartal complex, which official website is https://za-kvartal.ru/)
6	Aviasales mirrors	Sites that mimic the names of Aviasales competitors' websites as well as domains that include misprints of the word 'Aviasales'	101	авиабилет (flight tickets), москва (Moscow), санкт, петербург (Saint Petersburg), крым (Crimea)	scyskaner.ru (imitation of Skyscanner.ru) afiasales.ru (aviasales.ru misprint)

Cluster №	Cluster name	General description	№ of websites	Key words examples	Website examples
7	YandexNews portal	YandexNews portal and its domains	3	сша (the USA), новость (news), коронавирус (Covid), ситуация (situation)	yandexnews.by newsyandex.by
8	Airports imitating sites	Sites usually containing word ‘airport’ in the domain, however just advertising Aviasales	25	авиабилет (flight ticket), рейс (flight), авиакомпания (airline), время (time)	irkutsk-airport.ru (real Irkutsk airport: https://iktport.ru/ru/)
9	Not available sites	Sites, hosting of which has expired	39	сервис (service), срок (time), регистрация (registration), домен (domain)	kupimbilet.ru ticketall.ru
10	City portals	Sites aimed at citizens and not tourists. Usually with message boards and forums	211	россия (Russia), вакансия (vacancy), новость (news), резюме (CV)	kaliningradlife.ru (Kaliningrad city portal)
11	Abandoned blogs II, part 1	Empty sites, presumably, based on the hosting name, abandoned travel blogs that now only contain a link to Aviasales	31	авиабилет (flight ticket), москва (Moscow), билет (ticket), электронный (electronic)	cbultoz.blogspot.com floclood.blogspot.com
12	Abandoned blogs III	Empty sites, presumably, based on the hosting name, abandoned personal blogs that now only contain a link to Aviasales	4	авиабилет (flight ticket), москва (Moscow), билет (ticket),	boburintaras.blogspot.com tsvitkovevgen.blogspot.com

Cluster №	Cluster name	General description	№ of websites	Key words examples	Website examples
13	Wikipedia	Wikipedia library and its domains	3	избранный (favorite), статья (article), медиа (media), премия (prize)	ru.wikipedia.org wikiredia.ru
14	City portals II	Sites aimed at citizens and not tourists. Usually with news and weather on the front page	19	погода (weather), информация (information), найти (find), портал (portal)	alchevsk.ua
15	Delo.ru business portal	Delo.ru business portal	3	ростов (Rostov), конференция (conference), новость (news)	prokat-krasnodar.ru deloru.ru
16	Abandoned blogs II, part 2	Empty sites, presumably, based on the hosting name, abandoned travel blogs that now only contain a link to Aviasales that were not attributed to Abandoned blogs II cluster	14	авиабилет (flight ticket), москва (Moscow), билет (ticket), электронный (electronic)	aslofetop.blogspot.com sapoeohe.blogspot.com
17	Forums created through Mybb hosting	Variously themed forums	19	форум, создать, mybb, бесплатный	animebb.ru 9bb.ru

Table 14 DBScan clustering. Description of Russian language clusters

Important to note that the number of websites in DBScan cluster is not equal to initial number of affiliates as DBScan can consider those sites that do not belong to any cluster as outliers and, thus, remove them. However, DBScan Clustering results are relatively similar to those of K-Means. Thus, a number of clusters like miscellaneous websites, city portals, Wikipedia, airports imitating sites, potentially fraudulent sites and Aviasales mirrors are detected by both

algorithms. Nevertheless, DBScan was also able to detect a cluster of sanatoriums and forums created on Mybb platforms. Both algorithms were prone to forming small clusters that included only one site with several variations of its domain.

6.3.2. English language websites

For the English dataset the number of principal components chosen was 16. 16 components explain 86% of the variance (Figure 8):

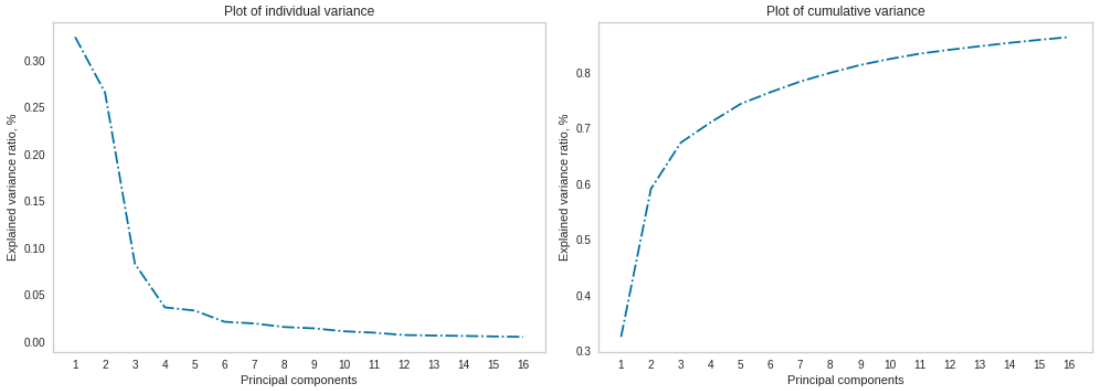


Figure 8 Plots of principal components' individual and cumulative variance in the English language dataset

As in the case of the Russian websites, the choice of the clusters number lied in the range of 0 to 40 clusters. Based on the elbow curve the number of clusters chosen is equal to 8 (Figure 9):

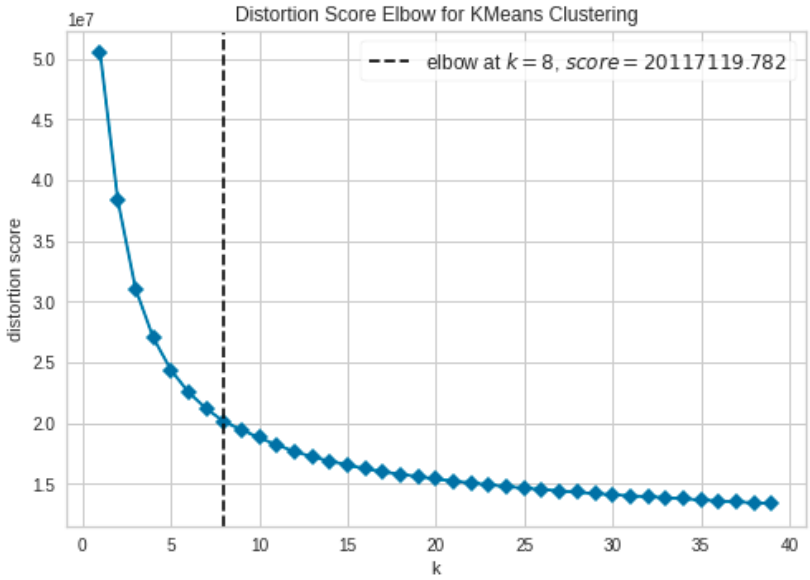


Figure 9 Elbow curve for English language websites

Thus, the general description is presented in the following table (Table 15):

Cluster №	Cluster name	General description	№ of websites	Key words examples	Website examples
0	Miscellaneous websites	Mix of travel and non-travel sites	7032	Get, travel, hotels, domain	bergoglium.com (religious tourism) intervitrina.ru (financial site)
1	Miscellaneous websites	Mix of travel and non-travel sites	13297	Travel, read, hotel, city	qatarhandball2015.com (handball tournament) internationalfbb.blogspot.com (shopping blog)
2	Miscellaneous websites	Mix of travel and non-travel sites	16521	Travel, best, hotels, read, new	elmundoconmigo.com (broken link) homeandaway.co.in (travel blog)
3	Miscellaneous websites	Mix of travel and non-travel sites	4787	Travel, best, hotels, world	cadzandbad.nl (hotels) goldfish-dodecahedron-ckpx.squarespace.com (travel blog)
4	Miscellaneous websites	Mix of travel and non-travel sites	247	Media, public, destinations	oppasharing.com (Japanese site) followmetohungary.com (Japanese site)
5	Miscellaneous websites	Mix of travel and non-travel sites	1323	Travel, booking, best, hotels, search	dealsandcouponsonline.com (coupons), clmhome.com (broken link)
6	Miscellaneous websites	Mix of travel and non-travel sites	13697	Travel, new, best, get, hotels,	eternal-guild.com (online game) cumbriancottagelets.co.uk (cottages)
7	Miscellaneous websites	Mix of travel and non-travel sites	8171	Travel, hotels, get	hktravelers.blogspot.com (backpackers forum) pinkskiesandparadise.com (expired hosting)

Table 15 Description of the English language clusters

The clusters are named in the table ‘Miscellaneous websites’ because no clear pattern was detected within the groups and specific cluster names could not be assigned. Each cluster presents the mixture of travel-related and not travel-related websites and, thus, each cluster is similar to each other. On the contrary to the Russian language clustering there are no small groups consisting of the variations of the same website.

In a 2-dimensional space the clusters can be presented as following (Figure 10):

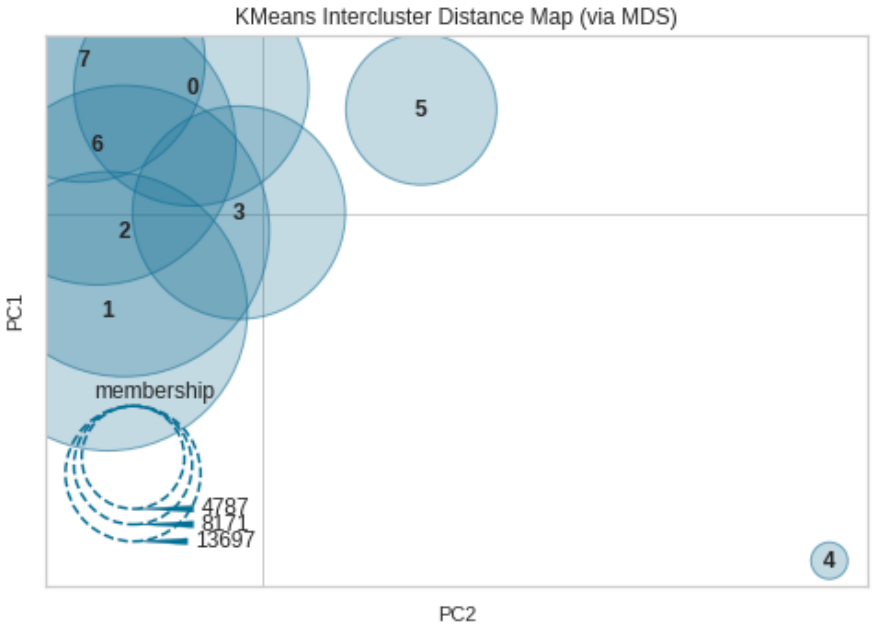


Figure 10 English websites intercluster distance map

In the graph it is seen that the main amount of clusters intersect and, thus, they are hard to discern. Indeed, the same keywords are used for almost all clusters. For example, the words ‘travel’, ‘hotels’ or ‘home’ are attributed to almost all clusters. This leads to variously-themed sites being united as one cluster. Cluster 4 and cluster 5 stand out, though. Indeed cluster 4 is the only cluster with the unique set of keywords: media, public, destinations. Nevertheless, it prevailingly consists of mixture of Japanese websites that occasionally use English words. The appearance of such cluster also emphasize the problem of language detection and existence of irrelevant observations in the data. Cluster 5 has a number of coupon and promo codes sites within it, however, they are still mixed with other-themed sites. Overall, no distinct hidden patterns were found from the data.

DBScan clustering was also performed with arbitrary parameters ε-neighbourhood equal to 3 and minPoints equal to 10. DBScan detected 5 clusters in the data, however, as DBScan removes the observations that are considered to be outliers from the dataset not all affiliates were analyzed (Table 16):

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	57861	travel, home, hotel, world	bergoglium.com (religious travel), italyvacationpackages.it (travel to Italy)
2	Forums	Forums on various topics	12	forum, view, posts, post, thread, feed	masterrussian.net, tenerifeforum.org
3	Metasearch engines	Websites like Aviasales and its competitors	17	booking, price, search rooms, find, compare, best	booking.com
4	Events	Landing pages of various events that receive profit from selling tickets	33	coin, home, holiday, official	trianglepremierleague.com (local soccer league matches), internationalfestivaloflife.com (food festival)
5	Blogs	Blogs devoted to various topics	14	new, read, word, review	beadeegee.com (personal blog),

Table 16 DBScan. English websites group division

DBScan could not break travel-themed sites into smaller groups and, thus, Cluster 0 includes 57 861 website, which amounts for 86% of the initial dataset. Therefore, cluster 0 is too general to be able to make any conclusions about it. Nevertheless, DBScan was able to detect several small groups: forums, metasearch engines, events and blogs.

6.4. Data classification

For the supervised learning part Aviasales decided not to base class division on clustering results. Thus, from clustering the company received the general outlook on how the data can be divided, however it wanted to create a sort of minimum valuable product model that could have been further used and developed internally by the employees. Therefore the company introduced the following base groups:

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

However, 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 in the case of the English dataset, the langdetect package did not manage to perfectly determine the language ‘error’ class was added for this data. The assumption is that this can be connected to the nature of the alphabet meaning that langdetect could not perfectly discern between latin letters used in various languages, for example, in English, Spanish or German, while cyrillic letters were more distinguishable. Moreover, as English is the worldwide language, a number of websites utilize English words together with their own language words or use English words in the websites name, thus, confusing the algorithm. The example of such a website is sister.travel:

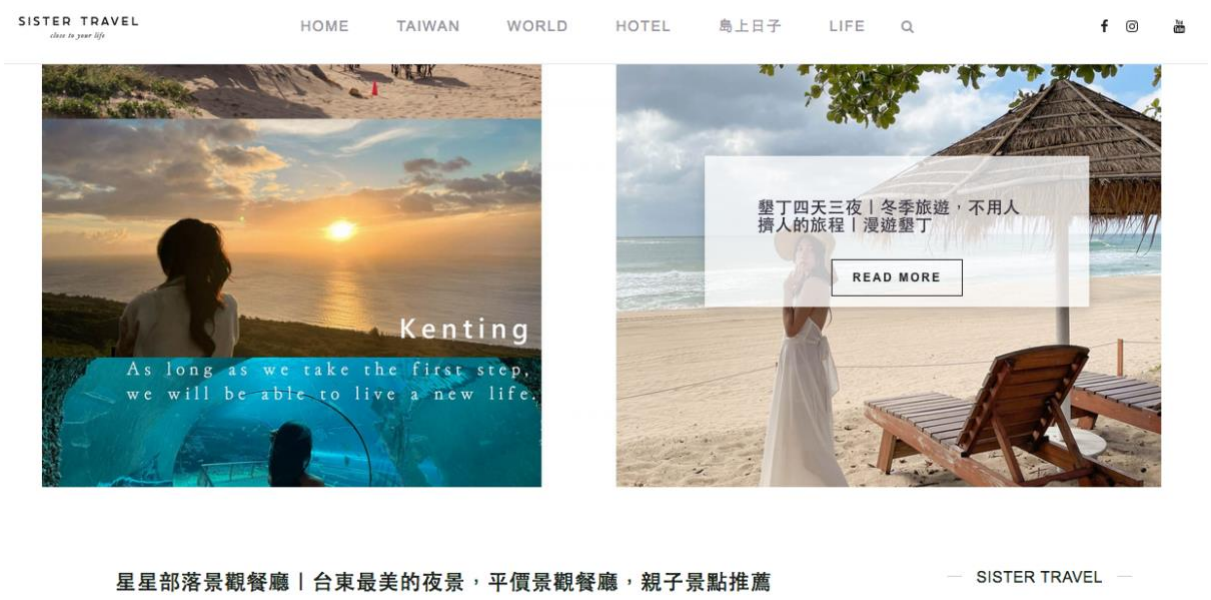


Figure 11. sister.travel main page

For the supervised part it was decided to leave broken links in the data labeled as ‘not available’ to train the machine to discern them as a separate group. Furthermore, a peculiar group of sites was detected in the Russian language dataset: sites that did not show their content but instantly automatically redirected to Aviasales. The company also wanted to leave this group of sites in the dataset and defined them as service sites. However, these sites were different from the rest of the service class sites and confused the model prediction. Therefore, they were attributed to the technical class ‘service1’.

Thus, finally, seven classes were created: content sites, cashback or promo codes sites, service sites, service1, other, 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.

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 i.e. missing values. 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 17):

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 17 English classification dataset after parsing

Missing values were dropped from the dataset. Thus, it was transformed in the following way (Table 18):

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 18 English classification dataset after missing values drop

To perform the model the data was split into training and test sets with the application of Python scikit-learn package train_test_split function. 80% of the data was assigned to the training set and 20% of the data was assigned to the test set. The distribution of classes among dataframes was the following (Table 19):

Class name	Initial distribution of classes across the dataframe	Distribution of classes across the dataframe after missing values 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
Service1	95	95	70	25
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
Service1	0	0	0	0
Cashback/promo codes	73	51	44	7
Not available	298	31	27	4
Error	332	233	181	52
Other	89	61	52	9

Table 19 Distribution of classes throughout the dataset, test set and train set

The largest number of sites in the datasets belonged to the ‘content’ class with ‘service’ and ‘error’ being runner-ups. Moreover, ‘error’ class was not detected in the Russian language

dataset: all manually labelled websites even though some of them belonged to Ukranian domains were written in Russian language. In the case of the English language websites no ‘service 1’ class belonging websites were detected. Only 6 sites are included in the Russian language ‘cashback and promo codes' class as a result of its poor representation in the initial dataset.

6.4.1. Gradient boosting classifier

The first model to try was the general Gradient boosting classifier. 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. In order to avoid overfitting of the parameters, randomized search was applied with 3-fold cross-validation. For Gradient boosting model the following parameters were applied and tested (Table 20):

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
max_depth	a parameter that sets a maximum value for nodes in each individual tree (weak model)	10, 40, None	None	None

Parameter name	Parameter meaning	Tested values	Finally chosen value Russian language dataset	Finally chosen value English language dataset
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 20 Parameters of the applied Gradient Boosting model

The Gradient Boosting model produced the following classification reports (Figure 12):

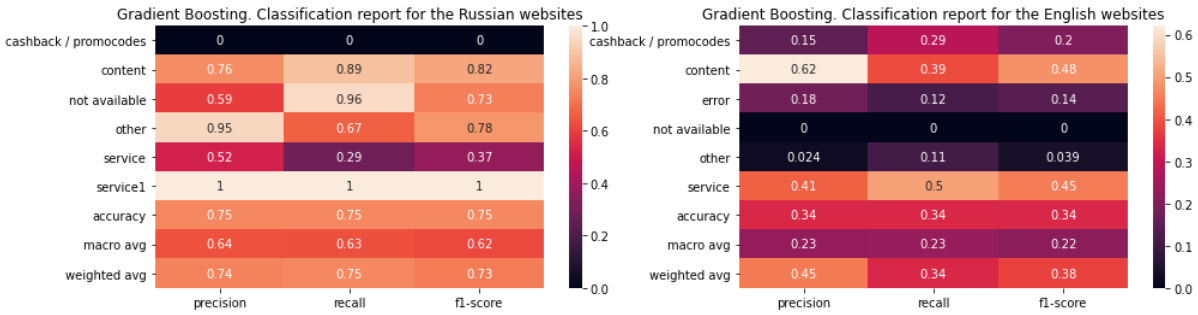


Figure 12 Classification reports for Russian and English language websites after application of Gradient Boosting model

In the case of the Russian language dataset no ‘error’ class sites were detected and in the English language dataset no representatives of ‘service 1’ class were found. Therefore, these classes are not included in the classification report. F-score is chosen as the main metric of model quality because it combines both precision and recall scores and is less affected by sample imbalance.

In the case of the Russian language ‘content’, ‘not available’’ and ‘other’ classes achieved relatively good F-score with its values of 0.82, 0.73 and 0.78 respectively. Unfortunately ‘service’ class showed only 37% F-score. Presumably, this can be explained by the broad nature of the term service: a lot of different websites fall into this category: from wedding on Cyprus (e.g. http://svadbavmire.ru/) to online bets (e. g. http://stavochka.com/). As this class forms the basis of and ‘cashbacks and promo codes’ were not predicted at all. The poor result of cashback and promocode class prediction is connected to the fact of its minimal representation in the labelled dataset.

In comparison to Russian language websites all classes in the English language dataset were predicted quite poorly. However, the top three by F-score classes are the basic classes offered by Aviasales: content, service and cashbacks and promo codes with F-scores of 0.48, 0.45 and 0.2 respectively. In the case of English websites ‘not available class is not predicted’. This once again can be attributed to the fact of the small size of a sample. All the missing values in the English language dataset belonged to the class ‘not available’, meaning it majorly decreases after missing values drop.

6.4.2. CatBoost Classifier

The second model applied was CatBoost Classifier. To increase the chances of obtaining the best model, randomized search with 3-fold cross-validation was also applied. In terms of randomized search a set of parameters was chosen arbitrarily and each possible combination of them was fitted to the model with loss calculations (Table 21). The final parameters chosen compile the models with minimum loss value.

Parameter name	Parameter meaning	Tested values	Finally chosen value Russian language dataset (model.randomized_search)	Finally chosen value English language dataset (model.randomized_search)
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

Parameter name	Parameter meaning	Tested values	Finally chosen value Russian language dataset (model.randomized_search)	Finally chosen value English language dataset (model.randomized_search)
l2_leaf_reg	Regularization parameter	1, 2, 3, 4 ,5, 7, 9	7	4

Table 21 Parameters of CatBoost model obtained with randomized search

Both Russian and English datasets showed the best results with number of iterations equal to 300, learning rate equal to 0.1 and tree depth equal to 6. However, the l2_leaf_reg parameter is higher for the Russian language dataset. l2_leaf_reg is the L2 (Ridge))regularization parameter of the cost function that controls the complexity of the model in order to avoid overfitting. The higher value of the regularization parameter in the Russian dataset indicates that it is more prone to overfitting.

For all other parameters default values were used. Thus, the following classification reports were achieved (Figure 13):

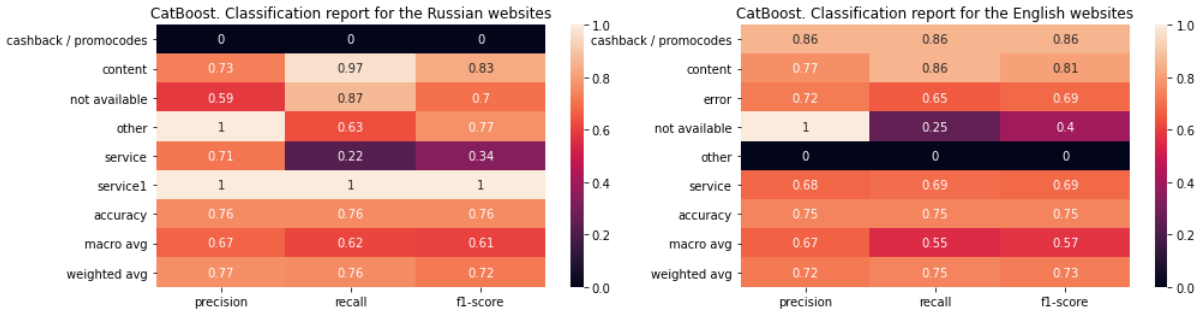


Figure 13 Classification reports or Russian and English language websites after application of CatBoost model

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-score (0.77), 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-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-score and accuracy scores of the executed models looks the following way (Table 22):

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

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

Table 22 Comparison of classification models

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.

7. Further analysis and data visualization

The main dataset was merged with an auxiliary one and, thus, the interactive Excel dashboard was formed to derive managerial insights (Figure 14):

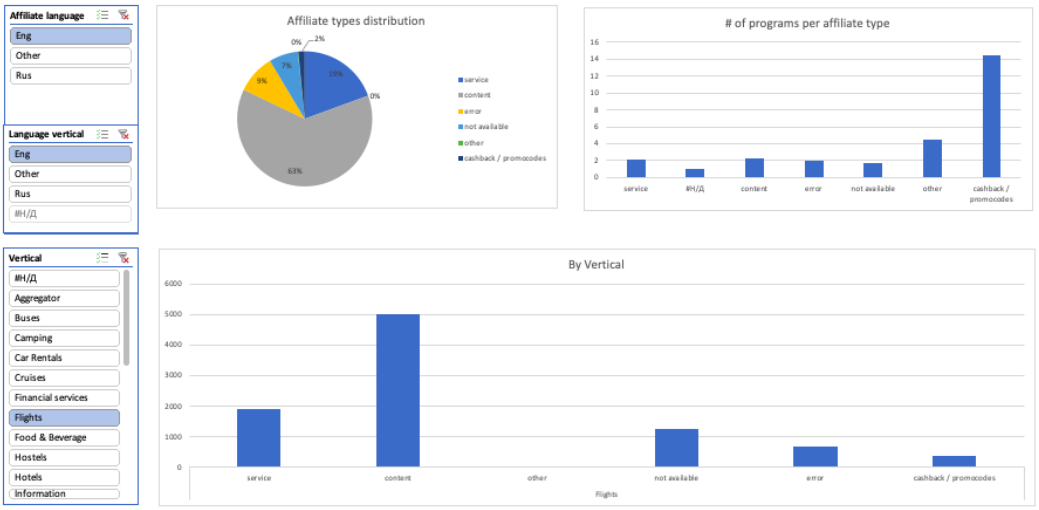


Figure 14 Interactive Excel dashboard

Important to note that the auxiliary dataset is also provided by Aviasales, however, it initially mainly focuses on advertisers (Table 7) and does not fully correspond to dataset that was used in terms of the Machine Learning models (Table 6). Thus, the number of unique affiliates in both dataset differ. In case of the Russian language websites the initial dataset included 6999 Russian websites, while in the auxiliary dataset there are only 3000 of them. In the case of English websites there are 50 815 unique affiliates instead of 67 490 in the initial dataset. Moreover, the auxiliary dataset includes various languages, however, in term of this Thesis only Russian and English websites are analyzed. Nevertheless, the company wanted auxiliary dataset to be further analyzed to see the analytics in relation to a certain list of advertisers it provided.

7.1. 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 15):

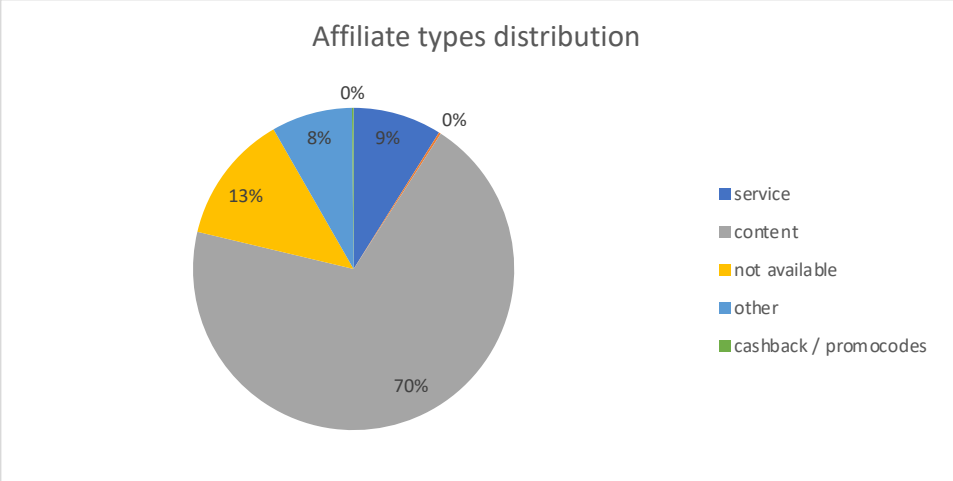


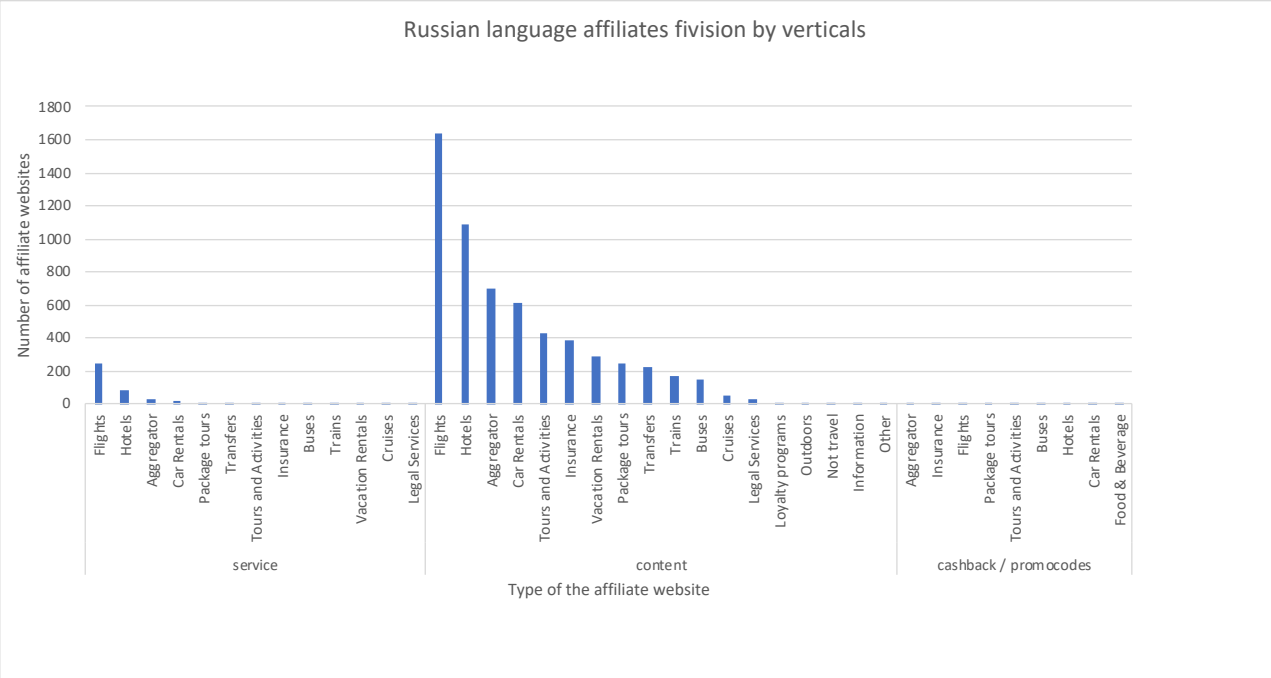
Figure 15 Distribution of Russian language affiliates according to their class

Important to note that 13% of the data are broken links. This partly reflects the quality of the participating affiliates: many of them rapidly 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 (niches of the market), 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 16):



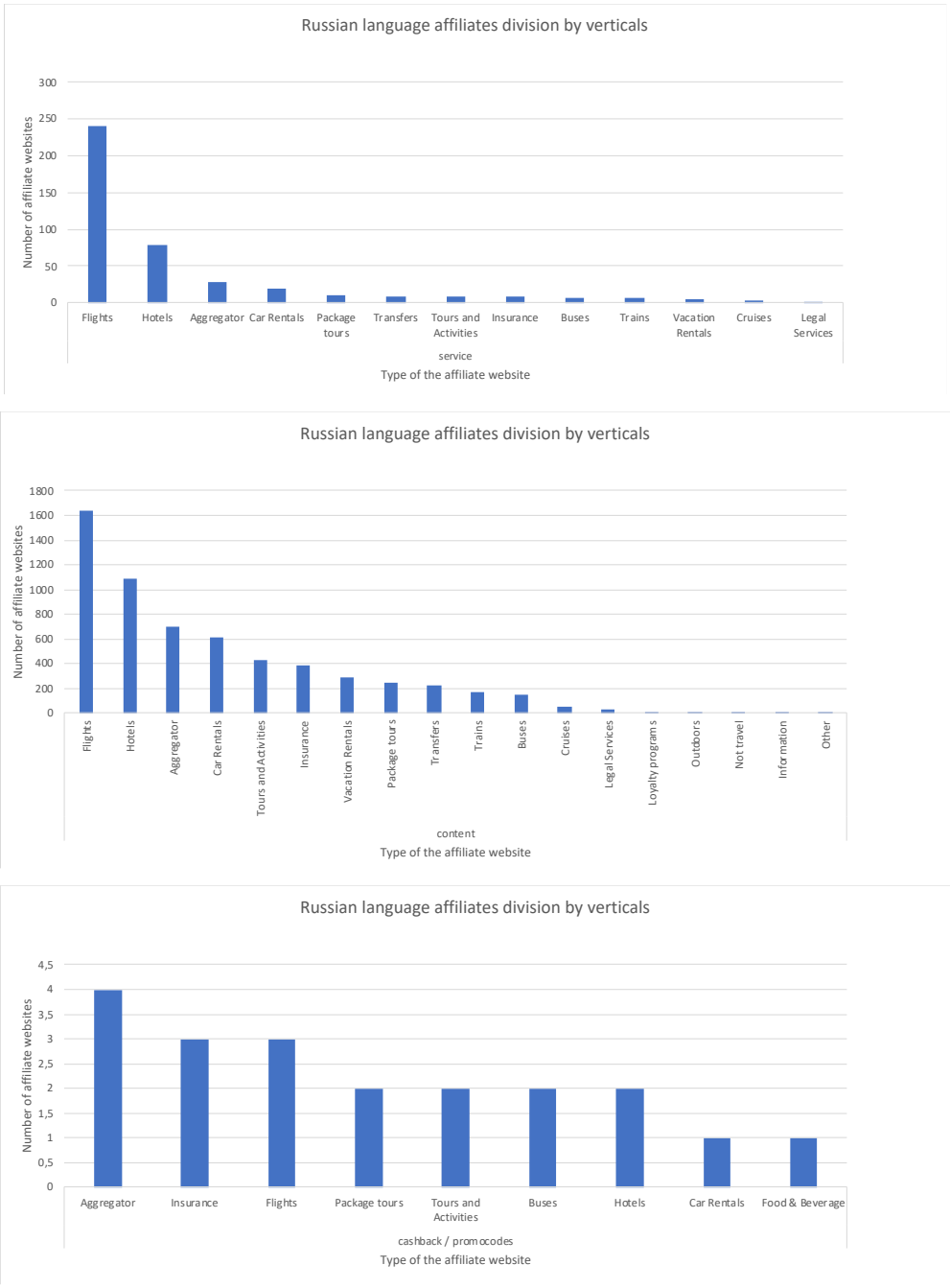


Figure 16 The structure of verticals by the Russian language site affiliate type (overall and zoomed by type)

It is also interesting to consider the analysis of the main verticals: “Flights” and “Hotels” to look at the data from different perspective (Figure 17):

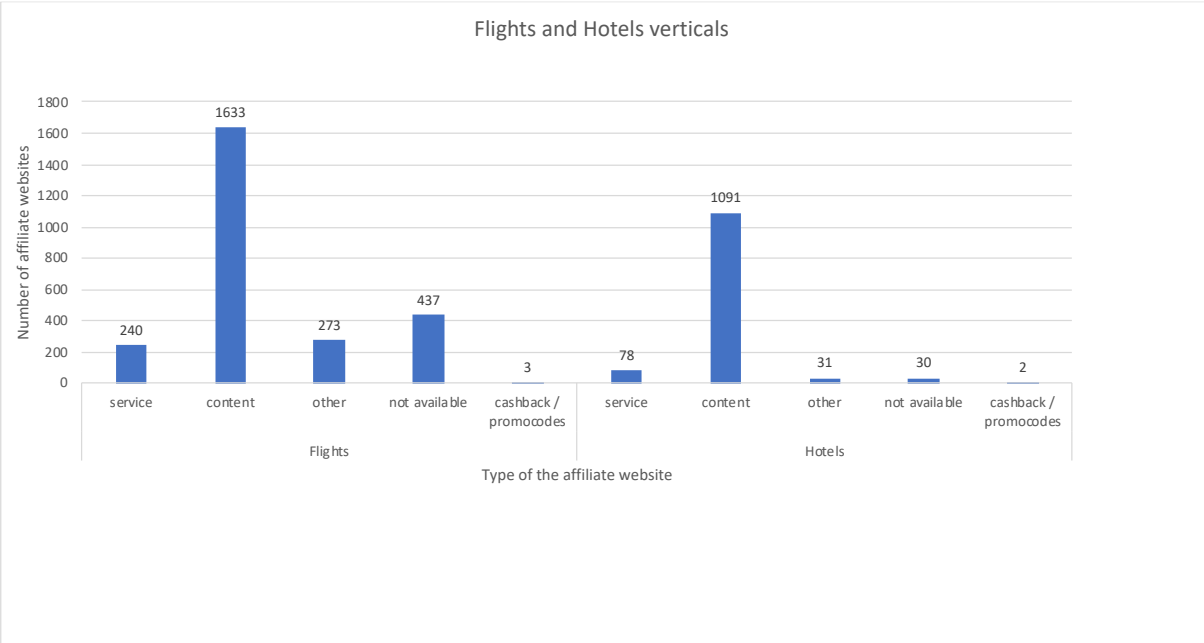


Figure 17 The structure of Flights and Hotels verticals

Flights have a significant number of not available links, which means that this category is the most susceptible to the appearance of affiliates that quickly become abandoned, for example, due to hosting expiration.

Another interesting information to look into is the number of affiliate programs a certain type of affiliate participates. 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 18):

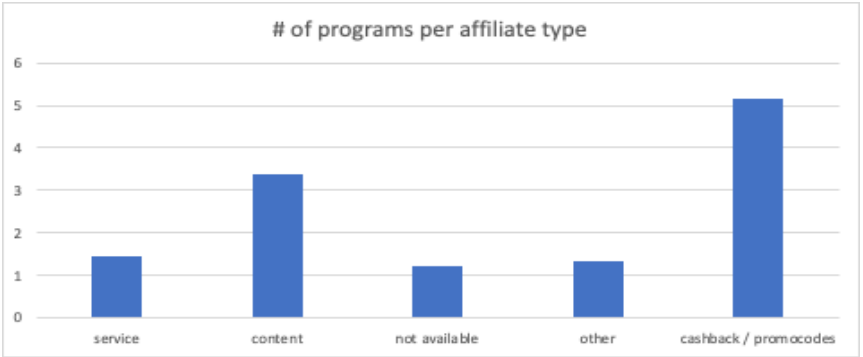


Figure 18 The number of programs per affiliate type (Russian websites)

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.

7.2. 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 19):

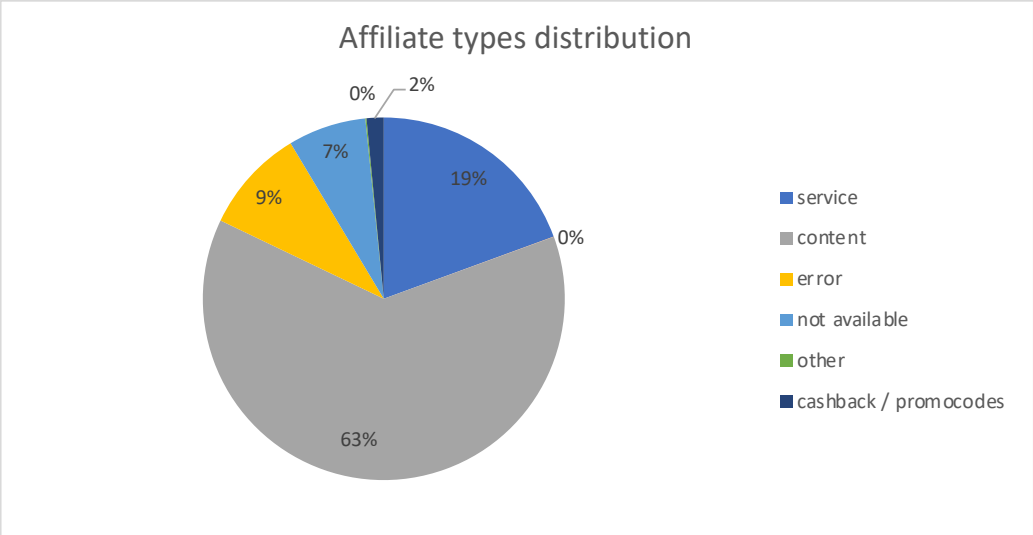


Figure 19 Distribution of English language affiliates according to their class

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 20):

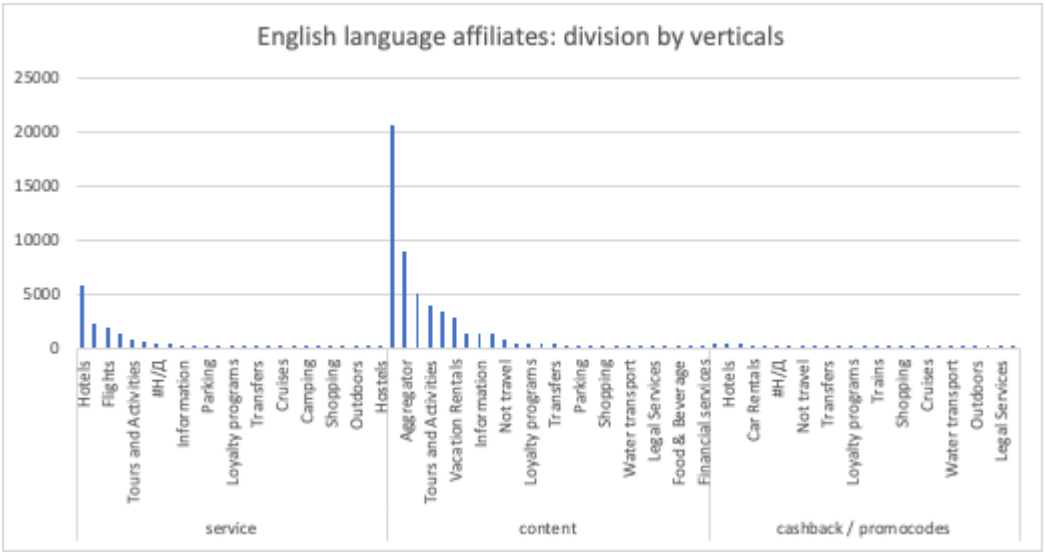


Figure 20 The structure of verticals by the English language site affiliate type.

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 21):

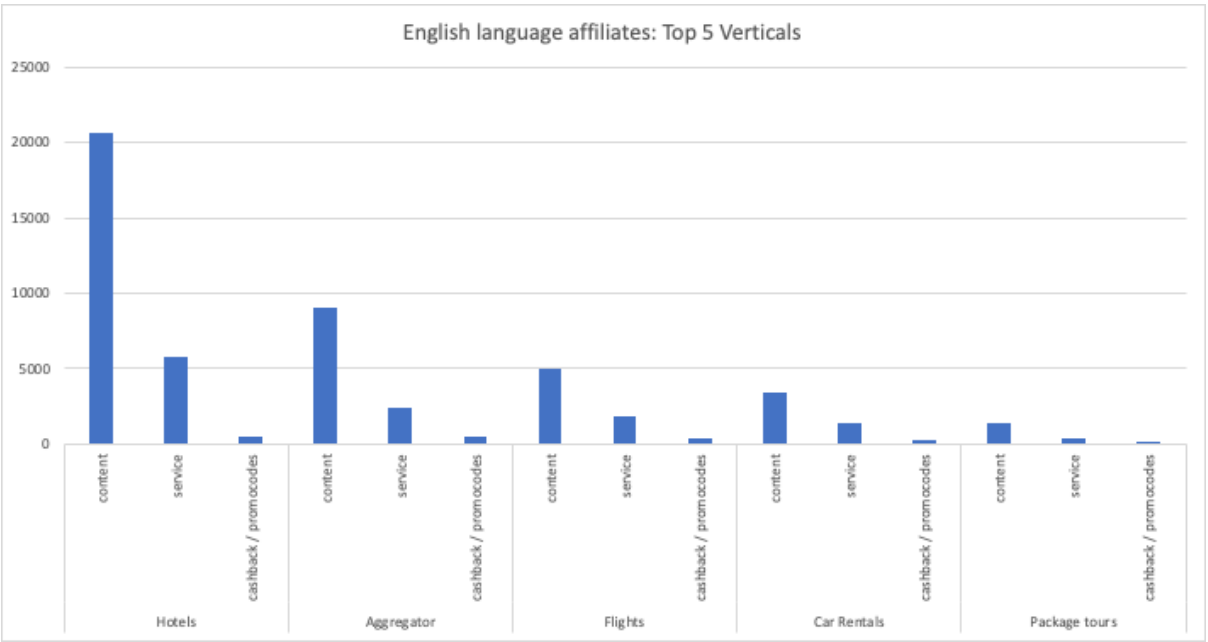


Figure 21 The structure of classes across Top-5 verticals

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 22):

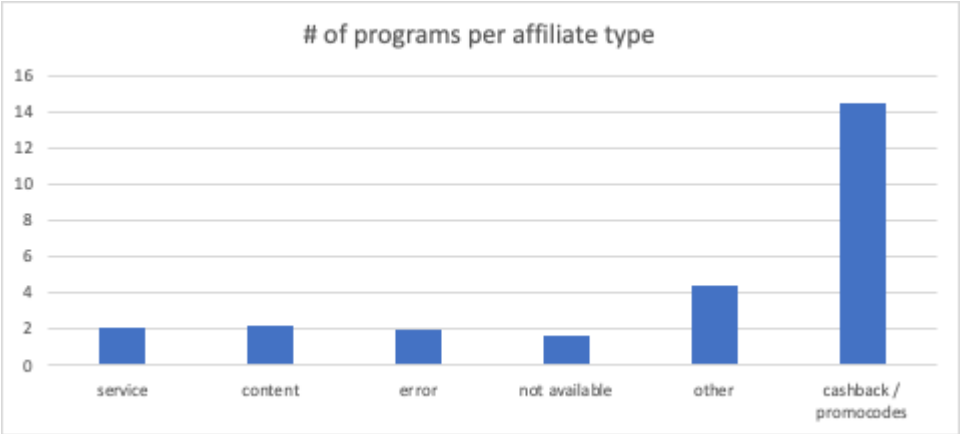


Figure 22 The number of programs per affiliate type (English websites)

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 are 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.

8. Managerial application and further directions of research

Clustering 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 clustering 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 cluster analysis showed that the differences among affiliate exist and affiliates can be divided into subgroups.

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 consumers. 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 (service1 class) also may not imply long-term relationships and bring benefit for Aviasales. Thus, these sites do not invest effort into content development but rather try to 'catch' a user to make him click on the link in order to receive commission from Aviasales. Straight redirection from a different site to Aviasales websites may create a perception of Aviasales website being a mirror or a fraudulent site. 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 hotels
- 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

9. Conclusion

All in all, this thesis 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. Moreover, it implemented a real-life case study from one of the biggest travel aggregators – Aviasales and analyzed the company's 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 pandemics, the travel sector still presents a wide range of opportunities for affiliate marketing. The industry is highly-competitive and, thus, urges advertisers to look for new ways to stand out. Moreover,

the industry changes under the influence of digitalization. Thus, more and more consumers buy travel services via mobile devices and the Internet, and, therefore, digital marketing tools that include affiliate marketing are starting to become more and more important.

Despite the growing popularity of affiliate marketing and increase in the data availability, Machine Learning tools are currently implemented in the field only to a limited degree. Thus, the research included 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 thesis also described implementation of Natural language processing algorithms including language detection.

In terms of this thesis the affiliates were divided into two datasets – English and Russian language websites. This approach allowed to determine the affiliates' structure within the program taking into account semantic peculiarities of the data. Moreover, the thesis considered various types of the affiliates and their relations with travel sub-industries.

To achieve the goals presented in the introduction the paper introduced clustering and classification algorithms that allowed both to determine hidden patterns of the data and take into account Aviasales viewpoint on affiliates analytics.

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 included 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 travel sub-industries which are flights and hotels since they represent the most popular consumer queries in terms of travelling. Moreover, content websites were defined as the widespread type of affiliates. 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 may diminish their value in consumers’ viewpoint 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 benefit can be achieved from such partnership.

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 promoted advertiser's link the most?
- What type of payment mechanism (CPC, CPA) is better for advertisers?
- Which affiliates present the most opportunities for monetization?

In the end, despite the thesis being one of the few dedicated to this vast and complex topic of affiliates in the travel industry, the stated research goals were completed.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews. Computational Statistics, 2(4), 433–459. https://doi.org/10.1002/wics.101

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

Alpaydın, E. (2014). *Introduction to machine learning*. Cambridge, MA: MIT Press

Aviasales. (2021). *O kompanii [About the company]*. Aviasales.ru. https://www.aviasales.ru/about.

Baidin, I. (2018). *Istoria Travelpayouts: ot idei do vyplaty pervogo milliarda [Travelpayouts history: from the ide to the first milliard]*. Vc.ru. Retrieved 19 June 2021, from https://vc.ru/aviasales/52867-travelpayouts.

Bartlett, P., Freund, Y., Lee, W. S., & Schapire, R. E. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics*, *26*(5), 1651-1686.

BERT. Huggingface.co. (2021). https://huggingface.co/transformers/model_doc/bert.html.

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

Bhavsar, K., Kumar, N., & Dangeti, P. (2017). Natural language processing with Python cookbook : over 60 recipes to implement text analytics solutions using deep learning principles (1st edition). Packt.

Bowd, C., Belghith, A., Proudfoot, J. A., Zangwill, L. M., Christopher, M., Goldbaum, M. H., ... & Weinreb, R. N. (2020). Gradient-Boosting Classifiers Combining Vessel Density and Tissue Thickness Measurements for Classifying Early to Moderate Glaucoma. *American journal of ophthalmology*, *217*, 131-139.

Bremner, C., & Popova, N. (2020). *Digital Travel Innovation Across the Traveller Journey*. Euromonitor International.

Celebi, M. E. (2015). Partitional Clustering Algorithms. Springer International Publishing. https://doi.org/10.1007/978-3-319-09259-1

Chadjipadelis, T. (2021). Data analysis and rationality in a complex world . https://doi.org/10.1007/978-3-030-60104-1

Chadjipadelis, T., Lausen, B., Markos, A., Lee, T. R., Montanari, A., & Nugent, R. (2021). Data Analysis and Rationality in a Complex World. Springer International Publishing AG.

Chernikova, A. (2015). *Kak rabotayut v Aviasales [How they work in Aviasales]*. The Village. https://www.the-village.ru/business/office/174343-ofis-aviasales.

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.

Deloitte. (2020). *2019 Travel and Hospitality Industry Outlook*. https://www2.deloitte.com/us/en/pages/consumer-business/articles/travel-hospitality-industry-outlook.html

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.

Downs, M., York, J., Clayton, B., Tradgett, C., Hall, D., & Gregoriadis, L. (2008). Affiliate marketing. *Journal of Direct, Data and Digital Marketing Practice*, *9*(3), 304–.

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

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

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

Forrester Consulting (2016). *New Affiliate Marketing Research 2016*. Go.rakutenadvertising.com. https://go.rakutenadvertising.com/new-affiliate-marketing-research-2016.

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

Gomer, R., Rodrigues, E., Milic-Frayling, N., & Schraefel, M. (2013). Network Analysis of Third Party Tracking: User Exposure to Tracking Cookies through Search. Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 1, 549–556. https://doi.org/10.1109/WI-IAT.2013.77

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

Hardeniya, N. (2016). *Natural Language Processing: Python and NLTK*. Packt Publishing.

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

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.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis.Machine Learning, 42, 177–196.

Herrera, F., Charte Ojeda, F., Rivera, A. J., & del Jesus, M. J. (2016). Multilabel Classification Problem Analysis, Metrics and Techniques . Springer International Publishing. https://doi.org/10.1007/978-3-319-41111-8

Hossan, F., & Ahammad, I. (2013). Affiliate Marketing: The Case of Online Content Providers in Bangladesh. *World*, *3*(2).

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

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

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

Jaadi, Z. (2021). 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

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

Li, J., Si, Y., Xu, T., & Jiang, S. (2018). Deep Convolutional Neural Network Based ECG Classification System Using Information Fusion and One-Hot Encoding Techniques. *Mathematical Problems in Engineering*, *2018*, 1–10. https://doi.org/10.1155/2018/7354081

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

Liu, F., & Deng, Y. (2021). Determine the Number of Unknown Targets in Open World Based on Elbow Method. IEEE Transactions on Fuzzy Systems, 29(5), 986–995. https://doi.org/10.1109/TFUZZ.2020.2966182

Kazmina, I., Zhukova, K., Uzbekova, I., Petukhova, L., Borodina, V., Titova, Y., . . . Yakobvenko, D. (2020). 20 samyh dorogih kompanij Runeta [20 most expensive companies in Runet]. Forbes. Retrieved March 06, 2021, from https://www.forbes.ru/biznes-photogallery/393345-20-samyh-dorogih-kompaniy-runeta-reyting-forbes?photo=9

Kumar, A., & Jain, M. (2020). Ensemble Learning for AI Developers Learn Bagging, Stacking, and Boosting Methods with Use Cases (1st ed. 2020.). Apress. https://doi.org/10.1007/978-1-4842-5940-5

langdetect. PyPI. (2021). https://pypi.org/project/langdetect/.

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).

MarketLine Industry Profile. (2020). *Global Travel and Tourism*. (Report No. 0199-2806) MarketLine. https://advantage-marketline-com.ezproxy.gsom.spbu.ru/Analysis/ViewasPDF/global-travel-tourism-112483

Meng, W., Yu, C., & Liu, K.-L. (2002). Building efficient and effective metasearch engines. ACM Computing Surveys, 34(1), 48–89. https://doi.org/10.1145/505282.505284

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

Nimmermann, F. (2020). Congruency, Expectations and Consumer Behavior in Digital Environments (1st ed. 2020.). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-28421-3

NLTK. Natural Language Toolkit — NLTK 3.6.2 documentation. Nltk.org. (2021). Retrieved 20 June 2021, from https://www.nltk.org/.

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.

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

Overview of CatBoost - CatBoost. Documentation. Catboost.ai. (2021). https://catboost.ai/docs/concepts/about.html.

Pakhomov, V. (2010). *Mif № 7. Napisanie e vmesto ë – grubaya orfograficheskaya oshibka.* [Myth № 7. Writing e instead of ë is a serious spelling mistake]. Gramota.ru. Retrieved 20 June 2021, from http://gramota.ru/class/istiny/istiny_7_jo/.

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.

Pourahmad, S., Basirat, A., Rahimi, A., & Doostfatemeh, M. (2020). Does Determination of Initial Cluster Centroids Improve the Performance of K-Means Clustering Algorithm? Comparison of Three Hybrid Methods by Genetic Algorithm, Minimum Spanning Tree, and Hierarchical Clustering in an Applied Study. Computational and Mathematical Methods in Medicine, 2020, 7636857–11. https://doi.org/10.1155/2020/7636857

Prussakov, G. (2016). *20 Affiliate Marketing Stats That Will Blow Your Mind*. AM Navigator. Retrieved 19 June 2021, from https://www.amnavigator.com/blog/2016/04/27/20-affiliate-marketing-stats-will-blow-mind/.

Prussakov, G. (2015). *Analysis of 550 Best Affiliate Programs Reveals Top 20 Niches*. Affiliate Marketing Navigator. https://www.amnavigator.com/blog/2015/09/25/analysis-of-best-affiliate-programs-top-20-niches/.

Razrabotka. (2017, December 7). *Opensource v Yandexe: Catboost — novoe pokolenie gradientnogo boostinga* [Opensource in Yandex: Catboost a new generation of gradient boosting] [Video]. Youtube. https://youtu.be/Q_xa4RvnDcY

scikit-learn: machine learning in Python — scikit-learn 0.16.1 documentation. Scikit-learn.org. (2021). https://scikit-learn.org.

Semerádová, T., & Weinlich, P. (2020). Using Google Analytics to Examine the Website Traffic. In Website Quality and Shopping Behavior (pp. 91–112). Springer International Publishing. https://doi.org/10.1007/978-3-030-44440-2_5

Smit, E. G., Van Noort, G., & Voorveld, H. A. (2014). Understanding online behavioural advertising: User knowledge, privacy concerns and online coping behaviour in Europe. *Computers in Human Behavior, 32*, 15-22.

Travelpayouts. (2021). *Getting started*. Travelpayouts Help Center. https://support.travelpayouts.com/hc/en-us/articles/203955593-Getting-started.

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

Vlase, S., Marin, M., & Öchsner, A. (2019). Eigenvalue and Eigenvector Problems in Applied Mechanics (1st ed. 2019.). Springer International Publishing. https://doi.org/10.1007/978-3-030-00991-5

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

Wang, S., & Yu, G. (2001). *Advances in Web-Age Information Management: Second International Conference, WAIM 2001, Xi'an, China, July 9-11, 2001. Proceedings* (Vol. 2). Springer Science & Business Media.

Wu, J. (2012). *Advances in K-means clustering: a data mining thinking*. Springer Science & Business Media.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*(1), 1-37.

Yeung, A. (2021). *BERT - Tokenization and Encoding*. Github.io. Retrieved 20 June 2021, from https://albertauyeung.github.io/2020/06/19/bert-tokenization.html.

Yu, K., Guo, G.-D., Li, J., & Lin, S. (2020). Quantum Algorithms for Similarity Measurement Based on Euclidean Distance. International Journal of Theoretical Physics, 59(10), 3134–3144. https://doi.org/10.1007/s10773-020-04567-1

Zheng, Y., Wang, W., Chen, B., Zhang, L., Phangthavong, S., Su, Z., ... & Xiao, G. (2019). Determining the number of instars in potato tuber moth Phthorimaea operculella (Zeller) using density-based DBSCAN clustering. *Journal of Applied Entomology*, *143*(10), 1080-1088.