UNIVERSITY OF TURKU

ABSTRACT

	Bachelor's thesis
X	Master's thesis
	Licentiate's thesis
	Doctoral dissertation

Subject	Marketing	Date	8.12.2019
Author	Veera Engren	Student number	
		Number of pages	84
Title	Classifying Search Queries To Digital Customer Purchasing Funnel Stages: A Machine Learning Approach		
Supervisors	Prof. Rami Olkkonen, prof. Aino Halinen-Kaila		

Abstract

Understanding customer journeys is essential for marketers in order to guide consumers through the purchase funnel and respond more effectively to the consumers' information needs in different stages of their buying journeys. Internet and social media have had a significant effect on consumer decision process and search engines have a central role in the consumer purchase journeys in the digital era. Also, year after year search engine marketing gains the biggest revenues in digital advertising.

While targeting advertising according to the consumer purchase journey stages is considered important, manual efforts to classify all the keywords or search terms are not feasible. One search advertising account can include tens of thousands of keywords and consumer search queries are continuously evolving. This study presents a supervised machine learning approach to automatically classify consumer search queries into customer journey stages.

Automated solutions are now a popular approach in the field of marketing, too, because of the vast amount of data available for decision-making. As machine learning becomes more and more usable and available for marketing purposes, it is important for practitioners to understand its possibilities and be aware of its limitations. For one of the cornerstones in supervised machine learning applications is the creation of training data, the study examined factors that affect the quality and validity of the training data and therefore the success of automatic classification. All in all, machine learning cannot provide valid results without representative training data.

The research data consisted of 190 245 search queries gathered from the Google Ads account of a Finnish e-commerce company over a period of five years. Two different machine-learning algorithms, Random Forest and XGBoost, were applied and their results compared. The classifiers were able to find distinct classes and classify search queries accurately while no significant differences in the performance of the two classifiers were observed. The machine learning approach to large-scale automatic classification of search queries shows successful results and this method and these categories related to the purchase funnel can be used in analyzing customer behavior in different stages of the online purchase process or targeting advertising.

Key words	Digital Marketing, Machine Learning, Search Engine Advertising, Purchase Funnel, Consumer Journey, Classifier, Algorithm, Random Forest, XGBoost, Google Ads

TURUN YLIOPISTO

TIIVISTELMÄ

	Kandidaatintutkielma
X	Pro gradu -tutkielma
	Lisensiaatintutkielma
	Väitöskirja

Oppiaine	markkinointi	Päivämäärä	8.12.2019
Tekijä	Veera Engren	Matrikkelinumero	
		Sivumäärä	84
Otsikko	Classifying Search Queries To Digital Customer Purchasing Funnel Stages: A Machine Learning Approach		
Ohjaajat	prof. Rami Olkkonen, prof. Aino Halinen-Kaila		

Tiivistelmä

Kuluttajan ostopolun ymmärtäminen on oleellista, jotta markkinoijat pystyisivät paremmin tarjoamaan kuluttajalle relevanttia tietoa kussakin ostopolun vaiheessa, ja täten viemään kuluttajan koko ostofunnelin (ns. ”ostosuppilon”) läpi. Internetillä ja sosiaalisella medialla on ollut suuri vaikutus kuluttajien ostokäyttäytymiseen. Digitalisaation myötä myös ostoprosessi on digitalisoitunut ja hakukoneita käytetään aktiivisesti päätöksenteossa. Hakukonemarkkinoinnilla onkin valta-asema digitaalisessa markkinoinnissa vuosi toisensa jälkeen.

Hakukonemainonnan kohdentamista kuluttajan ostopolun mukaisesti pidetään tärkeänä, mutta kaikkien avainsanojen ja hakutermien luokittelu manuaalisesti ei ole järkevää. Mainostilit sisältävät runsain määrin avainsanoja, ja kuluttajien käyttämät hakutermit muuttuvat jatkuvasti ajan myötä. Tutkielma esittää automatisoidun keinon luokitella hakutermejä kuluttajan ostopolun vaiheisiin käyttäen ohjattua koneoppimista.

Koska saatavilla olevan datan määrä on nykyisin valtava, ovat automaatioratkaisut suosittuja myös markkinoinnin alalla. Koneoppimisen käytettävyyden parantuessa tulisi markkinoinnin ammattilaistenkin ymmärtää sen tuomat mahdollisuudet sekä siihen liittyvät rajoitteet. Opetusdatan luominen on yksi ohjatun koneoppimisen kulmakivistä, ja siksi tutkielmassa tarkastellaan eri tekijöitä, jotka vaikuttavat opetusdatan laatuun ja sitä kautta automaattisen luokittelun onnistumiseen. Koneoppiminen ei pysty tuottamaan valideja tuloksia ilman edustavaa opetusdataa.

Tutkimusaineisto koostui suomalaisen verkkokaupan Google Ads –tililtä haetuista hakutermeistä, joita oli yhteensä 190245 viiden vuoden ajalta. Datan luokitteluun testattiin kahta eri algoritmiä, Random Forestia ja XGBoostia, ja näiden tuloksia vertailtiin keskenään. Molemmat luokittelijat pystyivät erottamaan selkeitä luokkia ja luokittelemaan hakutermit hyvin tuloksin. Koneoppimisen soveltaminen laajamittaiseen automaattiseen hakutermien luokitteluun osoittautui toimivaksi. Tätä metodia ja ostofunneliin liittyviä kategorioita voidaan hyödyntää mainonnan kohdentamisessa sekä kuluttajakäyttäytymisen analysoimisessa ostoprosessin eri vaiheissa.

Avainsanat	kuluttajan ostopolku, koneoppiminen, hakukonemarkkinointi, digitaalinen markkinointi, ostofunneli, luokittelija, algoritmi, Random Forest, XGBoost, Google Ads

UNIVERSITY OF TURKU
Turku School of Economics

CLASSIFYING SEARCH QUERIES TO DIGITAL CUSTOMER PURCHASING FUNNEL STAGES: A MACHINE LEARNING APPROACH

Master´s Thesis
in Marketing

Author:
Veera Engren

Supervisors:
prof. Rami Olkkonen
prof. Aino Halinen-Kaila

8.12.2019
Turku

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

Table of contents

List of figures

List of tables

1 INTRODUCTION

1.1 Introduction to the topic

Understanding customer journeys[1] is essential for marketers in order to guide consumers through the purchase funnel and respond more effectively to the consumer information needs in different stages of their buying journey (Wolny & Charoensuksai 2014). Internet and social media have had a significant effect on consumer decision process mainly through abundance of information and electronic Word of Mouth (e-WOM). Consumers utilize to a great extent product reviews and user recommendations, as well as other user generated content available online to support their purchase decisions. (Owusu et al. 2016.) Hence, the customer journeys have become more complex consisting of multiple touch points and mixing offline and online channels (Haven 2007; Wolny & Charoensuksai 2014; Rangaswamy et al. 2009).

There is a lot of discussion on whether the traditional purchase funnel is valid in digital environment (Hall et al. 2016; Jansen & Schuster 2011). Others think the funnel should be abandoned entirely (Noble et al. 2010), while others think the web has just transformed the funnel (McMillan 2007; Court et al. 2009; Haven 2007; Wolny & Charoensuksai 2014, 318-319). Nevertheless, practitioners, along with academicians, have emphasized the importance of delivering consumers the right content at the right moment in relation to the their purchase process stage (Petrik 2014; Vanarsdall 2016; Wolny & Charoensuksai 2014, 325; Rangaswamy et al. 2009). In addition, researchers propose that consumer search queries and responsiveness to ads vary according to the consumer's purchase funnel stage (Jansen & Schuster 2011).

Search engines have a central role in the consumer purchase processes, and it is the main tool for information search (Rangaswamy et al. 2009, 49; Su et al. 2018, 547). KPMG reports that the use of online channels is very active in the research and evaluation phases globally (KPMG 2017). In Great Britain the majority of consumers, 78% this year, use Internet to search for information on products and services (Statista 2019). According to the common understanding search engine marketing is more relevant and hence more effective than other types of digital marketing because the initiative is on the consumer side (Zhang et al. 2018, 663). And still today, the "old fashioned" text ads are the most effective ad form according to the marketers (Neely 2019a). This shows also in the European digital advertising market, which has shown continuous growth.

[1] Customer journey and purchase funnel refer to the purchase process or path that consumers go through when making a purchase decision consisting of different stages from need recognition to purchase (Lemon & Verhoef 2016, 71). These concepts will be further discussed in the study at a later stage.

The total digital advertising expenditure in Europe totaled €48.0 billion in 2017 from which search engine marketing has held the largest share (45.7% in 2017) already since 2009. (IAB Europe 2018.)

Search engine marketing practitioners as well as researchers acknowledge the importance of targeting advertising according to the consumer purchase journeys. (Hudson & Hudson 2013; Wolny & Charoensuksai 2014; Hadrien 2015a; Swan 2018; Qian 2017). One method for targeting, mainly suggested by practitioners, is mapping keywords to the purchase funnel stages. There is a wealth of academic research on user intent and search query classification but not from the perspective of the consumer purchase process. Because of the vast amount of data, manual classification is not really a feasible solution in many classification tasks and therefore this study presents a supervised machine learning[2] approach to automatically classify consumer search queries into customer journey stages.

Although not a new thing, machine learning is a hot topic at the moment. It's been around for ages but now it is becoming more and more applicable because of the technical advancements in the data processing capacity of computers. Machine learning is already a part of our everyday lives and it is applied for example in the fields of finance, accounting, medicine, e-commerce, just to mention a few. Also, new areas of application emerge constantly. For instance, just recently Mackmyra whisky distillery in Sweden released the world's first AI-created whisky. It was created using machine learning in cooperation with Finnish tech company Fourkind and Microsoft (Dedezade 2019). As the amount of data available for marketers in the digital era keeps growing and growing, also the automated and machine learning solutions become more attractive in the field of marketing. It is not humanly possible for example to manually spot hateful and malicious comments through thousands of tweets, and machine learning can help with that. Machine learning is also definitely a hot trend in the area of search engine marketing. It can be used for bidding and budget automations or better attribution modeling, for example. (Neely 2019a.) Hence, it is important for marketers, as well, to understand all the possibilities machine learning provides and be aware of its limitations.

[2] Machine learning is based on artificial intelligence and it can be described as designing systems that can predict outcomes by learning from input data (Bell 2015, 2). The concept will be defined more comprehensively at a later stage in the study.

1.2 Purpose of the study and research methods

The purpose of the study is to apply machine learning methods to automate a large-scale classification of user search queries into different consumer purchase funnel stages in digital environment.

The main research question is how the searches of search engine users can be automatically classified to different stages of digital purchase funnel based on search query terms using machine learning. This can be divided into two sub questions:

1. What are the key considerations for the classification of text-based supervised machine learning training data?
2. What are the key differences in different machine learning technique results for collected sample data?

The research data consists of 190 245 search queries gathered from the Google AdWords (currently known as Google Ads) account of a Finnish e-commerce company over a period of several years. Search query is a word or phrase that the search engine user typed into the search field, and the query term was recorded to the AdWords campaign results data if the user ended up clicking a company's search advertisement that was shown on the search engine results pages as a result to the user's search. Hence, only the query term data from sponsored search was available for this study.

The user's intent can be interpreted from the search query (Kathuria et al. 2010, 565) and hence, they are a valuable source for analyzing consumer purchase process. Moreover, search engines are a crucial link between companies and customers and vital to the success of many online businesses (Jansen & Mullen 2008, 114, 122), as majority of consumers use search engines at some point of their purchase process. Therefore, using search query data provides researchers and businesses a way to better understand the relation between the search terms used on search engines and the consumer purchase process. This also indicates that the amount of data can be tremendous. In previous research search term classification to purchase process stages has been carried out, but manually (Jansen & Schuster 2011). This is, however, extremely time consuming and requires significant resources. Especially regarding search queries, the number of variations is infinite, and queries are ever changing (Gomes 2017). For example, vast majority of Google searches are specific long-tail queries, that consist of several words, and over 90 per cent of them get only maximum ten searches per month (Soulo 2018). Considering these facts, large-scale classification manually is not feasible.

For these reasons, the objective was to automate the classification by applying machine learning methods and develop a model that could effectively and accurately perform the classification. The researcher herself does not have required skills in programming, therefore the actual programming and model development was executed by a research assistant. Consequently, neither is the model development or running the mod-

els discussed in detail in this paper, though an outline is presented. Supervised machine learning methods are applied in this study and the paper introduces commonly used methods without going to the computational and mathematical details. Also, the machine learning process, model selection and evaluation are discussed. Machine learning is based on artificial intelligence and it can be described as developing systems that can predict outcomes by learning from input data (Bell 2015, 2). Supervised machine learning requires labeled training data. In the creation of training data content analysis was applied, as search terms were manually classified into purchase funnel stages that were constructed based on theory, whereas machine learning can be considered as automated content analysis. While one of the cornerstones in supervised machine learning applications is the creation of training data, the study aimed to bring forward aspects to be considered in the creation of training data. The main steps of the empirical part of the study included data collection, determining the classification criteria and categories, human coding, programming and training of automated classification algorithms and evaluation of the classification results. This reflects the typical machine learning process. (Raschka 2015, 11; Okazaki et al. 2014, 467.)

The central theory in the study is the theory on consumer purchase process, which is also know as customer journey (Tax et al. 2013) or purchase funnel (Hoban and Bucklin, 2015). These terms are used synonymously in this study. The classical purchase process models are presented in the paper and the applicability of the models in online context is discussed. Previous research on search term classification mainly excludes the post purchase stage, which is generally referred to as post purchase evaluation (e.g. Engel et al. 1968) in the traditional purchase process models. However, the importance of the post purchase behavior and customer loyalty is recognized among both academics and practitioners (Gounaris & Stathakopoulos 2004; Tellis 1988; Wijaya 2012; Vanarsdall 2016). Therefore, the brand loyalty stage was acknowledged in this study.

The research has implications for businesses that are looking for new, more efficient ways to process web metrics and to utilize and analyze online user data. The end goal for this is to gain better understanding of customer behavior in different stages of the online purchase process. This is to target both marketing and resources more effectively and segment audiences according to the purchase process stages, not only for search engine advertising but also for search engine optimization and content marketing purposes.

The thesis constructs as follows: the classical purchase process theories are presented in the chapter 2.1 and chapter 2.2 discusses the effects of digitalization on these models and finally in chapter 2.3 introduces prior research on consumer search behavior and query classification while also providing the basis for the classification framework of this study. Chapter 3 discusses machine learning introducing the main concepts in chapter 3.1 and chapter 3.2 presents the commonly in text classification used supervised ma-

chine learning algorithms, what the machine learning process involves and how to evaluate the machine learning results. Chapter 4 consists of the empirical part of the study first presenting the research context and the data in chapter 4.1. Chapter 4.2 explains the process of creating the training data for the machine learning classification including the classification criteria and challenges related to the task. The outline of execution of the automated search query classification is given in chapter 4.3. Chapter 5 presents the results of the machine learning classification and finally in chapter 6 conclusions, limitations, reliability and validity of research are discussed.

2 CONSUMER BEHAVIOUR AND THE PURCHASE FUNNEL

2.1 The classical models and theories

Understanding consumer buying behavior is essential for companies to be successful on the markets. Consumer buying behavior refers to, according to the definition used by Stankevich (2017) in her article, “the process by which individuals search for, select, purchase, use, and dispose of goods and services, in satisfaction of their needs and wants". Engel et al. (1968, 5) define consumer behavior as “the acts of individuals directly involved in obtaining and using economic goods and services, including the decision processes that precede and determine these acts”. The key models on consumer behavior describing consumer decision making processes or customer purchase journeys date back to the 1960s while still being topical in marketing research. The common idea behind the existing various purchase process models is to describe the path consisting of different stages consumers go through when making a purchase decision from the initial stage of need recognition to purchase. (Lemon & Verhoef 2016, 71.) The benefit of defining and understanding hierarchical processes is the possibility it gives to understand and predict customer behavior and, also, it helps in planning marketing communication strategies (Barry & Howard 1990, 107).

These purchase process models have a strong connection to hierarchy of effects models, like the classic AIDA model (Attention, Interest, Desire, Action) originally introduced by E. St. Elmo Lewis. The model was developed to illustrate the stages a consumer is taken through by the seller and later the model was used to explain the effects of advertising or marketing communication. (Lemon & Verhoef 2016, 71; Wijaya 2012, 76.) First advertising raises consumer’s attention and he/she becomes aware of the product, then he/she becomes interested in the product followed by desire for the product and finally the consumer ends up buying the product (Im et al. 2019, 26).

Later, similar expanded model was introduced by Lavidge and Steiner (1961). According to them “advertising may be thought of as a force, which must move people up a series of steps”. The model consists of seven steps (awareness, knowledge, liking, preference, conviction, purchase) that consumers take to proceed from a stage of a total unawareness of a product or service to the final stage of performing the actual purchase. In the beginning, the fully unaware potential customers stand near the bottom of the stairs. Then they become aware of the existence of the product after which they gain knowledge of the product and what it offers. In these cognitive stages consumers obtain product information through advertising. Next step involves developing favorable attitudes towards the product and when moving on one step further these attitudes evolve to a preference over other products. During these affective stages ads shape consumers’

attitudes and feelings towards the product or brand. One step away from the purchase includes consumers who have along with preference developed a desire to buy and think that the purchase would be worthwhile. At the final step the actual purchase takes place. The last two steps are conative, which means a consumer has an active motivation or intention to buy stimulated by ads, for example point-of-purchase offers. (Lavidge & Steiner 1961.)

There are several different models describing customer purchase journeys but generally these models include three fundamental stages: pre purchase, purchase, and post purchase. Pre purchase contains all the phases that consumer goes through before the purchase such as need recognition, information or alternative search and consideration. Purchase stage comprises the purchase decision and the purchase event itself. While having a lot of significance, it is temporally the most short-lasting stage. Post purchase stage covers the entire customer experience and interactions with the brand such as using the product, interaction with customer service, brand loyalty and customer engagement like word of mouth. (Lemon & Verhoef 2016, 76.)

One of the most recognized consumer decision process models was created by Engel, Kollat and Blackwell in 1968 (Hall et al. 2016, 54). According to their approach, the events that precede and follow the purchase must be explored in order to understand the act of purchasing itself (Engel et al. 1968, 7). The model describes the consumer buying process that consists of five most widely accepted stages that are problem recognition, search, evaluation of alternatives, purchase and post purchase (Darley et al. 2010, 95).

Each stage is affected by values, attitudes, personal characteristics and past experiences of the consumer. After recognizing the need the consumer searches for information and alternative brands or products. This stage relates to problem-solving because of the conscious attempts to discover several appropriate options. Next stage involves searching even more information on the alternatives. This usually occurs in nonroutine purchases and when the risk of a wrong decision is perceived high. However, both this and the previous information search stage can be bypassed if the costs of search, i.e. time and energy, outweigh the gains. This could be, for example in case of habitual and routine purchases. The buying process can also come to a halt after the search stage if no option seems suitable. Subsequently, the previous steps lead to a purchase decision, i.e. to the decision to make the purchase or not to make it. This purchase model does not end to the purchase, but the process continues with the purchase outcomes. These can include perceived doubt against the purchase, which triggers additional information search and post purchase evaluation, or the purchase can trigger a need for additional actions such as further purchases related to the purchase just made. Nevertheless, the outcomes of the purchase are stored in consumer's memory and influence the future buying processes. (Engel et al. 1968, 47-51.)

Another influential process model was Howard and Seth's model describing consumer brand choice behavior (Lemon & Verhoef 2016, 71; Stankevich 2017, 9). Like many other models, the model is based on the assumption that brand choice is systematic and a rational cognitive process. Rather than consisting of consecutive steps that the consumer takes, though, the model consists of different components one of which is the consumer response variables. These process outcomes are ordered to create a hierarchy that goes as follows: attention, comprehension, attitudes, intention and purchase behavior. Other components in the model include various factors that affect the consumer buying behavior such as personal traits, brand preferences, motives, earlier purchase experiences and barriers like high price and product unavailability. Also, impulses from the environment, such as marketing communication or word-of-mouth, affect the buyer. The model suggests three levels of consumer decision-making. In extensive problem-solving consumer has low brand preference and decision-making requires active information search. The more extensive the search, the longer it takes to get from the initiation of the process to the completion. In limited problem-solving consumer has a few preferred brands and is likely to seek information mainly to compare these different brands. In routine response behavior consumer already has accumulated enough experience and information to eliminate any uncertainty towards brand choice and hence brand preference is high. (Howard & Sheth 1969, 467, 470-481.)

The consumer decision-making or buying process is traditionally often illustrated in a shape of a funnel.

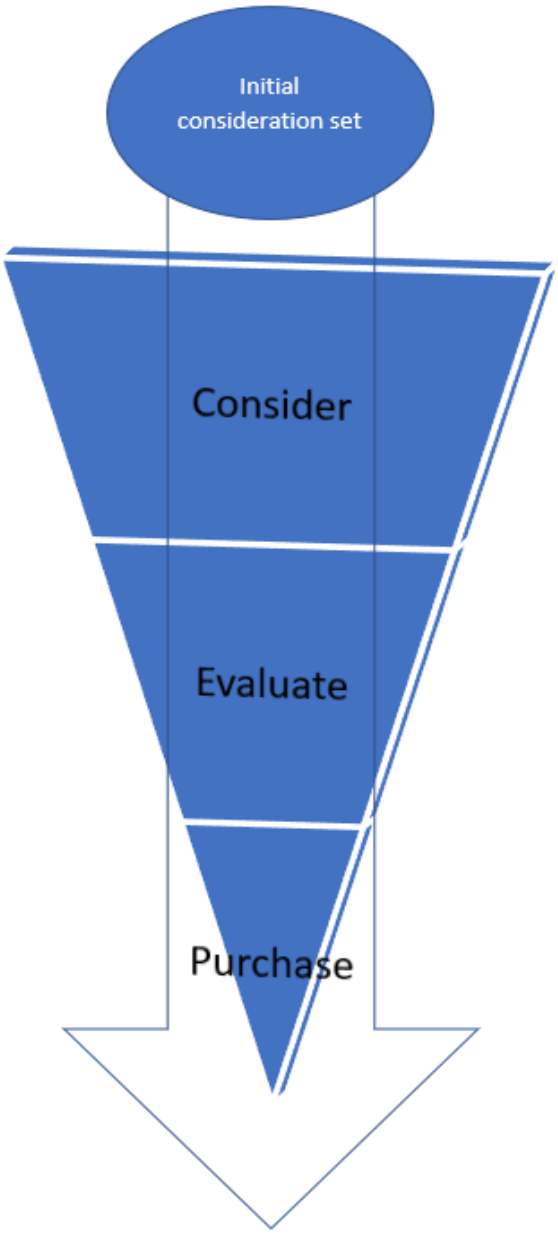


Figure 1 Traditional purchase funnel according to Hudson & Hudson (2013)

The buyer begins the process with initial large set of products or brands in mind and systematically narrows this set down until he/she has come to a decision which product or brand to buy. (Hudson & Hudson 2013, 208-209.) One example of the buying funnel with this logic is depicted in Figure 1 above.

2.2 Digital purchase funnel

Internet and social media have had a significant effect on consumer decision process and there is a lot of discussion on whether the traditional purchase funnel is valid in

digital environment (Hall et al. 2016; Jansen & Schuster 2011). Others think the funnel should be abandoned entirely (Noble et al. 2010), while others think the web has just transformed the funnel (McMillan 2007; Court et al. 2009; Haven 2007; Wolny & Charoensuksai 2014, 318-319) considering that human characteristics remain the same regardless of whether the decision making happens online or offline (Punj 2012, 800). Nevertheless, practitioners, along with academicians, have emphasized the importance of delivering the right content to consumers in web at the right time in relation to the their purchase process stage (Petrik 2014; Vanarsdall 2016; Wolny & Charoensuksai 2014, 325; Rangaswamy et al. 2009). Also, the importance of knowing the search goals of consumers to better respond to the consumer information needs is widely featured in the academic literature (Rose & Levinson 2004; Jansen et al. 2008a; Dai et al. 2006; Ashkan & Clarke 2009; Su et al. 2018) and often this goal is linked to the consumer purchase process or customer journey (Im et al 2016; Im et al 2019).

Indeed, search engines have a central role in online purchase processes since it is the main tool for consumers to search for information (Rangaswamy et al. 2009, 49.) KPMG reports that the use of online channels is very active in the research and evaluation phase globally (KPMG 2017). In Great Britain the majority of consumers, 78%, in the year 2019, use Internet to search for information on products and services (Statista 2019). Results from a survey conducted by Hotchkiss (2004) state that search engines are used more likely in the research stage of the consumer purchase process and the usage decreases as the consumer approaches to the actual purchase. Jansen & Schuster (2011) made similar findings in their study where they classified search phrase terms to the purchase funnel stages. Further, the results showed differences in the search engine usage depending on the familiarity of the product or retailer. In case of a high level of familiarity, the consumer would often go directly to the retailer website but without prior preferences or familiarity, search engines were the primary means for product research. Differences were also found between high or low involvement product purchases. With high involvement purchases the path to purchase is usually more complex and longer while low involvement purchases follow shorter and more direct path. (Hotchkiss 2004, 8-9.)

When analyzing query phrases in combination with the standard online marketing metrics, such as impressions, clicks and purchases, Jansen & Schuster (2011) found unexpectedly that awareness terms generated the most actions and the biggest sales revenue together with the purchase terms. However, the number of ordered items was double in awareness compared to the purchase stage, which indicates that the items purchased in awareness stage were less expensive. Based on their analysis, the researchers suggested that each stage of the buying funnel could lead directly to a conversion depending on the product or service, and the consumer's personal attributes. The consumer would enter the purchase funnel and stop the progressing as soon as a satisfactory

solution is found. To make a lower cost purchase consumers would not be willing to spend any unnecessary time or energy to conduct expanded search but as the price becomes higher they would perform more comprehensive search and go through all the buying funnel stages. (Jansen & Schuster 2011.) These findings are consistent with what is presented in traditional marketing theories. Already (Lavidge & Steiner 1961, 60) suggested that the bigger the psychological and economic commitment related to the purchase, the longer it will take to go through the purchase process steps and conversely, the smaller these commitments, the faster the purchase process can be completed. Consumer does not necessarily even go through all the stages. For example, in habitual buying or impulse purchase the consumer can go directly from the need recognition or awareness to the purchase. (Lavidge & Steiner 1961, 60; Kotler & Keller 2012, 166; Engel et al. 1968, 48-53.)

The line between online and offline is blurred (Rangaswamy et al. 2009, 53) and consumers can mix offline and online channels during their purchase journeys. They might be viewing physical products in-store but conclude the purchase journey to a purchase online. Or vice versa, they might conduct product research prior to purchase online but buy the product in-store. (Wolny & Charoensuksai 2014, 318.) Also, the awareness stage can partly take place offline because online search is usually initiated by a need (Jansen & Schuster 2011, 7). This can be seen from the actual consumer behavior according to KPMG report where approximately half of the consumers gained product awareness through some offline channel (KPMG 2017). Considering these aspects, web is thought to have complicated the traditional buying process (Wolny & Charoensuksai 2014; Haven 2007; Darley et al. 2010; Lemon & Verhoef 2016, 80). In between the awareness and purchase, the funnel is affected by new touch points that marketers do not control, such as product reviews or recommendations from other buyers (Haven 2007, 2).

Web is also considered to compress the buying funnel by shortening the path from message delivery to the purchase and by converging the communication channels, as well. A consumer could, for example, complete the whole purchase journey through one online banner ad. Seeing the ad can raise the awareness and interest, which makes the consumer to click the banner. He or she ends up to the seller page where the consumer can research the advertised product and finally complete the purchase, all this during the same session. (McMillan 2007.) Often practitioners in search engine marketing field use "simplified version", i.e. a shorter version, of the traditional five-stage purchase funnel consisting of, for instance, three stages: awareness, research and comparison, purchase (Hadrien 2015a).

The importance of post purchase stage, also called loyalty stage, is ever more important in digital environment (Haven 2007; Vázquez et al. 2014; Noble et al. 2010; Zhang et al. 2018). However, many models stress purchase as the end result for the cus-

tomer journey and overlook the long-term effects customer loyalty can have (Lemon & Verhoef 2016, 85). Brand familiarity or loyalty increases the likelihood to a positive response to marketing messages and it is a major factor in purchase behavior. Current purchase affects loyalty and thus future purchases. (Tellis 1988, 135, 142.) In turn, consumers who are on their way towards purchase utilize product reviews and comments, as well as other user generated content available online to a great extent to support their purchase decision. Hereby, in the digital era the experiences from current purchase affect not only the future purchases of that particular consumer but also purchases of other consumers through the reviews and comments they write online. (Vázquez et al. 2014 68-70; Okazaki et al. 2014, 467-468; Owusu et al. 2016.) The majority of consumers prefer recommendations and reviews from peers, family and friends over sponsored advertisements to support their purchasing decision (Vázquez et al. 2014, 68-69). In addition, consumers do search information post purchase, too and thus are exposed to search marketing ads. This is however ignored in prior research. (Zhang et al. 2018, 664.)

Smith and Rupp (2003) model the online consumer behavior through a decision-making process model that consists of three stages: input stage, process stage and output stage. The main factor in the input stage is the consumer's need recognition while there are external influences affecting the consumer behavior. These include the website marketing efforts and underlying sociocultural factors. For example, if online shopping is common among friends and family the person is more likely to also buy online. The process stage concentrates on how consumers make decisions. The consumer's motives, attitudes and personal traits affect his/her decision-making. These psychological factors determine how the external influences affect the need recognition, information search and evaluation of alternatives. The output stage consists of the purchase and post purchase evaluation. The consumer's experience from the buying process influences his/her psychological factors, which in turn influence the next decision-making process the consumer initiates and its outcomes. (Smith & Rupp 2003.) It is noteworthy that basically the model proposes the same concept as the traditional models presented in the previous chapter, but in the online environment taking into account specific web related factors such as online security and the convenience of online shopping.

In the Consumer Decision Journey model introduced by Court et al. (2009) instead of the traditional funnel shape, the purchase process is presented as a continuous loop where consumers continue their journey after the purchase by experiencing, advocating and bonding with the brand that leads to repeat purchases. Contrarily to the traditional tapering funnel idea, consumers may as well increase their consideration set during the decision process because of the vast amount of information available nowadays (Hall et al. 2016, 55). Another loop-like model that takes into account the influence of Internet and social media is proposed by Noble et al. (2010). This customer life-cycle model

describes the relationship between consumer and a brand during the process of consumers discovering new brands, products or needs, exploring their options, making new or repeat purchases and engaging with the product and brand. The model also emphasizes that loyalty, traditionally seen as the final stage in the purchase funnel, is not something that follows automatically after purchase but requires continuous work from marketers to earn and retain customer or brand loyalty. (Noble et al. 2010.)

Although there is a lot of criticism towards the traditional purchase funnel and many think the real purchase behavior of consumers does not conform to the funnel model (Hall et al. 2016; Noble et al. 2010; Court et al. 2009; Haven 2007; Hudson & Hudson 2013), the current research does also provide support for the purchase funnel, for instance based on consumers' search behavior and search intent (Im et al. 2019; Im et al 2016; Rutz & Bucklin 2011; Bronnenberg et al. 2016; Jerath et al. 2014). Also, Jansen & Schuster (2011) state based on their research results that the purchase funnel seems to represent actual online consumer behavior and it is suitable framework for classifying search queries, although the consumer actions did not conform to the buying funnel form based on the campaign metrics. These aspects are discussed in more detail in the next chapter.

2.3 Mapping keywords to customer purchase funnel stages

Many marketing practitioners acknowledge the important role of the purchase funnel in online environment and it is widely discussed in industry publications how the funnel should be used in targeting search engine advertising or optimizing online content (Swan 2018; Petrik 2014; Vanarsdall 2016; Hadrien 2015a). Also, researchers propose that consumer search queries and responsiveness to ads vary according to the consumer's purchase funnel stage (Jansen & Schuster 2011). However, academic research on the topic is very limited. Instead, considerably more research exists on examining the search engine user intent and the need behind a web search. It is acknowledged that different user intents or search goals indicate different likelihoods to purchase and levels of responsiveness to marketing messages (Zhang et al. 2018, 664).

Popular approach has been to study the user intent by classifying user search queries into informational, navigational, and transactional queries (Broder 2002; Jansen et al. 2008a; Lewandowski et al. 2012; Kathuria et al. 2010; Rose & Levinson 2004). According to Broder (2002) navigational queries aim to a visit to a certain web site that the user already has in mind, the intent behind informational queries is simply to find information and transactional queries refer to performing a transaction like a purchase, using online services or downloading material, for example images, songs, videos.

While transactional query by definition could refer to higher purchase intent (Kathuria et al. 2010, 578), Ashkan & Clarke (2009) and Lewandowski et al. (2012) introduced the concept of commercial queries, in addition. The intent behind these queries is solely commercial - to purchase a product or service immediately or later in the future. On the other hand, commercial intent can be seen completely as an additional dimension, besides these three other categories Broder (2002) introduced, and further divided into two commercial activity phases that reflect the stage of commercial activity a user is in, either research or commit. First users search for product information and reviews, for instance, and at commit phase they are ready to make a purchase decision and will enter a transaction page or complete the purchase offline. (Dai et al. 2006, 820-830.)

Where the above-mentioned taxonomy concerns web search in general, Su et al. (2018) introduced query-based user intent classification specifically for product search purposes. They proposed three categories called Target Finding, Decision Making and Exploration. Unlike the categories introduced by Broder (2002), these categories can be considered sequential as they represent increasingly more explorative search activities and less specific search target when proceeding from Target Finding to Exploration. At Exploration category the target specificity is low, and the user does not have a specific product in mind, only perhaps the product category or several categories. The purpose of exploration might be just casual with no immediate purchase need, for example exploring seasonal sales, or purposeful, for example looking for birthday gifts. The Decision-Making category is defined by user's immediate purchase need and the search target is more specific, i.e. the product type is known. The user would compare several products of this type in prior to making a purchase decision. Target Finding is described as the user having a specific product in mind, perhaps even a certain brand and the keywords used are specific. Product comparison does not usually take place in this category, except price comparison between retailers. (Su et al. 2018, 549.) These categories could well be comparable to the purchase funnel stages used in this study. Exploration is similar to the awareness stage, Decision Making similar to research and evaluation of alternatives, and Target Finding to the purchase stage.

Although the categories used by Su et al. (2018) are more corresponding to the purchase process used in this study, could the taxonomy created by Broder (2002) also used as a directive guideline in this study in classifying search queries into purchase funnel stages. Informational queries refer primarily to the awareness or research stages where users search information to find the best solution to their problem. Since transactional queries refer to performing a transaction such as a purchase and include for example queries with the word *"buy"* (Jansen et al. 2008a, 1256), these could be considered as purchase stage queries. Navigational queries could be associated with the brand loyalty stage since they are used to reach a specific known website and contain for example company names (Jansen et al. 2008a, 1256; Rose & Levinson 2004, 15). On the other

hand, the purpose of transactional queries is to locate a website where further interaction takes place (Broder 2002, 6), and in the case of this study it can be assumed that many users visiting the focal company website by entering the company name to the search engine, do this with a purchase intent since it is an online store site. For this reason also, it could be presumed that majority of the queries in the research data of this study have a commercial intent. Also, the source of the research data supports this presumption. The search query data originate from search engine marketing campaigns of one specific company, which means that every classified query comes from an ad click. Ashkan & Clarke (2009, 74) found that queries categorized as commercial received more ad clicks compared to the ones categorized non-commercial. They also separated commercial queries into commercial-navigational and commercial-informational and found that commercial-navigational queries had the most clicks.

Thereby, these so called navigational queries, i.e. queries including the brand name, in the research data of this study could have been classified to the purchase stage. However, it was considered essential to include the class of brand loyalty, which meant that these queries were not assigned to the purchase class but to the brand loyalty instead. According to Broder (2002, 5) search engine users use navigational queries to enter a specific website that they know because they have already visited it before or they think this site exists. Also, Hotchkiss (2004, 8) found that a high level of familiarity would make the searcher navigate straight to the specific website. In addition, Zhang et al. (2018, 663) found that more than half of the branded search queries came from users that had already made a purchase.

There are substantial arguments supporting the inclusion of brand loyalty stage in the purchase funnel classification and why this group is important to the marketers to recognize. Many studies show that search queries containing brand name generally result to higher conversion rates and generate more return on marketing investment than generic product queries (Im et al. 2016; Ghose & Yang 2008; Rutz & Bucklin 2011). Also, the very recent study of Im et al. (2019) confirms that brand terms indicate higher purchase probability. The effect is emphasized when the search query is a combination of a brand term and an experience good (Im et al. 2016, 199). This is a noteworthy finding since the focal company of this study operates in the experience gift industry. While post-purchase searchers are less responsive to ads for products they already purchased, they are still more likely to click on ads for other relevant or complementary products (Zhang et al. 2018, 663). Hence, this finding adds to the importance of recognizing the different search engine user groups for more efficient marketing. Altogether, brand loyal consumers can be considered as a valuable customer group as they have a higher tendency to purchase and they are less price sensitive, and in addition they even might recommend the brands they prefer to others (Im et al. 2016, 190, 199; Ghose & Yang 2008, 249; Vázquez et al. 2014, 68-69).

As can be noted from existing literature, user intent and search query classification is extensively studied but not from the perspective of the consumer purchase process. Very little research exists from the purchase funnel point of view. Hotchkiss (2004) conducted a survey on consumer search behaviors, this being the first studies examining the purchase funnel within the field of search engine marketing. Later, Jansen and Schuster (2011) studied whether the consumer's current stage of the buying funnel can be deduced from the search query he/she submitted. They presented criteria to classify queries into four purchase process stages labelled as awareness, research, decision and purchase, after which they compered the standard search marketing metrics of the keywords in each stage. They were able to distinguish between different stages at the individual query level and find differences in consumer behavior at each stage. Ergo, their results suggest that the purchase funnel is suitable framework for classifying search queries. The query classification was based on idea that in the beginning of the purchase process consumer search queries are broad and generic and the further in the funnel the consumer advances the more narrow and specific the queries become. (Jansen & Schuster 2011.) The complete classification criteria used in the study is shown in Table 1.

Vázquez et al. (2014) and Navas-Loro et al. (2018) classified user-generated content into consumer purchase funnel stages called awareness, evaluation, purchase and post purchase experience. They hypothesize that consumers use different expressions and linguistic elements in the different stages of the purchase funnel. Awareness stage would include texts that exhibit consumer's first contact with the brand or express opinions about the advertisements or brands. Texts in evaluation stage indicate active research about a brand or product and interest in acquiring a product or service, as well as expressing brand comparison or preference over another. Texts that particularly refer to the decision to buy belong to the purchase stage. Post purchase comprises texts that refer to a past purchase or possession of a product, as well as to an actual user experience. (Vázquez et al. 2014; Navas-Loro et al. 2018.)

Jerath et al. (2014) found in their study on consumer click behavior at a search engine that consumers can be classified into segments representing their involvement from low to high. Segments were not predefined, but they emerged from the data and the number of segments with the best distinguishability was four. Identified segments correspond to a purchase funnel model in how consumers move toward a purchase through different stages of involvement. The segments of involvement follow the purchase funnel shape, as the segments representing lower involvement are larger in size than the ones representing higher involvement. Also, lower-involvement consumers use more popular keywords in their search, whereas higher-involvement consumers use less popular keywords, i.e. keywords that have low search volumes. Such queries result to greater click-through rates and higher propensity to click sponsored links. This implies

that consumers who use low volume search queries are more invested in their information search and hence, closer to a purchase. (Jerath et al. 2014.)

In a very recent study, Im et al. (2019) examined inferring the user intent based on the search queries they made through a search session consisting of one or more queries and clicks, and in addition, across multiple sessions. The objective was to capture the purchase process of a single item of a unique individual. Based on the characteristics of search sessions, the researchers were able to identify six different groups indicating the user purchase funnel location and thus propensity to buy. They found that two of the groups have significantly higher likelihood to purchase than the other groups. By examining the breadth and depth of user search queries they concluded that users whose queries were broad and less focused and the diversity of queries were high were likely to be earlier in the funnel, whereas less broad and more specific queries predicted low-funnel user and consequently higher likelihood to purchase. (Im et al. 2019.)

Rutz and Bucklin (2011) made similar findings that consumers might first undertake online search using more popular generic keywords and later end up purchasing the product using more specific branded keywords. Other researchers have noted, too, that broad or generic search queries are more likely used in the upper purchase funnel and narrowed down to more specific in the lower purchase funnel, closer to the purchase (Li et al. 2016, 844; Spink et al. 2001, 228; Bronnenberg et al. 2016; Hotchkiss 2004, 20). Song (2009) defined broad queries as queries that cover several subtopics and clear, i.e. specific queries as queries that cover the narrower topic. An example of this from the research data would be a term "*present*" as a broad query and "*30th birthday present*" as a subtopic. Hafernik and Jansen (2013) listed several attributes indicating specificity of a query, for example query including an URL, comparisons, locations or place names, numbers, like product model numbers and names. Short keywords or queries are generally more frequently used and refer to more generic search while longer search phrases are less popular and typically indicate more specific search (Ghose & Yang 2008, 245).

Phan et al. (2007) and Hafernik and Jansen (2013) found a correlation between query length and the degree of specificity of a query. As the length of a query decreased the generality of the search increased, the threshold between broad and narrow being three words (Phan et al. 2007, 709-710). Hence, longer search queries can indicate higher purchase intent (Agarwal et al. 2011, 1079). Queries that include terms indicating commercial intent along with brand terms, and addition to that, a product term, are linked close to the purchase decision – closer than commercial or brand terms used alone (Im et al 2016, 199). As noted before, brand terms and commercial terms can be a signal of purchase intent as such, but when they are used together along with the product, the longer and the more specific the query becomes. Although, contradicting the previously mentioned findings, Skiera et al. (2010) found that the 100 most-searched keywords, which usually are not long tail, generate the most searches, clicks, as well as conver-

sions, i.e. purchases. However, they refer to the number of conversions not to the value of conversions. Referring to the Jansen's and Schuster's (2011) approach to the online purchase process, early funnel keywords can generate conversions as well, but late funnel keyword conversions might be more valuable.

Dai et al. (2006, 832) composed a list of terms that clearly convey commercial or purchase intent, such as "*price*", "*cheap*", "*buy*", "*sale*", "*purchase*", "*deal*", "*discount*", "*bargain*" and "*retail*". Also Im et al (2016) and Bronnenberg et al. (2016) observed that price appears more likely in the later stages of the search or purchase decision process when the consumer already has reduced the set of preferred alternatives. However, only small part of search queries contains these explicit commercial intention indicating terms (Dai et al. 2006, 830).

By exploring the different purchase funnel models and query classification criteria used in prior studies, the framework for this study was built. In addition, other elements related to the query or user intent, such as query length and specificity, were utilized in the query classification to the purchase funnel stages. Table 1 compiles an overview of different classification criteria relevant to this study used in previous studies, although none of them could directly be applied to this study.

Table 1 Class definitions, criteria and theories used in training data classification

Source	**Classification task**	**Category**	**Class Definition & Classification criteria**
Academic:			
Jansen & Schuster 2011	Classifying user search queries into the purchase funnel stages to determine the purchase process phase the user is in when submitting the query.	Awareness	The consumer is both conscious of a need and conscious of a desire to address that need with a product or service. The user is searching for general knowledge. The queries: • are the broadest of all queries in the buying funnel. • do not contain a brand name
		Research	The consumer engages in an information seeking process to address a need, including determining the correct product considering factors like affordability of

			the product. The consumer has decided on the type of product they want. The queries: • are still broad, but more focused than Awareness queries • do not contain brand name • contain specific product
		Decision	The consumer defines a purchase set (i.e. a limited options of possible products, services or brands and enters a decision making process among this set. The consumer is comparing alternatives. The queries: • are more focused than Awareness and Research queries • contain specific product and partial brand name • do not contain full brand/company
		Purchase	The consumer has made the decision to purchase (or not) with possible comparisons of price etc. are the most focused of any stage in the buying funnel. Contains specific product and full brand name/company
Vázquez et al. 2014 Navas-Loro et al. 2018	Classifying user-generated content in social media (tweets: Navas-Loro et al. 2018; different social media channels: Vázquez et al. 2014) into consumer decision journey stages to	Awareness	The first contact of the consumer with the product or brand, either with a desire to buy or not. Texts convey consumer's interest through opinions e.g. on advertisements or knowledge of the brand. Example texts:

	locate the consumer, i.e. the text author in the purchase funnel.		"I love Hyundai's ad.'' "I like the videos in Nike's YouTube channel."
		Evaluation	The consumer already knows the product or brand and evaluates it in comparison to other corresponding products or brands. Texts indicate interest, active product research and/or express preference. Example texts: "Looking for a second-hand Kia Sorento in NY, please send me a DM." ''Well, I'd rather fly with Emirates than with Ryanair.''
		Purchase	This stage contains texts that clearly express the decision and intention to buy or refer to the exact moment of the purchase. Example texts: "I've finally decided to switch to Movistar." "Buying my brand new blue Citroen right now!"
		Postpurchase	At this stage texts refer to a past purchase or to an actual user experience, implying to a possession of a product, also when there is no opinion expressed. Example texts: ''I bought a 2002 Citroen two days ago." ''I've been using a pair of Nike for the past two years, and I'm delighted."

Su et al. 2018	Classifying user intent based on their product search queries	Exploration (EP)	User does not have a specific target in mind. The target is unknown or known at a product category level. The user may or may not have an immediate purchase need and is browsing either casually to e.g. kill time, or purposefully to e.g. search for birthday gifts. Example queries: clothes, shoes
		Decision Making (DM)	The user has an immediate purchase need and has some idea what to buy. Target is known at product type level but the user would typically explore and compare related products of different brands or models in order to make a purchase decision. Example queries: wind coat, long sleeve T- shirt, rechargeable lamp
		Target Finding (TF)	The user has a specific target in mind with an immediate purchase need, and knows at least the product type and name but usually also the brand. Typical search in this category includes direct purchase and specific keywords. Users normally do not need to compare different products any longer, except price comparison. Example queries: Zara wind coat, iPhone 7, 3M head-wearing mask
Non-academic:			

Swan (2018)		Awareness / low intent	Consumers who are at the early stage of their journey. They use non-branded, broad search terms. Example queries: Hiking shoes, phone cases
		Considera-tion / mid intent	Consumers at this stage use more specific, long-tail keywords and may already engage in compari-son-shopping. Example queries: Northface hiking shoes
		Decision / high intent	High-intent shoppers who are searching for a specific product. Example queries: North face one trail hiking shoes womens
Hadrien (2015a,b)		Awareness	Prospect customer is in early stage of the research process, uses less relevant broad terms and has low probability of con-version. Example queries: shoes laptop computers
		Research & Compari-son	Prospect customer has a better idea of what he/she is looking for and therefore closer to buying. Example queries: mens shoes cheap acer laptops
		Purchase	Prospect customer is ready to buy, has high probability of con-version and uses long-tail key-words, which are very specific in

			nature. The further down the funnel, the more action verbs the queries will include. Example queries: red Nike mens running shoes buy Acer Aspire 15.6” laptop

As already stated, none of these classifications presented in Table 1 could directly be applied to this study. The class membership of a query depends on the type of product and whether it is high- or low-involvement product (Vázquez et al. 2014, 80; Jansen & Schuster 2011, 14). Depending on the product, the purchase process and thus the search process could differ. The researchers note that some purchase process stages may occur offline, like awareness or purchase (Bronnenberg et al. 2016; Jansen & Schuster 2011) and search engines are typically more likely used in the research stage (Hotchkiss 2004, 8). This is why, for example, the purchase stage in this study refers to the purchase decision and not to an actual purchase. Also, it is challenging to infer the intent behind a query since the same query can have several meanings and goals (Zhang et al. 2018, 664; Dai et al. 2006, 836).

A framework consisting of five purchase process stages, including the post-purchase stage of loyalty, was chosen for this study. However, during the manual coding for machine learning classification it was noticed that with that many classes the distinctness of the classes was inferior and the classes needed to be reduced. As Okazaki et al. (2014, 474) noted in their study, to reach acceptable results for precision in automatic classification a good number of classes is generally three to four. Therefore, the purchase funnel stages used in this study were reduced by one, so that there would be four classes in total, including the loyalty stage. The final purchase funnel class structure used in labeling the training data for machine learning included the stages of awareness, research, purchase and brand loyalty. The classes with definitions specific to the context of this study are presented in table 2 in chapter 4.2.4.

3 USING MACHINE LEARNING IN TEXT CLASSIFICATION

3.1 Machine Learning as a research method

Machine learning applications are becoming more and more important in numerous fields and they can already be seen in our everyday lives. Just to mention a few examples, they are employed in e-mail spam filters, in speech recognition, for example Apple's Siri, in advertising systems to match ads with keywords or to predict ad click through rates. E-commerce uses it for online shop product recommendations, medicine and healthcare utilize machine learning for example in cancer diagnosis, financial institutions use it for fraud detection and predicting stock prices. (Chen & Guestrin 2016.) As its applications, machine learning itself is multidisciplinary by nature. It combines artificial intelligence, probability and statistics, information theory and neurobiology among others. (Mitchell 1997, 2.)

Altogether machine learning enables to spot patterns in large amount of data and make predictions (Raschka 2015, 1). One of the most well-known definitions by Tom Mitchell (1997, 2) goes: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." Task refers to the problem that needs to be solved by the machine learning algorithm, for example a classification task. In order to solve the task, algorithm requires data to learn from. Thus, experience refers to the learning process, which can be guided or self-learning, in other words supervised or unsupervised learning. To know whether the learning process has been effective, i.e. the algorithm performed the task well after learning from the data, the performance of the algorithm has to be measured with a measure like accuracy or error rate. (Mitchell 1997, 5-6; Mayo 2018.)

The main types of machine learning are supervised and unsupervised learning. The seminal difference between supervised and unsupervised learning is that in the former the results, for instance the class labels, are known beforehand while in the latter the result is unknown. (Raschka 2015, 2-6.) Unsupervised learning is used to discover structures or patterns in unlabeled data and it could be seen more as data mining than actual learning (Bell 2015, 3-4). With clustering technique, data can be sorted into different groups based on the data object features. Any group or cluster that emerges as a result of the analysis consists of objects with a certain degree of similarity while being dissimilar to the objects in other clusters. Clustering can be used, for example, to discover different customer groups. (Raschka 2015, 7.)

A similar task of dividing data objects into different groups or categories is known as classification in supervised learning. In this case, the resulting categories are determined

in beforehand and the purpose is to learn from labelled training data, which consists of a set of samples. (Raschka 2015, 3.) By learning from the training data machine can then make predictions about previously unseen data (Bell 2015, 3). Hence, supervised learning is an inductive process where learning algorithm makes prediction for all input values based on the training data (Rasmussen & Williams 2006, 165). In binary classification there are only two possible classes for machine to distinguish between and a typical example is e-mail spam filtering. The machine learning algorithm learns a set of rules based on observations made from the training data that consists of e-mails that are correctly labelled as spam or not-spam. Based on these rules, the algorithm then predicts the class label of every new e-mail. Multiclass classification, in turn, refers to a classification task with several categorical classes, which is the case in this study, since there are more than two stages in the purchase funnel. (Raschka 2015, 3-6.) All the purchase funnel category labels also must be presented in the training dataset in order to have the algorithm to recognize every category. Multiclass labeling with supervised machine learning methods has been used in many social media studies and other text-based classification tasks (Okazaki et al. 2014; Habernal et al. 2014; Vázquez et al. 2014; Salminen et al. 2018a). In the next chapter the machine learning process and some of the most commonly used supervised learning techniques are discussed.

3.2 Machine learning framework and methods

Since supervised machine learning methods are employed in this study, the focus of this chapter is the supervised machine learning process and the algorithms typically used in classification tasks. The different stages of the process will be discussed and different algorithms presented.

3.2.1 Machine learning process

Typically automated text classification problems can be seen as a process of several phases that are feature extraction and dimension reductions referring to the data preprocessing followed by classifier selection and evaluation (Kowsari et al. 2019, 1-2). Okazaki et al. (2014) proposed a framework for opinion mining that included determination of a classification scheme and categories, human coding, programming of the automated classification algorithm, and evaluation of the classification results. Vázquez et al. (2014) reported the process of their study for automatically classifying short user-generated texts into the stages of the consumer decision journey that started with gathering the text data, creating criteria for classification and human coding, then followed by

data pre-processing, such as normalizing the gathered texts, developing and training the classifiers and finally evaluating the performance. Similar framework was followed also in this study, as the main steps included collecting the search terms, defining classification categories and criteria, human coding, data pre-processing, classifier development and training and evaluation of the performance. Figure 2 below presents the process generally followed in predictive modeling.

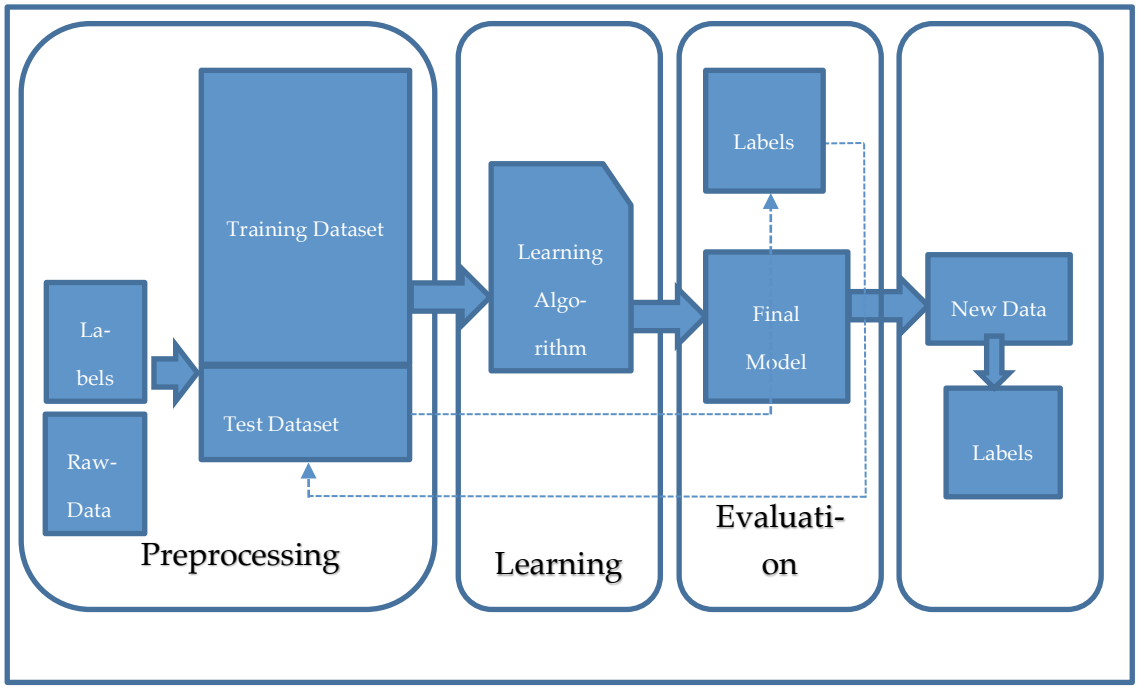


Figure 2 Typical workflow diagram using machine learning in predictive modelling according to Raschka (2015)

Any machine learning application process begins with collecting and pre-processing of the data, which is moreover a pivotal measure (Bell 2015, 17-18). Pre-processing contains data cleaning, feature extraction and selection and sampling (Raschka 2015, 11). Processing the data prior to the actual learning algorithm application is essential to obtain optimal results (Bell 2015, 43). Typically, texts and documents are unstructured data sets that first must be converted into a structured set for predictive modelling and this refers to feature extraction (Kowsari et al. 2019, 1-2). Dimensionality reduction can be done for example through feature selection techniques to avoid excess complexity (Okazaki et al. 2014, 475; Kowsari et al. 2019, 2). Furthermore, the data is divided into two separate sets, training and test set. This is to first train and optimize the machine-learning algorithm with the training set and finally to evaluate the performance of the final model with the test set to verify its applicability to new data. Most commonly used division ratios are 60:40, 70:30 or 80:20. The suitable ratio depends on the size of the total dataset, and the bigger the test set, the more valuable information is detained from

the learning algorithm. However, the estimation of the generalization error becomes more inaccurate, the smaller the test set gets. (Raschka 2015, 12, 109.) The data pre-processing will be discussed in more detail in chapter 3.2.2.

The learning phase in Figure 2 refers to the machine learning model development, training and optimization. There are several algorithms to choose from and each has its benefits and disadvantages, since different models perform differently on different data with different characteristics (Bell 2015, 42-43). Therefore, several models should be tested to find the best performing one to each particular task or problem (Raschka 2015, 12). Moreover, many practitioners state, that even more important than choosing the right algorithm is the data available for the algorithm (Schönleber 2018; Raschka 2015, 50, 99). The training dataset has to include features that are informative and discriminating enough because the learning algorithm will use these features in learning the data labels (Habernal et al. 2014, 694, 698). In Practice, the model selection, optimization, validation and evaluation go hand in hand and same techniques may be applied (Raschka 2015, 174; Rasmussen & Williams 2006, 106).

Selecting the best performing model requires evaluating the performance of different models with suitable metrics. While commonly used with classifiers, different validation processes are needed for model selection depending on the type of model (Okazaki et al. 2014, 475). Evaluation can be done simultaneously with the model training using different cross-validation techniques. In addition to evaluate model generalization performance cross-validation techniques are used for selecting the optimal parameter values to further improve the model performance. (Coussement & Van den Poel 2008, 317-318.)

Once the best model for the problem at hand is chosen and its performance optimized a final evaluation of the model performance is executed by running the model on the test dataset that was separated from the training data in the beginning. This shows how well the model will perform its actual task to predict new unseen data. (Coussement & Van den Poel 2008, 317.) This is the phase called evaluation in Figure 2. In the final stage of the process the entire unlabeled dataset is fed to the trained model to be classified. Different evaluation metrics and performance measures are presented in chapter 3.2.4.

3.2.2 Textual data pre-processing and feature extraction

As stated earlier, the cleaning of the input data and extracting representative and informative features from the data for the machine learning algorithm are essential measures at the beginning of the predictive modelling process. (Kowsari et al. 2019, 3; Habernal et al. 2014, 694) Textual data often contain noise such as irrelevant content,

spelling mistakes, abbreviations, jargon words and emoticons (Salminen el al. 2019, 205; Vázquez 2014, 73). The quality of the data strongly affects the performance of the classifier; hence majority of the time could be used for data cleaning in a machine-learning project (Bell 2015, 35-36). The purpose of raw text cleaning is to convert data into a form that is more suitable for processing at a later stage. This includes for instance separating special characters and punctuations from words and removing stop words, which are words that do not contain any significance, such as 'and', 'the', 'or', 'is', 'again', 'this', and thus do not affect classification (Coussement & Van den Poel 2007, 871-872; Salminen el al. 2019, 205; Kowsari et al. 2019, 4). Although, semantic features that imply the meaning or intention can also be important features for classification purposes, for example when classifying positive and negative emotions based on words. Then elements like special characters and emoticons might not be removed in the data pre-processing. (Salminen et al. 2018a, 336.)

While there are a wide variety of methods that can be applied to data cleaning and pre-processing, some of the most commonly used text normalization techniques include stemming and lemmatization (Habernal et al. 2014, 697). These methods retrieve the root of the word so that different conjugated forms of a word can be grouped together and processed as a single element (Vázquez 2014, 75; Coussement & Van den Poel 2007, 872). The difference between these two methods is that lemmatization retrieves the grammatically correct basic form of a word while stemming simply removes suffixes or prefixes to retain the root word that does not have to be a valid word in the language (Jabeen 2018). Although, for instance the basic form of stemming has not really been considered as a suitable method for the Finnish language due to its morphologically highly complex nature (Kettunen et al. 2005, 477). The advantage of these procedures is that they reduce the vocabulary size and hence, they provide a way to reduce dimensionality (Habernal et al. 2014, 697).

For predictive modeling purposes the unstructured textual data must be converted into a structured feature space after the data cleaning. Commonly used techniques for feature extraction are weighted words and word embedding. (Kowsari et al. 2019, 2, 4.) The most basic weighted word feature extraction technique is term frequency (TF), also known as bag-of-words. In this model, a text, such as a sentence, is considered as a bag of its words, disregarding grammar, word order and the semantic relationship between the words that makes the model simple. However, here also lays the main challenge while two sentences with a different meaning but consisting of the same words have exactly the same vector representation. (Kowsari et al. 2019, 6.) In TF the frequency of occurrence of each word is measured and used as a feature for a classifier algorithm. The logic is that the more frequently a word occurs in a text, the more important it is in characterizing the content and, thus representative of a certain class. (Coussement & Van den Poel 2007, 873.)

Yet, texts might contain overly common words, like the stopwords, that appear so often that they are not discriminative anymore (Coussement & Van den Poel 2007, 872). To overcome this, inverse document frequency (IDF) was introduced to be used together with term frequency. The commonly used method is well known as TF-IDF. (Kowsari et al. 2019, 7.) IDF reduces the effect of these common words by inversely assigning higher weight to the rear words. The underlying thought is that the more infrequently a word occurs, the more discriminating that word is. (Coussement & Van den Poel 2007, 873.) The problem of losing syntactic relation between words can be avoided by using n-gram technique. While bag-of-words processes text separating each word, i.e. using 1-gram, with n-gram technique consecutive words can be joined together and processed together as a token. For example, 3-gram combines three consecutive words of a sentence that comprise a token. The features generated this way could perceive more information than using single words. (Kowsari et al. 2019, 5.) For example, Salminen et al. (2018a) successfully applied n-grams in combination with TF-IDF for feature extraction to classify hateful comments in social media.

With word embedding methods, also the semantic meaning of the words can be taken into account (Salminen et al. 2018a, 336). While weighted words methods are based on counting word appearances in text, word embedding techniques learn from sequences of words by taking into consideration the semantic similarities between words (Kowsari et al. 2019, 47-48). Furthermore, a novel word embedding method called FastText considers the morphology of words, which refers to word conjugations, for example (Kowsari et al. 2019, 9). Probably the most advanced method by far, though, is a new technique called contextualized word representations that are able to model polysemy, which means that the same word has different meanings in different contexts. Ergo, these models learn based on the context of the word in a text and thus incorporate both the characteristics of word use, i.e. syntax and semantics, and how the uses vary in different contexts. (Kowsari et al. 2019, 10,48.)

The feature selection procedures serve for several goals. A smaller set of well-chosen and highly predictive features decreases the computing demands for the classifier and improves classification accuracy whereas an extensive feature set containing much irrelevant or noisy data increase the likelihood of overfitting. (Habernal et al. 2014, 694-695; Caigny et al. 2018, 766.) Thus, this is also known as dimensionality reduction (Okazaki et al. 2014, 475).

3.2.3 Typical classification algorithms

The perceptron, developed by Ronsenblatt, is one of the first classifiers and it had an important role in the history of machine learning. It is a linear binary classifier and

strives to minimize classification errors. (Bishop 2006, 192-193.) It constructs a decision boundary that divides data points into two categories among linearly separable set of samples, which means that the samples can be correctly classified by a straight line (Mitchell 1997, 86-87). Because of the requirement of perfectly linearly separable data, the use of the perceptron is typically not recommended in practice (Raschka 2015, 55).

Logistic regression is one of the earliest machine learning methods (Taniguchi et al. 2018, 1) and the model is widely used for binary classification (Coussement & Van den Poel 2008, 316). Despite its name, the model is not really for regression but for classification, as said (Bishop 2006, 205). It performs very well when the classes are linearly separable and it is also possible to extend the application to multiclass classification by using One-versus-the-Rest technique, which means training a single classifier per class, with the samples of that particular class considered as positive samples and the samples from all other classes considered as negatives. Logistic regression is a probabilistic model, which means that it predicts the probability of a positive event, i.e. the probability that a certain sample belongs to the given class. (Raschka 2015, 28, 56-57.)

Another popular machine learning algorithm is the support vector machine (SVM) that is based on neural network technology (Coussement & Van den Poel 2008, 314). SVMs are generally used for instance in image recognition but they are also suitable for text classification. The algorithm can be seen as an extension to the original perceptron model. (Bell 2015, 140.) While in perceptron the aim is to minimize classification errors and find the least terrible decision boundary (Bishop 2006, 193), in SVMs the purpose of the model optimization is to maximize the distance between the hyperplane and support vectors. This refers to the margin between the decision boundary dividing the classes and the training data samples situated nearest to this boundary on each side, the positive and negative class (Coussement & Van den Poel 2008, 314). Large space between the categories enhances the model prediction confidence and generalization performance (Bell 2015, 146).

When comparing SVMs and logistic regression, in practice, they often produce similar results. Logistic regression is, however, more sensitive to outliers on the data because it aims to maximize conditional probabilities, whereas SVMs are mostly concerned with the data points that are closest to the decision boundary. Yet, logistic regression is a simpler model that can be implemented and updated more easily and this is especially relevant when working with streaming data. (Raschka 2015, 74.) SVMs can be used to perform nonlinear classification tasks, as well, by using kernel methods, which means transforming linearly inseparable data from lower dimension to higher dimension when it becomes linearly separable (Coussement & Van den Poel 2008, 314). This feature reinforces the popularity of this model among machine learning practitioners (Raschka 2015, 75). Essentially SVM is a two-class classifier but can also be

transformed to a multiclass classifier by combining multiple binary SVMs for example with the same One-versus-the-Rest technique as logistic regression (Bishop 2006, 338)

K-nearest neighbor classifier (KNN) is used for text classification in many research domains (Kowsari et al. 2019, 27). It is essentially different from the learning algorithms introduced earlier. It is an instance-based learning method and instead of learning a discriminative function from the training data when they are presented, it simply stores the entire training dataset. (Raschka 2015, 92-93.) The learning or generalization takes place when a new instance must be classified by retrieving a set of similar samples from memory and classifying the new instance based on them. Because of delaying the training data processing until a new instance must be classified, the instance-based methods are also called lazy learning methods. (Mitchell 1997, 230.) K refers to the number of neighboring samples that are nearest, i.e. most similar, to the unclassified data observation. KNN classifier finds the k nearest neighbors of that observation and then assigns the class label based on the class labels of the neighbors. (Kowsari et al. 2019, 27.) Determining the right value for k is pivotal for good model performance as it affects the balance between the model over- and underfitting (Raschka 2015, 95). KNN can naturally handle multiclass classification tasks (Kowsari et al. 2019, 28) and it is fast and robust to noisy data. However, it is limited by memory constrains since most of the computation happens at the point of classification (Debaere et al. 2018, 763).

Decision tree is the most classical technique used in several domains for classification including text and data mining (Okazaki et al. 2014, 476; Kowsari et al. 2019, 32). While various algorithms have been developed for decision tree learning, they all follow the similar process of splitting data repeatedly into smaller and purer subsets constructing a flow-chart-like tree graph, as the model name implies (Bell 2015, 47; Caigny et al. 2018, 764; Kowsari et al. 2019, 32). The process begins at the top of the tree in the root node that has no parent nodes. Node means the conjunction where the data splitting takes place. The data is divided over to the child nodes based on a splitting criterion that determines the optimal splits. (Caigny et al. 2018, 764.) This means that each feature of the data samples is evaluated to determine which one of them alone can best classify the samples to different categories. The samples are then sorted to different child nodes along branches corresponding to each possible value of this feature. Typical example is a node where the selected feature for splitting the data is weather outlook. There are three possible values for this feature: sunny, overcast or rainy. A branch is grown from the node for each of the values and the data is then divided into these three different branches according to the data observation value for the outlook. After this the same process is repeated in each child node and the new split is made based on the best feature at that point in the tree. (Mitchell 1997, 53, 55.)

The process goes on until no further splits are possible, i.e. all the samples at each node belong the same category. These terminal nodes are called the tree leaves that con-

tain the resulting class labels. However, decision trees can spot even the smallest nuances on the data. Therefore it is advisable to govern the splitting process by pruning the tree and set a maximum limit for the tree depth. Otherwise the tree model can grow very deep and complex, which can result to overfitting, meaning that the model does not generalize well to new data. (Caigny et al. 2018, 764; Raschka 2015, 81.) The strong point of decision tree algorithms is their simplicity and efficiency; they perform well with reasonable computing power and they are fast both in learning and prediction (Bell 2015, 45-46; Kowsari et al. 2019, 32). Also, they do not require much data preparation and they can handle numerical as well as categorical data (Bell 2015, 46).

Random Forest is one of the machine learning models used in this study and it can be seen as an ensemble of decision trees. Ensemble learning methods are used to obtain better predictive performance by combining several weak learning algorithms. (Kowsari et al. 2019, 51-52.) Indeed, due to their good classification performance, random forests have gained great popularity. Random Forest offers a solution for the instability of decision trees. (Coussement & Van den Poel 2008, 316.) The model is based on bagging technique, which is used to decrease overfitting and to improve classification accuracy (Raschka 2015, 220).

As the name suggests, Random Forest consists of several decision tree classifiers. Randomness is added as random selection of features and randomly selected inputs are used when growing each tree. (Breiman 2001.) A set of samples is randomly chosen from different parts of the complete training set and the tree is grown from this bootstrap sample. At each node, where the data splitting is made, only a random subset of features is considered to determine which feature provides the best split, instead of evaluating them all. (Raschka 2015, 90.) Finally, after growing number of trees, i.e. a forest, the class label is assigned based on the majority vote of individual trees (Breiman 2001). Random forests correct for individual decision trees' tendency to overfit to their training set, therefore there is usually no need to set a limit for maximum tree depth (Raschka 2015, 90).

The key parameter to consider is the number of trees (Coussement & Van den Poel 2008, 316). Generally, the better the performance gets, the more trees there are (Raschka 2015 90). However, though being fast learners, random forests are quite slow making predictions and a larger forest is more complex and hence, slower. Therefore, the total number of trees should be limited. (Kowsari et al. 2019, 33) Another important parameter to optimize is the number of randomly chosen predictive features (Coussement & Van den Poel 2008, 316). In addition, the size of the bootstrap sample can be optimized. By adjusting the sample size the bias-variance tradeoff of the model can be controlled. A larger sample set decreases the randomness and hence increases the likelihood of overfitting, i.e. variance. In contrast, smaller sample sets decrease overfitting but also reduces the model performance, which refers to higher bias of the model.

(Raschka 2015, 91.) In general, random forests are robust to outliers and noise (Breiman 2001).

Another tree based learner, a novel XGBoost (Extreme Gradient Boosting) was the other classifier used in this study. As the name suggests, this model uses boosting technique, gradient boosting to be exact. (Chen & Guestrin 2016.) Both bagging and boosting are based on majority voting and they are generally used in ensemble learning models and text classification (Kowsari et al. 2019, 20-21). Unlike bagging, boosting observes the performance of previous classifiers and increases weights for incorrectly classified samples (Debaere et al. 2018, 763). As models are added on top of each other, the next model corrects the errors the previous one made, until the prediction results are accurate (Reinstein 2017). XGBoost is an additive tree model that incorporates several techniques and algorithmic optimizations that make it highly adaptive, powerful and scalable. The model uses similar feature subsampling as Random Forest, but it is sparsity-aware. This means that when a feature value needed for the split is missing, the sample is classified into the default direction, which is learnt from the data. (Chen & Guestrin 2016, 3-5.)

Altogether there are several parameters that can be optimized while the number of trees together with learning rate being the main parameters affecting the model complexity and model generalization. By reducing the influence of each individual tree, the model overfitting can be prevented. (Nielsen 2016, 44, 73-74.) Together all the attributes provide an end-to-end system that prevents overfitting in several ways by actively considering the bias-variance tradeoff when fitting models, and that can handle large scale classification problems with limited computing resources (Nielsen 2016, 92; Chen & Guestrin 2016, 7). XGBoost has been successfully implemented for example in web text classification, customer behavior prediction, ad click through rate prediction and product categorization, but as well in other domains (Chen & Guestrin 2016, 1).

3.2.4 Model optimization and evaluation

One of the most typical problems in machine learning is overfitting, which means that the model performance is good with the training data but not with the test data that is new data for the model. It is said that the model does not generalize well (Mitchell 1997, 123, 110). Overfitting is also known as a high variance, which is an indicator for consistency of the model prediction and when variance is high, it means that the model is sensitive to randomness and noise in the training data. High variance can result from having too many parameters that make the model too complex and hence, captures all the nuances and outliers in the data. Underfitting, on the contrary, means that the model is not complex enough to capture data patterns and, hence, does not perform well on

new data. Underfitting refers to having a high bias. Bias is the measurement of the systematic error of a model and it indicates how much the predictions deviate from the correct values in general. (Raschka 2015, 65-66.)

The bias-variance tradeoff and concurrently the model complexity are controlled with the model parameters (Bishop 2006, 32). There are parameters that are learned from the training data and parameters that are separately optimized, also known as hyperparameters or tuning parameters (Raschka 2015, 185). The main goal of optimizing these parameters is to achieve the best predictive performance (Bishop 2006, 32). The hyperparameters of different algorithms were introduced in previous chapter together with the classifiers, for instance the depth parameter of a decision tree. Using a separate validation set to compare the performance of different models is typical for finding the best performing model and also for determining the optimal parameter values (Nielsen 2016, 25).

However, a separate test set is already set aside for the final model evaluation and the data available for training and testing in many cases is limited. Nevertheless, it is ideal to have as much data for training as possible, but a small validation set does not provide good estimates. Commonly used technique to avoid this problem is cross-validation when model validation takes place simultaneously with learning. (Bishop 2006, 32.) Hence, different cross-validation methods are also used for model selection and estimating the model generalization performance (Rasmussen & Williams 2006, 111, 108). For example, in k-fold cross-validation the training data is randomly divided into k number of groups, which are then used for evaluation one at the time while rest of the data is used for training. Finally, the aggregate performance score of the model is obtained from the average performance of all the groups. (Salminen el al. 2019, 206.) In order to find the best performing parameters, the cross-validation procedure is repeated with several different parameter values, which is known as grid search (Coussement & Van den Poel 2008, 318). Using k-fold cross-validation in combination with grid search is a common technique for optimizing the performance of a classifier (Raschka 2015, 187). These two techniques were used in this study, as well, to validate and improve the performance of the machine learning models.

The model variance and bias can be detected and analyzed with learning curves that show the relationship between training set size and evaluation metric, for example accuracy, on the training and validation sets. The learning curve is a good tool for evaluating how to improve model performance. In a case of high bias, both the training and validation accuracy is low and it does not seem to improve with more training samples. Underfitting can be fixed by adding or creating more features or decreasing the degree of regularization, while this increases the model complexity. (Raschka 2015, 179-181.) With regularization parameters the model can be given more freedom to learn more complex patterns, or the learning can be discouraged by restricting the freedom to avoid

too much complexity (Bishop 2006, 10-11). If the training accuracy is high whereas the validation accuracy is low, the model suffers from high variance, i.e. overfitting, instead. Then the model complexity needs to be reduced or more training data collected. Complexity can be reduced by increasing the degree of regularization or decreasing the size of the feature set that reduces the noise in data. (Raschka 2015, 179-181.)

As a final step of the machine learning process, the performance of the model must be evaluated using the test set, which comprises of unseen data. This is to ensure that the model is able to generalize well to new data. (Coussement & Van den Poel 2008, 317.) There are several methods for measuring the performance but none of the single metrics is able to capture all the strengths and weaknesses of a classifier (Kowsari et al. 2019, 3, 46). A confusion matrix is typically used for deriving different performance metrics. It shows the number of samples that the algorithm predicted both correctly to each class and incorrectly to each class. In a basic two-class classification problem these are referred to as True Positives, False Positives, True Negatives and False Negatives. (Chawla et al. 2002, 322-323.)

Accuracy is the simplest and the most common evaluation metric of a classifier, but it is not suitable for imbalanced data sets (Kowsari et al. 2019, 3). It shows the ratio of correct predictions to the total number of predictions. Similarly, the error rate shows the ratio of false predictions to the total number of predictions. (Chawla et al. 2002, 323.) Because accuracy, does not consider the individual class performance, it might provide good overall values for performance even if the model always predicted the most common class neglecting the minority class (Coussement & Van den Poel 2008 317)

ROC curves (receiver operating characteristics) are graphical tools for classifier performance evaluation representing true positive and false positive rates (Kowsari et al. 2019, 46-47). These metrics are more useful for imbalanced class problems while they consider the individual class performance by comparing correct positive predictions with the total number of positives and incorrect positive predictions with the total number of negatives instead of the total number of predictions. (Coussement & Van den Poel 2008, 317; Raschka 2015, 192.) The optimal model performance on the ROC curve is where all positive samples are correctly classified as positive and no negative samples are misclassified as positive (Chawla et al. 2002, 323).

To obtain more comprehensive assessments, the metrics of precision, recall and F-score are widely adopted by researcher for multiclass classification tasks (Vázquez et al. 2014; Salminen et al. 2018a; Habernal et al. 2014) and these are the metrics employed in the model evaluation in this study, as well. Recall is actually the same as true positive rate and indicates the completeness of the prediction expressing how many samples of the positive class were labelled correctly. Precision indicates the prediction exactness expressing the portion of correct predictions in a class. (He & Garcia 2009, 1277.) Often, F-score that is the harmonic mean of recall and precision is used to provide the

overall evaluation of how the model avoids misclassification and how precise it is (Salminen el al. 2019, 205).

4 MAPPING SEARCH TERMS TO THE CUSTOMER JOURNEY STAGES

4.1 The research context and the research data

The research data originates from a Finnish e-commerce company called Elämyslahjat Oy. The company offers experience gifts and it is a leading experience gift online shop in Finland. Currently, company's wide product range includes over 1 300 experience gifts, for example hot air balloon flights, tandem parachuting, rally driving, different kind of dining experiences and spa treatments. (Elamyslahjat.fi/Tietoa meistä.) Elämyslahjat is part of an international concept that operates altogether in eight countries. The operations in Finland started in 2010. The company website gathers experiences from wide range of service providers and the customers can purchase a gift card to the experience of their choice from the website. The company mission is to change the gift giving culture from material gifts to immaterial experiential gifts. (Elamyslahjat.fi/Yritys.)

The research data was collected from the company's Google AdWords account over a period of January 2012 – December 2017 by manually exporting it from the account in a CSV file format. From CSV data was transferred into Excel format. The data consisted in total of 190 245 search queries that users typed in to the search engine and that led to a click of a search engine advertisement of the company.

Google AdWords is a platform used for search engine advertising. There are also other forms of ads that can be created in AdWords, such as display banner ads, but for this study only the search advertising is relevant. By using an AdWords account a company can create ads that are shown to search engine users along with the organic, non-sponsored, search results when they type in queries (Google Help Center). Through the search queries advertisers obtain valuable information on consumer interests giving the possibility to target ads directly to these interests (Misra et al. 2006, 1). Prior research has observed, though, that search engine users have a bias against sponsored results and favor non-sponsored links over the sponsored regardless of their relevancy (Jansen & Resnick 2006, 1959). Considering this, incorporating the research company web site's organic traffic search queries to the research data, would have offered valuable additional information and perhaps a different point of view to the study. Unfortunately, this was not possible for this study. Hence, the research data consists solely of data acquired from the company's paid search advertisement campaigns.

The ads that are created in AdWords are linked to keywords defined by the advertiser. Keywords are words or phrases that potential customers might use when they are searching for products or services the advertiser offers. When search engine user types in a query term that an advertiser has defined as a keyword to its ads, the ad can be

shown to the searcher among the search results and as a wanted result, searcher clicks the ad and ends up to the company web site. Commonly used price strategy is to pay for each ad click and the advertiser defines the maximum sum it is willing to pay for the click. (Google Help Center.) This is why search engine advertising is also known as keyword advertising or pay-per-click (PPC) advertising (Jansen et al. 2008b, 2). However, the query term does not have to be exactly the same as the keyword defined by the advertiser. For long, Google has counted in close variants such as words with misspellings and plurals and the latest changes in the algorithm indicate that there has been a shift towards consumer intent in the way the algorithm pairs up ads and search queries. Now, for example, also synonyms, paraphrases and different word orders are considered as close variants. (Neely 2019b.) Therefore, advertisers do not have to include every single possible term the searchers might use in their keyword lists. That also means that the data used in this study for search queries that were used to enter the company web site is much larger than the list of keywords.

A search query can also trigger several keywords or phrases when the search engine algorithm then determines which of an advertiser's ads is served in a response to the query (Jansen 2011,181). Because it is hard to predict any exact search queries, especially when 15 percent of daily searches are new (Gomes 2017), this kind of broad-match keyword type is commonly used. However, it might not always result to the best match between the query made by the searcher and the keyword and ad chosen by the search engine algorithm when considering the searcher's intent. In the research data of this study this issue emerged when the search queries and the assigned keywords were examined. For instance, exactly the same search query typed in Google *"30 vuotislahja miehelle"* (30th birthday present for a man) was in one case matched with a keyword *"lahja miehelle"* (present for a man) and in other case with *"synttäri lahja"* (birthday present), which is inconsistent.

In another example the mismatch of the query and the keyword would have led to a misclassification. The very specific search query *"1 kpl lahja elokuvalipun hinta tampere"* (1 piece present movie ticket price Tampere) was matched with a very broad keyword *"lahja"* (present). While the searcher's query refers to the end of the funnel, purchase stage, the keyword instead refers to the very first funnel stage of awareness. Thus, it can be presumed that search queries made by the search engine users give more accurate data on consumer intent than the keywords, which is the central idea in this study. Also, the algorithm that matches queries, keywords and ads has gone through several changes during the data collection period from year 2012 to 2017 and it will be ever changing. This makes the search query data more reliable for the purpose of this study and for applying machine learning methods.

Google has the worldwide dominance on the search engine market with 92.46% share. Bing holds the second place with 2.45% share while Yahoo! being third having a

market share of 1.82%. In Finland Google's dominance is even bigger, 97.1%, and on a third place with Yahoo! is DuckDuckGo. (Statcounter 2019.) There are several ad networks, but Google AdWords is the most popular, and one obvious reason to that is the high volume of searches on Google. Over 3.5 billion searches are made every day. (Raehsler, 2019.) If the company, whose data is applied in this study, had search engine advertising additionally in some other ad network, it would be interesting to include that data to see if there are any differences, for instance, in the search behavior.

According to IAB Europe's 2017 AdEx Benchmark report (2018) the European digital advertising market has showed continuing growth and the market size has doubled during the five-year period of 2012 – 2017. In 2017 the total digital advertising expenditure in Europe totaled €48.0 billion and search engine advertising had the largest share (45.7%), and it has been this way since 2009. (IAB Europe 2018.)

4.2 Creating the training data

4.2.1 Content analysis

Content analysis is a widely used method in marketing research (Okazaki et al. 2014, 469.) According to the original definition of Berelson[3] (1952) content analysis is "a research technique for the objective, systematic and quantitative description of the manifest content of communication". Later Stone et al.[4] (1966) defined content analysis as "a research technique for making inferences by systematically and objectively identifying specified characteristics within a text", hereby acknowledging the inferential nature of coding and categorizing textual material. (Krippendorff 2004, 19, 25.)

Moreover, according to Krippendorff (2004) drawing conclusions is pivotal in content analysis. In content analysis the researchers make objective interpretations from available texts or other material, like images or recorder speech, to find an answer to their research questions while the researchers' own background and experience combined to theoretical knowledge affect the conclusions they come to. Texts and data have different meanings in different contexts and they are examined in relation to the context constructed by the researcher. However, as a research technique, the analysis is expected to be reliable, as replicability is the most important form of reliability. (Krippen-

[3] Original source: Berelson, Bernard. (1952) *Content analysis in communications research*. Free Press. New York.
[4] Original source: Stone, Philip J. – Dunphy, Dexter C. – Smith, Marshall S. – Ogilvie, Daniel M. (1966) *The General Inquirer: A computer approach to content analysis*. MIT Press. Cambridge.

dorff 2004, 19, 38, 24, 33, 18.) It means that different researchers, as well as researchers working at different points in time, should be able produce similar results when applying the same technique to the same data (Salisbury 2001, 67). Therefore, the underlying context of the analysis and conclusions must be explicated (Krippendorff 2004, 24). Unlike Berelson, Krippendorff (2004, 87) also sees that quantification is not pivotal in content analysis, but rather quantitative and qualitative approaches are both substantive for the analysis of texts because text is qualitative by nature to begin with. As a process, content analysis consists of several steps involving developing categories for coding content, training coders, coding the material and finally analyzing the results statistically (Salisbury 2001, 67).

There are different approaches to content analysis. In inductive approach, also known as conventional content analysis, the analysis begins with examining of data. The researcher becomes absorbed in the text and let the categories emerge directly from the data and these categories are then used to organize the data into meaningful clusters. Inversely, existing theory and prior research forms the basis for deductive content analysis that is also known as a directed approach. The categories for classification and their operational definitions are determined based on theory prior to the analysis. (Hsieh & Shannon 2005, 1279-1281.) This is also the approach taken in this study since the categories for classification and initial coding criteria were created based on existing theory on the purchase funnel stages. Further, these initial criteria were modified as the manual coding proceeded (Hsieh & Shannon 2005, 1286).

Krippendorff (2004) argues, though, that neither deductive nor inductive approach is focal in content analysis, but rather content analysis is making abductive inferences from data. It involves examining what texts reveal about the phenomena in that context and finding the most likely explanation through logical reasoning. Content analysis is not just describing the attributes of texts but also their meanings. (Krippendorff 2004, 36, 346, 344, 85.) Indeed, the manual coding done prior to the automatic classification might be considered as abductive. Inferences that were abductive by nature were made when the researcher reasoned in which stage of the purchase funnel the search engine user was based on the search query.

Nowadays, there is a vast amount of digitalized data available online for academic and commercial marketing research. However, the data are predominantly unstructured and qualitative by nature, which makes it noisy and difficult to quantify and transform to valuable information and knowledge. (Netzer et al. 2012, 521.) Also, because of the abundance of data, the traditional manual content analysis methods can hardly be used (Okazaki et al. 2014, 468). The wide availability of automated tools for text processing tasks has made content analysis more popular research method again (Krippendorff 2004, 17). While manually executed content analysis often suffers from low replicability, mainly due to inconsistent coding practices, more objective results can be obtained

from machine coding and the classification can be repeated consistently, once the machine learning algorithms are programmed (Okazaki et al. 2014, 469).

In addition to objectivity, computers can effectively and efficiently process large amounts of data while for a human, the analysis of qualitative data is very time consuming. Computers, though, lack the contextual awareness of a human and the ability to make sophisticated interpretations based on the context. (Canhoto & Padmanabhan 2014, 1-2.) These interpretations can also be hidden in the process of coding done by a human in addition to the end results of a study. Due to these different abilities and advantages content analysts can be seen working in a synergy with computers, them being part of the methodology. Computers systematically and reliably process large volumes of text and humans understand and interpret these written documents giving them a meaning. The purpose of methodology is to provide the means of describing the research process itself. It enables researchers to plan, execute and critically evaluate their analyses. The reasoning behind decisions and inferences made must be explicit to others, so that the research may be replicated with similar results. This also refers to the validity of a research. (Krippendorff 2004, 36, 14, 81, 340, 313.)

Computer-aided content analysis commonly utilizes automated tools for text processing or text mining (Okazaki et al. 2014, 470). Text mining means deriving relevant and useful information from unstructured text (Netzer et al. 2012, 523). Typical text mining tasks comprise for example text categorization or clustering, information extraction, sentiment analysis and document summarization (Tang & Guo 2015, 69). Commonly applied text mining techniques involve machine learning (Okazaki et al. 2014; Netzer et al. 2012), which is also the method used in this study. Yet, the training of a machine learning algorithm is based on a data manually coded by a human (Okazaki et al. 2014, 472), which means that initially human analysis is required.

4.2.2 Manual classification of the research data

Prior to applying any supervised machine learning methods, a training data needs to be created to teach the machine the connections between each class and samples, in this case search terms, relevant to that class (Bell 2015, 3). Creating the training data required manual classification of the search query terms. The manual classification was executed by the researcher herself. Classification was conducted based on prior literature on customer purchase funnel and query classification presented earlier in the chapter 2 while considering the machine learning requirements. This included constructing the classes so that they were distinct i.e. well separable, and data samples incorporated in them had discriminating features (Raschka 2015, 97). Each class in the training data should provide a representative and comprehensive set of samples (Raschka 2015, 4)

and the classes should be well balanced; otherwise the machine learning model is not likely to generalize well (Bishop 2006, 45; Taniguchi et al. 2018, 1). The classes for search query classification were derived from the consumer buying process models developed by i.a. Engel et al. (1968) and they were as follows: awareness, research, evaluation of alternatives, purchase and loyalty. The generally accepted idea was followed that the search query specificity reflects the user intent specificity and the more specific the user intent the closer to the purchase the user is.

The 190 245 search queries exported from the AdWords account were presented in an Excel file where queries were placed in the rows one below the other. At the next column there was a dropdown menu with the purchase funnel stage options and the suitable class was manually selected from the dropdown menu for each search query term. The same dropdown menu was available again at the third column for a possible secondary class as many search terms could belong to several, at least to two different classes. Even when following the classification criteria based on literature, an unambiguous solution could not always be found. While the framework of the customer journey is established, the purchase process itself and consequently the embodiment of purchase funnel stages are dependent on the industry and product in question, for example low- vs. high-involvement products (Lavidge & Steiner 1961, 60; Vanarsdall 2016, Kotler & Keller 2012, 166, 173). Therefore, there is no universally applicable definition for the classes.

In addition, in manual classification the human factor has to be taken into consideration as the interpretations of the person implementing the classification affects the data analysis (Krippendorff 2004, 24-25; Taniguchi et al. 2018, 1-2). It is possible to have secondary classes in manual classification but machine, however, can only select one class. In the course of classification, the unclear cases of assigning the correct class to a search query term were discussed with an expert with experience both in search engine advertising in this particular company and the gift industry. The researcher herself does not have knowledge of gift industry but do have experience in conducting real search advertising campaigns in Google AdWords.

Because of the vast amount of data, the classification was started by sorting the query terms by most frequently appearing terms. From the total count of search queries more than 50 % was individual terms, which means that a search with exactly these words or phrases was conducted only once. The data consisted of altogether 124 558 different search terms. The most often used terms were *"lahja"* (present), which appeared 65 times, *"lahja äidille"* (present for mother), which appeared 55 times and *"lahja miehelle"* (present for a man), which appeared 53 times. All these terms are rather broad and refer to the beginning of the purchase funnel. In fact, the vast majority of at least the top 100 search terms did belong to the first two funnel stages, awareness and research, and these classes ended up being dominating.

Consequently, concentrating on classifying the top keywords resulted having rather uneven distribution of search terms among the classes. This is because the majority of high search volume query terms are broad, one- or two-word queries and the more focused phrases, with three or more words, that more likely refer to the end of the purchase funnel, usually have low search volumes (Soulo 2018). Therefore it is unlikely that such a long-tail keyword is included in the top search terms. Unevenly distributed classes in the training data could hinder the learning of the machine learning algorithm. The classifier would be less likely to classify terms to the classes with significantly less training data (Qian 2017; Chen et al. 2004, 2).

In order to get acceptable number of samples to all classes, the terms referring to the classes with fewer samples were searched from the data by filtering. Filtering was based on the class definitions and classification criteria. The classes with clearly too few samples were purchase and brand-loyalty. As it was decided for the classification criteria that, for example, words like "price" indicate purchase intent and therefore terms including this word were to be placed to the purchase class. Hence, the research data was filtered by word "price" and several different, mainly long tail, search queries could be located at once. Similarly, to find samples to the brand-loyalty class, the data was filtered by different brand names related to experiences or experience gifts, such as restaurant chains, adventure parks and other service providers. This required exploring the focal company web site to discover the different service providers as well as researcher's own knowledge of other Finnish service providers in the industry.

As proposed earlier in the chapter 2, the main principle in assigning class labels for the search terms was that when the searcher is in the beginning of the purchase funnel his/her queries are very broad and unspecific, mainly one- or two-word terms. The closer to the purchase decision the searcher gets, the more precise and longer the queries are. In manual classification the models presented in the chapter 2 were accommodated to the context of gift purchasing process. In the manual data coding process, the researcher had to make inferences and assumptions about the searchers' intentions in this underlying context, as the goal in the classification was to build a logical bath that describes the customer journey within this company through the search queries.

In awareness or problem recognition stage the customer has a general need to find a gift or get gift ideas and is looking for an answer to the question "what to buy as a gift". Queries in this stage are broadest and shortest of all queries included in the data. Below are examples of search queries found in the research data that were annotated to the awareness class:

- *lahja* (present)
- *mitä lahjaksi* (what to buy as a gift)
- *lahjavinkit* (gift ideas)
- *hyviä lahjaideoita* (good gift ideas)

In research or information search stage, the consumer has a need to find a specific gift, and searches information to find the best solution to the problem. He/she knows more specifically what kind of a gift he/she is looking for or to whom the gift is for and might be searching for example what to give as a present to a 30 year old man. Search queries in this stage are more specific than in the previous stage but still typically quite broad. Below are examples of search queries found in the research data that were annotated to the research class:

- *elämyslahjakortti* (experience gift card)
- *30 vuotislahja miehelle* (30th birthday gift for a man)
- *aineettomat lahjat* (immaterial gifts)
- *joululahja poikaystävälle* (Christmas present for a boyfriend)
- *romanttinen lahja* (romantic gift)

In evaluation of alternatives stage, the consumer has limited the options to choose from and is comparison-shopping to consider alternative experience gift options. Below are examples of search queries found in the research data that were annotated to the evaluation of alternatives class:

- *lentoelämys lahjaksi* (flight experience as a gift)
- *ravintolalahjakortti (restaurant gift card)*
- *teatterilahjakortti* (theatre gift card)
- *tandemlaskuvarjohyppy lahjaksi* (tandem parachute jump as a gift)

In the purchase stage, the consumer has found the right solution and is willing to buy. The search queries used indicate purchase intention. They can include words like "price" and "buy" because the assumption here was that the purchase decision has already been made or the consumer is close to making the decision and is for example just comparing the prices. Queries are the most specified in this stage of the purchase funnel. They specify the experience and possibly the place of purchase or the area where the experience should be available. It is taken into account that the actual purchase can happen offline, as well. Below are examples of search queries found in the research data that were annotated to the purchase class:

- *ravintola lahjakortti netistä Turku* (restaurant gift card online Turku)
- *elämyslahjakortti kuumailmapallo Helsinki* (experience gift card hot air balloon Helsinki)
- *drinkkikurssi hinta* (cocktail course price)
- *hierontalahjakortti hintavertailu* (massage gift card price comparison)

Loyalty stage is a post purchase stage where the consumer has made a purchase from the company before and is already familiar with the brand. He/she is returning for more purchases or other interaction and is making search queries directly with that brand name. In the course of labeling the samples, three subclasses were created to the loyalty class: retailer brand loyalty, supplier brand loyalty and competitor brand loyalty. This was to differentiate which brand is in question; the focal company Elämyslahjat, a company supplier or service provider brand, such as amusement and activity parks or restaurants, or competitor brand, whose products the focal company does not provide. In case the search term included a supplier or competitor brand name including the focal company brand, the term was labeled to the retail brand loyalty subclass assuming that would be the preferred place of purchase. Below are examples of search queries annotated to these three loyalty classes.

Retail brand loyalty:

- *www elämyslahjat fi*
- *elämyslahja fi*
- *elämys fi*
- *elämyslahjat*
- *lahjakortti elämyslahjat*

The first three examples obviously refer to the company web site even though it is not in its complete form, e.g. *"elämys"* (experience). The fourth example *"elämyslahjat"* could refer either to experience gifts in general or to the focal company. However, in this case it had to be annotated to the brand loyalty class for the consistency. In the last example, it is quite certain that the word *"elämyslahjat"* refers to the company judging from the preceding word "gift card" – the consumer wants to buy a gift card from the company Elämyslahjat.

Supplier brand loyalty:

- *flowparkin hinta turku* (Flowpark price Turku)
- *fastmotors elämyspaketti* (Fast Motors experience package)
- *helsingin kahvipaahtimo kahvitasting* (Helsingin kahvipaahtimo coffee tasting)
- *funpark tarjoukset* (Funpark offers)

As can be seen from the first example, when the word "price" was connected with a brand name, the query was annotated to the brand loyalty class.

Competitor brand loyalty:

- *ravintolalahjakortti muru (restaurant gift card Muru)*

- *greatdays elämyslahjakortit (Greatdays experience gift cards)*
- *elämystaikurit*
- *elämysmatkat*

The first example was annotated to this group because even though the focal company offers restaurant gift cards, they are not available to this specific restaurant called Muru. The rest are competing experience providers. Although, it could be seen from the search query phrases in the data that in many cases the consumers clearly confused the company Elämysmatkat with the focal company Elämyslahjat. Ergo, the actual customer intent referred to the retail brand loyalty class, but the queries had to be annotated to the competitor brand loyalty class for clarity.

Manual classification was challenging and some modifications were made to the classification criteria during the process. The class definitions evolved as the research data were examined and labeled. The aim was to ensure that the classification was logical and the classes were balanced as well as exclusive and discriminating enough. In the next chapter the challenges encountered in manual classification will be discussed in more detail.

4.2.3 Challenges in classification

Classifying text-based data is not a simple task for a human let alone for a machine. The process begins with defining the classes and even though there are purchase funnel theories and models as a basis they are still not directly applicable as such to every industry and context, as it was noticed here. For instance, it can be discussed, whether the customer journey in online context, and in the context of this study, always begins in the awareness stage, goes through all the stages and ends to the purchase stage. As it has been noted before, the funnel might differ in different contexts. It might be shorter or more complex or skip some stages. (Haven et al. 2007, 2; Hotchkiss 2004, 8-9; Jansen & Schuster 2011, 4, 14).

Also, classification requires finding out what the customer intention behind the search query is and that is probably company- or, at least, industry-specific. Furthermore, the class boundaries are not unambiguous. The research data included several terms that were in between two classes, so to speak, or could have been assigned to several different classes depending on how the intent behind the searcher's query or the logic behind the classification criteria was interpreted. This means that the human factor had a great role. Altogether, creating a proper training data is a time-consuming task and requires a lot of manual work. Although, once it is done, it can be leveraged later in

other classification or machine learning tasks within the same company or similar field of business. (Bell 2015, 3.)

Firstly, the boundary between awareness and research classes was quite challenging to distinguish. The initial idea to include only the most general terms regarding presents to awareness resulted in a very small class and in contrast to a huge research class. Then the challenge was finding the terms that could be added to awareness without creating any conflicts with the terms in research class, i.e. finding the boundary that rationally separates awareness from research stage. In fact, the same challenge was encountered with all the other class boundaries, as well. When the separation between research and evaluation of alternatives seemed good all aspects considered, the purchase class was almost empty. When data was added to purchase, the distinction between classes became ambiguous.

Thus, finding the balance between all the classification criteria proved to be difficult in word-based classification. Even when the logic behind the class division and the inferences made were clear from the human perspective, from the machine learning perspective the data would have offered contradicting information if there had been similar terms labeled in several different classes. As an example, the queries including a brand name were annotated to the brand loyalty class. However, in Internet consumers search a lot for others' experiences on products and brands when they are doing their research. As also in this case, the search query data included quite a few queries including a brand name and the word "experiences" and it would have been intuitive to annotate these queries to the research class. For the consistency though, they were annotated to the brand loyalty like all the other queries including brand names.

Purchase class was probably the most challenging to define in this context. It is quite difficult to determine whether the searcher is ready to purchase or just comparison-shopping. The only thing that can be deduced with certainty is that someone is either further from or closer to the purchase. Using the word "buy" in the query is a clear signal for buying but in the data it was more often connected with a brand name, which means the query had to be labeled as brand-loyalty. Also, labeling queries including a brand name to brand-loyalty required making assumptions. Without knowing the search history of that particular person, it is impossible to know for sure whether he/she already was familiar with the brand or just became familiar with it during the same session while first using some broader search terms.

To be able to automatically classify search query terms by using machine learning, however, the classes need to be explicit and semantically similar terms labeled to the same class while having approximately the same amount of data in every class. In practice, this proved to be a difficult equation to solve. To manage this some assumptions and compromises needed to be made. The Finnish language provided some additional challenge. For example, according to the classification criteria the terms including the

word *"hinta"* (price) would indicate intention to purchase, e.g. *"helikopterilento helsinki hinta"* (helicopter flight Helsinki price), and hence, should be assigned to the purchase class. However, the very similar word *"hintainen"* occurred in contexts that refer to the research stage rather than to purchase, e.g. *"minkä hintainen 50 v lahja pitäisi olla"* (How much should a 50th birthday gift cost). The meaning is different but the word itself is very similar, thus if these terms were assigned to different classes, would the machine be able to tell the difference. In this search query *"50-vuotislahjan hinta"* (50th birthday gift price) the word "price" is in exactly the same form but judging by the other words included in the query, the searcher is not yet in the purchase stage but searching for information.

Another example of a challenging term is the word *"elämyslahjat"*, which is the focal company name and therefore refers to the loyalty class according to the classification criteria. Yet, this same word in the Finnish language means experience gifts and the searchers might use it when they are looking for experience gift ideas in general. In some cases it was quite impossible to distinguish whether the consumer was searching for the brand or experience gifts in general. It could be presumed that if the word appears in a singular form *"elämyslahja"*, it refers to the generic term and therefore belonged to the Awareness class. The plural, in turn, would more likely refer to the brand. There is only one letter difference, though, between these two words. A human can tell the difference but how about a machine.

The research data also included some irrelevant terms that had nothing do with the company Elämyslahjat or gift seeking, for example "Katherine Kelly Lang husband" and "bar fight Turku". The terms might have ended up to the list of search queries by accident. Due to the way search engine algorithm matches queries and ads, the searcher's query could have launched the company's ad and the searcher then just accidentally clicked the ad link without having any intention to find gifts. Because the number of search terms is vast, it was not possible to go through all the data and remove these irrelevant terms before classification.

The size and diversity of the data can pose a challenge for an automated classifying to some extent, since it was not possible to include all possible examples in the training data and clean the data from all the irrelevancies. In addition to that, the input data was unclean with countless spelling mistakes and compound words sometimes written separately, sometimes together, because the data consisted of something the users had freely typed in without any restrictions. The data cleaning is one essential and time-consuming step in a machine learning project but the cleaner data can be supplied the better the results will be. (Bell 2015, 31-36, 43.)

4.2.4 The classification criteria and training data for machine learning

Finally, the classification done by the researcher was checked by another person to see whether the classification criteria were logic and rational. The distinction between the classes research and evaluation of alternatives was found to be too indistinct and the classification was not reliably repeatable. Hence, it was decided to combine these two classes consequently resulting to a reduced number of purchase funnel stages from four pre purchase stages to three pre purchase stages. Combining the two overlapping classes clarified the classification and reduced the number of ambiguous cases, which also meant that the probability of successful machine learning application was improved. This type of a purchase funnel model where there are three pre purchase stages is applied in previous studies (Hoban & Bucklin 2015), as well as among practitioners in online context (Petrik 2014; Vanarsdall 2016; Hadrien 2015a). Of course, having fewer classes somewhat limits the informativeness of the results because the class definitions were expanded, but in this case the success of the machine learning method had to be considered.

As the research and evaluation of alternatives were combined, were the definitions of other pre purchase classes modified as well to balance the sample distribution between classes. In the initial classification there were terms that were sort of in between the classes of awareness and research. For example, quite general search term like “romantic gifts”, “Christmas present” and “immaterial gifts” that previously were annotated to the research class were now moved to the awareness class. This meant that the definition of awareness stage was extended. The searcher already might have a better understanding of what kind of a gift he/she is looking for. This alteration clarified the search query labeling between these two classes and restrained the dominance of the research class. The search queries originally annotated to the evaluation of alternatives class were distributed between research and purchase classes. The more general terms were annotated to research and the more specific to purchase, which means that also the purchase class was extended to cover a bit more general terms than initially. For example, the term “theatre gift card” was moved to purchase instead of research with the idea that it is already a specific product.

The purchase funnel model that was eventually applied and the operational definitions for the classes are presented in Table 2. These classes also represent the dependent variables, i.e. the output of machine learning while they are dependent on the input of the machine learning classifier, i.e. the features that define the class label (Coussement & Van den Poel 2008, 319). The search query examples in the table are translations from Finnish to English.

Table 2 Classification framework with definitions and example queries

Class	Definition	Example search queries
Main classes		
Awareness	The searcher has a general need to find a gift or get gift ideas, and is looking for an answer to the question **what to buy as a gift**. He/she might also already have more specific intention to find an **experience as a gift**.	Gift ideas Gift online Experience as a gift Birthday present
Research	The searcher knows more specifically **what kind of a (experience) gift** he/she is looking for and/or **to whom** the gift is for – searching an answer to **what to buy as a present to a 30 year old man** or comparing a set of options.	30th birthday gift for a man Driving experience as a gift Christmas present for a boyfriend Food experience as a gift
Purchase	The searcher has made the purchase decision or is close to making the decision. The search queries indicate purchase intent including words like **"price"** and **"buy"** and/or **specifying the experience** that is going to be bought as a gift.	Buy restaurant gift card Rovaniemi Cocktail course price Massage gift card price comparison Experience gift card for a wedding couple
Brand loyalty	The searcher is already familiar with a brand and is making search queries directly with the brand name.	
Subclasses		
Retail brand loyalty (RBL)	The focal company's brand name Elämyslahjat is used in the search query.	elämyslahja fi elämyslahjat Gift card elämyslahjat Hotel gift card elämyslahjat
Supplier brand loyalty (SBL)	The focal company's service provider brand names are used in the search query. (supplier = a company whose products or services the focal company is selling)	Flowpark price Turku Fast Motors experience package Funpark offers Tony's Deli brunch gift card
Competitor brand loyalty (CBL)	The focal company's competitor brand names are used in the search query (competitor = another company selling gift cards or experiences)	Elämystaikurit Buy Silja Line gift card Kokemuskauppa chocolate tasting

Altogether, 4469 search terms were manually labelled, representing 2.35% of the total search query term count. All the terms that had a secondary class label were assessed once more following the revised classification criteria and the final class label was as-

signed and the secondary class removed, except for the brand-loyalty subclasses. This was necessary in terms of successful machine learning application. The detailed sample distribution to the classes can be seen below in Table 3.

Table 3 Distribution of classes in the training data

Classification	Class	Secondary class	Count of Search term	% of Total data	% of training data
Labelled training data	**Awareness**		844	0.44%	18.89%
	Research		1 763	0.93%	39.45%
	Purchase		505	0.27%	11.30%
	Brand loyalty	Competitor brand loyalty	191	0.10%	
		Retailer brand loyalty	204	0.11%	
		Supplier brand loyalty	962	0.51%	
	Brand loyalty Total		1 357	0.71%	30.36%
Labeled training data Total			**4 469**	**2.35%**	
Not Labeled data			185 776	97.65%	
Grand Total			**190 245**	**100.00%**	

Table 3 shows that after reducing the number of classes, the research class was still predominant, just a little less than before though. Besides that, the brand-loyalty class contained the second highest number of samples. Between the loyalty subclasses the distribution of samples was clearly focused on the supplier brands. Over 70% of the brand queries included a supplier brand name while competitor and the focal company brand names had only about 14% share each.

There is no definite amount how much training data is needed. It depends on the variability of the data, the classification problem and the machine learning model used. (Bell 2015, 3; Bishop 2006, 2.) Studies have shown that a small training as well as testing sample size can hinder the design and evaluation of the model performance and the more complex model the more training and testing samples are needed (Raudys & Jain 1991, 252, 262). A representative sample is needed so that the model can capture the relationships between the data features. If outlier cases are not included in the training data, the probability of the model to predict the class labels correctly for those cases is low. (Brownlee 2017.)

The training sample size is one factor affecting the classifier performance. It affects the model bias and variance that indicate the prediction errors a classifier makes. (Beleites et al. 2015, 29.) As can be seen from Table 3, the classes were still imbalanced after the modifications; hereby only the problem of overlapping classes was resolved. The imbalance of the classes might well represent the natural structure of the data and often,

in fact, real data is imbalanced (Taniguchi et al. 2018, 1). However, this could have a negative effect on the model learning because unevenly distributed data means that also the features that the algorithm uses for learning are unevenly distributed. Especially Random Forest, the algorithm used in this study, can be strongly affected by imbalanced data distribution and according to the study of Taniguchi et al. (2018) Random Forest is actually more sensitive to imbalanced data than just small data size. The algorithm strives to minimize the overall prediction error and hence is apt to concentrate more on the prediction accuracy of the majority class. Consequently, this can result in poor accuracy for the classes with less data still leading to good overall prediction accuracy. (Chen et al. 2004, 2.)

There are several methods that can be applied to overcome the problem of imbalanced datasets. Some common approaches include using different sampling techniques, Synthetic Minority Over-sampling Technique (SMOTE) or an approach based on cost sensitive learning (Chen et al. 2004, 1-2). In this method a penalty, a high cost, is assigned to misclassification of the minority class and the objective is shifted from the best prediction accuracy to the lower prediction cost instead (Chen et al. 2004, 1; Rocca 2019). The purpose in applying different sampling methods is to balance out data distribution by modifying the imbalanced data set. In oversampling selected samples from the minority class are replicated and added to the dataset to increase the total number of samples in the minority class. In under sampling, inversely, a set of samples is removed from the majority class, which reduces its dominance. However, removing samples might also prevent the classifier getting all the relevant information pertaining to the majority class. The risk in oversampling is model overfitting that leads to a lower accuracy on the unseen test data, despite the high training accuracy. (He & Garcia 2009, 1266-1267.)

Another way of doing oversampling is using the SMOTE algorithm that, instead of creating copies, creates synthetic, artificial, minority class samples based on the feature space similarities between existing samples. SMOTE improves learning by providing more related minority class samples and more information to learn from (Chawla et al. 2002, 328, 352; He & Garcia 2009, 1267) and, unlike with just replicating samples, synthetic samples make the classifier build larger and less specific decision regions. Thus, the overfitting problem is avoided and the model generalizes better. Although, as a downside the model might be over generalizing, i.e. underfitting. (Kotsiantis et al. 2006, 3; He & Garcia 2009, 1268.)

4.3 Applying machine learning methods to data classification

The actual programming and model development was executed by a research assistant since the researcher herself does not have required skills in programming. An outline of the model development and running the model is covered in this chapter.

There are more than two classes in the research data; therefore a multiclass classifier was developed and two different algorithms were applied. Automatic classification was executed in two stages: first, Classifier A performed classification by estimating the probability of a sample belonging to each of the main classes depicted in Table 2 and assigned the class with the highest probability to each sample. After that, all the samples in the brand-loyalty class were classified again by Classifier B to the subclasses presented in Table 2. This two-step approach was used to avoid any confusion between the classes that are hierarchical by nature.

As discussed in the chapter 3, before applying the classifiers, the data cleaning and feature extraction were executed. Two types of features were used as input. Feature set A including text data: search query, campaign name and ad group name, and feature set B included campaign metrics such as click through rate, cost per click and conversion rate. The data was cleaned from missing values and the continuous variables standardized. For the text data processing a list of Finnish common stopwords was added and TF_IDF performed on search queries. For the model training, the labeled training data was divided into two parts, where the larger part was for training and smaller for testing the model performance. In this case 80% of the training data was used for training the model and 20% for testing. Purchase funnel stages were the target or dependent variables.

Two machine learning algorithms, Random Forest and Extreme Gradient Boosting (XGBoost) were applied in this study. It is recommended to test and compare several algorithms to see which one performs best, since every learning task is unique and every algorithm has its advantages and disadvantages (Raschka 2015, 49). Both algorithms originate from decision trees and they are effective and widely applied machine learning methods in classification tasks (Breiman, 2001; Chen & Guestrin, 2016; Chen et al. 2004). Random forests have been proven to be paramount in marketing context compared to other more traditional classification methods (Coussement & Van den Poel 2008, 314). Also, boosting algorithms provide excellent results, especially in text classification (Coussement & Van den Poel 2007, 874). Random Forest usually requires a large amount of training data and has found to be sensitive to imbalanced data distribution (Taniguchi et al. 2018, 10-11) while using XGBoost this can be easily managed by setting one parameter to indicate dataset ratio (Lahiri 2018). Both algorithms, however, are robust to outliers and noise in data, in case of misclassified samples, for example (Breiman 2001; Gómez-Ríos et al. 2017).

Grid Parameter search was applied to improve model performance and to improve model validity a 10-fold cross validation was performed. There are two types of parameters in machine learning: those that are learned from the training data and those of the algorithm that are separately optimized (Raschka 2015, 185). Random Forest is easy to implement, as only two parameters need to be set. They are the total number of trees and the number of predictive features that are randomly chosen at each node (Coussement & Van den Poel 2008, 316). While the number of trees is one of the main parameters of XGBoost, as well, it requires more parameter tuning than Random Forest.

Together with learning rate, the other main parameter, number of trees affects the model complexity and generalization ability. The more trees, the more accurate the model becomes, which can lead to overfitting. This can be avoided by a lower learning rate value, as smaller learning rate values tend to improve generalization performance. Hence, an optimal number of trees can be discovered by evaluating the prediction accuracy or through cross-validation process. In addition to several other parameters, it is possible to regulate the parameters of individual trees with XGBoost. (Nielsen 2016, 44, 73-74.)

Parameter optimization is pivotal since the parameter values are of great significance to the entire model performance (Coussement & Van den Poel 2008, 313). K-fold cross-validation together with grid search are commonly used to find the optimal set of parameters for the model (Raschka 2015, 187). Different cross-validation methods are also used for model selection and estimating the model generalization performance (Rasmussen & Williams 2006, 111, 108). The grid search procedure includes trying different parameter combinations on the training set and identifying the value set that provides the best level of accuracy by using cross-validation. In a k-fold cross-validation the training data is divided into several smaller sections, i.e. folds. K refers to the number of folds, which in this case was 10. Each of the folds is used for testing at some point so that in the first iteration, the first fold is used to test the model while the remaining folds are used to train the model, and this is repeated until each fold has served as a test set once. Next, this procedure is repeated with different parameter values to identify the set that provides the best cross-validated performance. Then the selected values are applied to the actual model. The cross-validation method presents a more reliable view on the model performance while the actual test or validation set separated in the beginning can be left untouched until the final model evaluation. (Coussement & Van den Poel 2008, 314, 317-318.)

5 RESULTS OF AUTOMATIC DATA CLASSIFICATION

After the trained classifiers were applied to the labeled test data set, the performance of the machine learning algorithms were evaluated. This is also known as the model validation that exhibits the confidence of the model (Okazaki et al. 2014, 475). Evaluating the performance on unseen data is fundamental to verify that the classifier is able to generalize well (Coussement & Van den Poel 2008, 317), i.e. to predict accurately the remaining unlabeled data, as well.

A confusion matrix is typically used for evaluating the performance of machine learning algorithms. It shows the number of samples that the algorithm predicted both correctly to each class and incorrectly to each class. In a basic two-class classification problem these are referred to as True Positives, False Positives, True Negatives and False Negatives. Generally used performance measures predictive accuracy and error rate can be calculated from the matrix. (Chawla et al. 2002, 322-323.) However, these metrics are highly sensitive to imbalances in data and therefore are not good for skewed datasets.

Instead, the performance of algorithms can be evaluated more comprehensively with metrics called precision, recall and F-score (He & Garcia 2009, 1276-1277; Chawla et al. 2002, 326), which are also derived from the confusion matrix (Chen et al. 2004, 5). Precision indicates exactness of the prediction (He & Garcia 2009, 1277). It is the percentage of samples the model identified to a certain class that actually belonged to that class (Raschka 2015, 191-193). Recall measures completeness instead (He & Garcia 2009, 1277). It is the percentage of the class relevant samples in the test data that the model identified correctly out of the total number of samples belonging to that class. The maximum value for these key indicators is 1 meaning that the classifier performs perfectly while values close to 0 indicate that the classifier is highly imprecise. (Raschka 2015, 191-193.) F1-score is the harmonic mean of precision and recall and by combining these two metrics it provides insight into the effectiveness of the classifier (He & Garcia 2009, 1277; Rocca 2019).

The results of applying machine learning algorithms to the classification task show that both classifiers perform well in classifying the search query terms. The values of precision, recall and F1-score are presented in Table 4 below.

Table 4 The values of precision, recall and F1-score

Class	Precision		Recall		F1-Score	
Algorithm Used	Random Forest	XGBoost	Random Forest	XGBoost	Random Forest	XGBoost
Awareness	**0.95**	0.94	0.97	**0.98**	0.96	0.96
Research	0.98	**0.99**	**1.00**	0.99	0.99	0.99
Purchase	**0.88**	0.87	**0.94**	0.92	**0.91**	0.89
Brand Loyalty	0.98	0.98	0.91	**0.94**	0.94	**0.96**
RBL	0.96	0.96	0,98	0,98	0,97	0,97
CBL	0.94	**0,97**	0,94	**0,97**	0,94	**0,97**
SBL	0.99	**1,00**	0,99	0,99	0,99	0,99
avg / Total	0.96	0.96	0.96	0.96	0.96	0.96

It can be seen from the Table 4 that both classifiers achieve a high performance for the classes, as the values of precision and recall, as well as F1-score are close to 1 throughout the classes. This indicates that the classes in the training data were distinct and discriminating while the predictive features provided a good signal for the algorithms to separate between the classes.

The imbalance can also happen within classes, which refers to overlapping data among the classes, and that hinders classifier performance, too. Both classification errors and overlapping often occur close to class boundaries, thus supposedly the problem of overlapping classes and class imbalances or the low number of samples are related. (Kotsiantis et al. 2006, 9; Rasmussen & Williams 2006, 36.) Recall indicates the detectability of the classes and Table 4 shows high values of recall for all the classes, which suggests that there is no significant overlapping issue. Furthermore, Random Forest was able to detect research class perfectly. While recall is not sensitive to imbalanced data distribution, precision is (He & Garcia 2009, 1277). In several cases recall is slightly better than precision and a bigger difference between these values can be spotted in purchase class that is the minority class.

However, the data distribution between classes in this research dataset is not highly imbalanced and the number of samples in the minority class is not very low, not even in the subclasses of brand-loyalty. As Kotsiantis et al. (2006, 10) observed in their study, the effect of imbalances was reduced when the minority class contained more samples and hence was better represented and detectable. In case of unsatisfactory precision and recall values, the amount of labeled data could have been added as well, which of course would have required more time and human resources to execute the labeling.

There was no significant difference in the performance of the two different classifiers, Random Forest and XGBoost. The latter seems to have been slightly better as a

whole reaching the better value 8 times out of 21, while equal value between the classifiers was reached 8 times. Random Forest was able to predict purchase class slightly better than XGBoost. XGBoost, in turn, performed somewhat better in predicting brand-loyalty and the subclasses of brand-loyalty, competitor brand loyalty in particular, while Random Forest did not excel in any of the subclasses. This suggests that the boosting mechanism of XGBoost algorithm might work particularly well with classes that have small sample sets. Purchase class was the minority class for the main classes but altogether the subclasses retail brand-loyalty and competitor brand-loyalty contained the lowest number of samples, approximately 200 each. Also, the data was more imbalanced between the subclasses than between the main classes. The most significant difference in performance between the two classifiers could be detected in competitor brand-loyalty class where the training data was more limited and additionally more incoherent.

Overall, the research class and the subclass of supplier brand loyalty were the most accurately predicted classes, as their values were closest to 1. Also, from the main classes, research had the smallest difference between the values of precision and recall. Admittedly, research class had the largest amount of labeled data and among the subclasses supplier brand loyalty was the majority class. All of the brand-loyalty subclasses, too, had very little difference, if at all, between the two values. Interestingly, despite its large number of samples, loyalty main class did not perform as well overall and the difference between precision and recall was relatively high. The high precision value with the lower recall value suggests that the algorithms were able to make correct predictions, but they were not able to identify all the samples to this class. This means that the algorithms had some problems in detecting the loyalty class.

Purchase class attained the poorest performance values and the difference between precision and recall was rather big also in this class. This was somewhat expected since purchase was the minority class and also the manual classification for the training data was found to be the most challenging. Without any context or explicit commercial intent indicators, it was difficult to say whether the searcher is ready to purchase or just comparison-shopping. However, as Dai et al. (2006, 832) noted, only a small portion of search queries contain these indicators, which was the case also in this study. In addition, the decision to have brand loyalty class forced many queries containing e.g. the word “buy” to brand loyalty class because these terms often included a brand name as well. Contrarily to brand-loyalty, purchase reached better values for recall than for precision, which means that the algorithms classified more query terms to this class than belonged there.

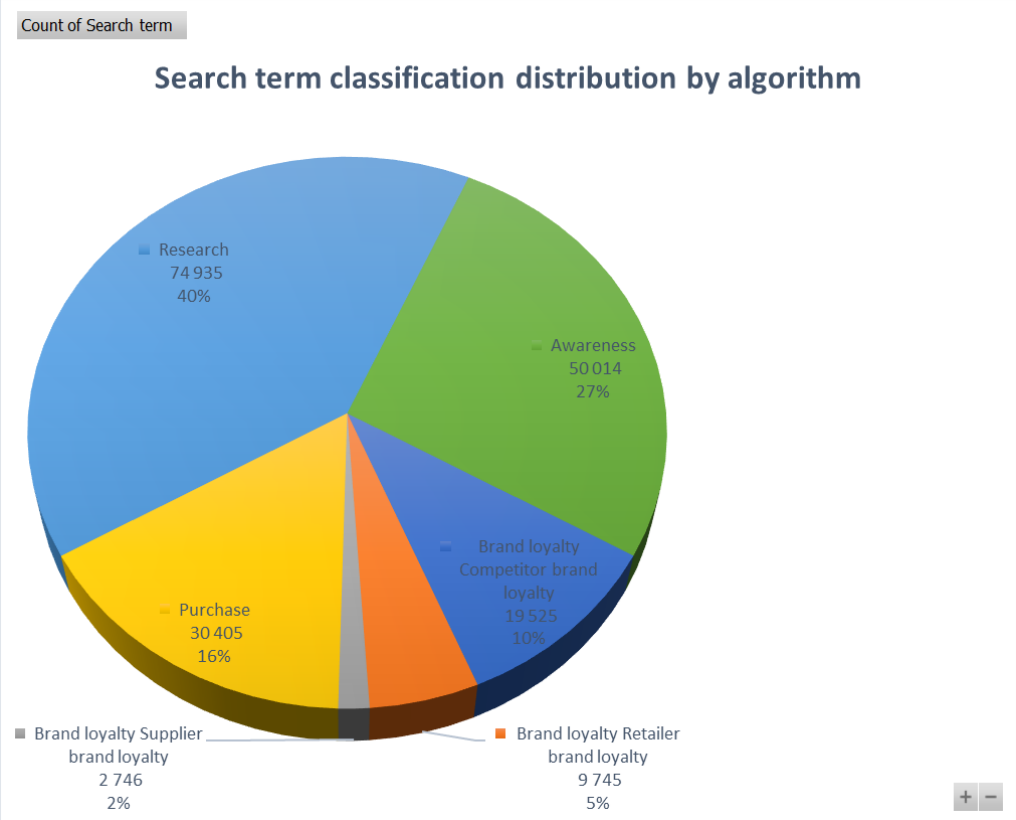
There were some confusion between purchase and research classes and some awareness terms were classified to research or purchase classes. The loyalty subclasses were mainly predicted very accurately. Classification mistakes occurred for example for

search terms that included the word “*ärrä*” which is colloquial term for the retail locale officially called R-kioski. Because this was not included in the training data, the algorithm did not know that R-kioski, or “ärrä”, is a competitor to the focal company in this research. The classification was inconsistent so that queries "*ärrä elämyslahjakortit*" (ärrä experience gift cards) were classified correctly as competitor loyalty while "*ärrä elämyslahja*" (ärrä experience gift) ja "*ärrä elämyslahjat*" (ärrä experience gifts) were classified to retail loyalty. This was probably because the second word refers to the focal company. Also, terms “*ärrä lahjakortit*” (ärrä gift cards) were classified to awareness. It seems that the algorithm ignored the word “ärrä” and only considered the second word. All in all, classifying queries to the loyalty subclasses requires deeper knowledge of the companies that operate on the market, namely which companies are suppliers and which are competitors.

As discussed earlier, the focal brand name itself might cause some confusion, i.e. when it refers to the company and when to experience gifts in general. Also, as suspected, the not purchase related terms caused some errors like ”*Elämyslahjakortti ei tullut perille*” (Experience gift card did not arrive) and ”*elämyslahjakortti ehdot*” (experience gift card terms of use). These were classified to purchase although they are clearly informational in nature and therefore not really related to the purchase process. When considering the purchase process, they would rather belong to the post purchase stage for they clearly imply that the purchase is already made.

Finally, in this study and in this context there is no need to stress the importance of either precision or recall but rather finding a good overall performance. This can be evaluated by examining the F1-score, and as stated earlier, the performance of XGBoost was slightly better overall, albeit no significant differences were discovered between XGBoost and Random Forest. To conclude, the performance metrics show good values, and according to these results a large-scale automatic classification of search terms into purchase funnel stages was successful with the chosen classifiers.

The distribution of all search queries to the different funnel stages can be seen from the Figure 3 below.



Class	Secondary class	Count of Search term	Of Search Terms (%)
Awareness		50 014	26.69%
Research		74 935	39.99%
Purchase		30 405	16.23%
Brand loyalty	Competitor brand loyalty	19 525	10.42%
	Retailer brand loyalty	9 745	5.20%
	Supplier brand loyalty	2 746	1.47%
Brand loyalty Total		**32 016**	**17.09%**
Grand Total		**187 370**	**100.00%**

Figure 3 Search term classification distribution by algorithm

As can be seen from the Figure 3, the search query classification performed by machine learning algorithms does not quite follow similar distribution as the manually labeled training data. Research is still the dominant class but awareness comes second and brand loyalty is only slightly larger than purchase stage. From the brand loyalty sub-classes the competitor class is remarkably larger than retailer and supplier brand loyalty classes. What also can be noted is that the distribution does not follow the traditional funnel shape. These findings correspond to the findings of previous similar studies. The class distribution between purchase funnel stages in the research of Jansen and Schuster (2011) looks roughly the same; research is significantly larger than other classes, the

next largest is awareness and last decision followed by purchase, though in this study these two classes are combined. Also according to the study of Hotchkiss (2004) the largest category is research and it is percentage-wise even larger than in this or in Jansen and Schuster's (2011) research. However, awareness is the smallest category in the study of Hotchkiss (2004) unlike in this and in Jansen and Schuster's (2011) research. The results imply that perhaps consumers in fact first increase their consideration set in the research or evaluation stage as for example Court et al. (2009) suggest. Although, it might not even be possible to capture the consumer's entire purchase journey through search queries because of the multiple touch points and the complexity of journeys.

When considering the search behavior online, it is discovered in many previous studies that broad or generic search queries are more likely used in the upper purchase funnel and narrowed down to more specific in the lower purchase funnel, closer to the purchase (Im et al. 2019; Li et al. 2016; Bronnenberg et al. 2016; Rutz & Bucklin 2011; Hotchkiss 2004) and specificity is found to be linked to query length so that shorter queries are more general and longer more specific (Hafernik & Jansen 2013; Phan et al. 2007). The query lengths of this research data are presented in Table 5 below.

Table 5 Length statistics for the search queries

	Average	Median	STDEV	% of searches 4 words or less	Count
Awareness	2.028	2	0.647	99.72%	50 014
Brand loyalty	2.520	2	1.034	96.21%	32 016
Purchase	2.608	2	0.837	97.50%	30 405
Research	3.129	3	1.062	91.07%	74 935
total	**2.647**	**2**	**1.028**	**95.30%**	**187 370**

According to the classification results of this study the search queries are on average longer in research stage than in awareness. However, after that the length starts to decrease. The average query length in awareness is 2 words, in research 3.1, in purchase 2.6 and in brand loyalty 2.5. The data is affected by some abnormally long search queries, up to 14 words that most likely are not made by actual user. That is why median was counted as well but it does not offer any more insight. The median number of words in a query is 2 in all other classes except in research it is 3. Additionally, the table includes standard deviation and the percentage of queries whose length is 4 words or less. This additional information supports the information provided by the average length. It can also be noted that not many queries exceed the length of 4 words. It has to

be considered, though, that the number of characters was not counted. Finnish language contains compound words that might be very long and therefore quite specific by nature. They would still be counted as a one word. It could be further explored if there are any other differences between the awareness and purchase queries.

6 CONCLUSIONS AND DISCUSSION

6.1 Theoretical and managerial implications

Understanding customer journeys is considered essential for marketers while the journeys have become more complex consisting of multiple touch points and mixing offline and online channels (Haven 2007; Wolny & Charoensuksai 2014; Rangaswamy et al. 2009). This helps marketers to target marketing activities more accurately in order to guide consumers through the purchase funnel, measure marketing effectiveness and respond more effectively to the consumer information needs in different stages of their buying journey (Wolny & Charoensuksai 2014; Im et al 2019; Dai et al. 2006). As search engines have a central role in consumer buying journeys (Su et al. 2018, 547), search engine marketing practitioners have raised the importance of targeting advertising according to the consumer purchase funnel stages by mapping keywords to these stages (Petrik 2014; Hadrien 2015a; Swan 2018; Qian 2017). However, because of the vast amount of data available, manual classification is not really a feasible solution. This study presents a supervised machine learning approach to automatically classify consumer search queries into customer journey stages.

The main research question was how the searches of search engine users can be automatically classified to different stages of the purchase funnel based on search query terms using machine learning. Two different machine learning algorithms were applied and their results compared. Also, the key factors in the classification of text-based training data for supervised machine learning were considered. The research data consisted of 190 245 search queries gathered from the Google AdWords account of a Finnish e-commerce company called Elämyslahjat over a period of five years. The main steps in the study included collecting the search terms, defining classification categories and criteria, human coding, data pre-processing, classifier development and training and evaluation of the machine learning performance.

Since creating a representational training data set for supervised machine learning tasks is with utmost importance (Schönleber 2018; Raschka 2015, 50, 99), several factors that affect the quality and validity of the training data and therefore to the success of automatic classification were observed from the theory as well as through execution. The first consideration is accurate and comprehensive classification criteria. Many researchers say that the manual coding is one of the most challenging and time consuming tasks in machine learning (Im et al. 2019; Vázquez et al. 2014; Lewandowski et al. 2012). This was evident in this study as well. The manual coding begun by going through the data to see what kind of search terms it contained altogether. Establishing comprehensive classification criteria provides the foundation for the classification task.

However, as the search query data is unique for every business, no earlier criteria could be followed exactly as they were, but they needed to be adjusted and modified as the classification progressed to be suitable for this purpose and for this business. Naturally these criteria could be then used later for the same company, however considering that search queries and consumer search behaviors keeps changing and evolving.

This leads to the second consideration, which are the qualities associated with the different classes or categories in the training data. Large imbalances can harm the machine learning classifier performance; hence the classes should contain approximately the same amount of data. The classes should also be representational of the whole data and include comprehensive set of samples because the supervised machine learning classifier learns through these samples. Lastly, the classes have to be well separable to avoid overlapping and misclassifications (Raschka 2015, 4, 97). In this study, for example the predefined classification criteria needed to be modified to fulfill these requirements. It was also noted, that having too many classes in this type of hierarchical classification caused overlapping and confusion, although using more classes might have offered more insight.

The previously mentioned factors concern all kind of supervised machine learning training data but in addition to these text-based data has its own special linguistic characteristics that need to be considered in the classification. Firstly, it is difficult to interpret consumer intent and search purpose solely based on short search queries without any further context and hence there is great risk for human error and bias. Other linguistic challenges included ambiguous, multifaceted and irrelevant queries. The data included several search terms that did not relate to a gift purchase or even the company but because of the large amount it was somewhat impossible to find and exclude them all. With ambiguous and multifaceted terms, it is important to acknowledge that all the similar terms should be assigned the same class even if in some cases the term seems to have a different meaning from a human perspective. As an example, in this study the search queries including the company name Elämyslahjat or different variants were difficult to classify logically since the word itself can have other meanings than just the brand name. Also, spelling mistakes such as compound words written separately or vice versa, which could affect the meaning of the word or search term.

Although the distribution of search queries did not quite follow the traditional purchase funnel shape, the machine learning algorithms were able to find distinct classes and classify search queries accurately. Two different algorithms were applied, Random Forest and XGBoost. The performance of XGBoost was slightly better overall, especially in the minority classes, albeit no significant differences were discovered between the two classifiers. To conclude, the performance metrics show good values, and according to these results a large-scale automatic classification of search terms into purchase funnel stages was successful with the chosen classifiers. This method and these categories

related to consumer purchase journey stages can be utilized in targeting search engine marketing by defining the campaign structure based on purchase journey stage. Afterwards by following the campaign metrics marketers can find the most profitable groups, identify good campaign keywords and set up optimal bids for the keywords. The categories can also be leveraged in search engine optimization or content marketing to improve the possibilities to reach the consumer at the right moment with right content in relation to his/her stage in the purchase funnel.

Automated solutions are a popular approach in the field of marketing in the digital era because of the vast amount of data available and automation offers scalability. As machine learning becomes more and more usable and available for marketing purposes it is important for practitioners to understand its possibilities and be aware of its limitations. This study suggests one possible application how to employ machine learning in marketing. While one of the cornerstones in supervised machine learning applications is the creation of training data, the study aimed to examine aspects to be considered in the creation of training data. All in all, machine learning cannot provide valid and valuable answers without good quality training data.

6.2 Limitations, validity, reliability and future studies

There are limitations that affect the validity and reliability of this study and the generalizability of the results. First, although the research data is large in size and from a long time span, it originates from one single company in e-commerce selling only one type of products. The language of the search query data is Finnish. In addition, the research data consists solely of data acquired from the company's paid search advertisement campaigns. Incorporating the company's web site's organic traffic search queries to the research data would possibly lead to different kind of outcomes, especially when consumers seem to avoid clicking paid search engine results (Jansen & Resnick 2006, 1959). The entire data set originates from Google's search engine, but it is the market leader, though, and clearly the most used search engine in Finland.

The data does not include any searches that did not result in a click leading to the company website, nor is it acknowledged if part of the purchase journey was completed offline. Hence it might not even be possible to capture the consumer's entire purchase journey through search queries because of the multiple touch points and the complexity of journeys. It was decided to use the search queries typed in by the search engine users instead of the ad campaign keywords because they were thought to give more accurate information on consumer intent and search purpose. Keyword lists are much more limited and created by the marketers.

In addition to the size and quality of the research data, in this study the size and quality of the machine learning training data created from the entire research data is just as central to the quality of the study. The machine learning algorithms require enough data to learn from and probably the greatest constrain there is that the manual coding is very time consuming. Also the amount of data should be evenly distributed among the different classes. In this study the classes were somewhat imbalanced, although the total number of data samples was sufficient. As discussed in chapter 4.2.4, there are several methods that can be applied to overcome the problem of imbalanced datasets and this can also be influenced by the choice of the algorithm, since some algorithms are more robust to imbalances and outliers in data. For example the other algorithm used in this study, XGBoost, is such and it could be observed from the results that it performed slightly better. The irrelevant terms in the data may therefore affect the algorithm performance as well, but as already mentioned before, due to the large amount of data it was somewhat impossible to clean the data completely from all the outliers. However, the both chosen algorithms are robust to outliers and errors, as discussed earlier in chapter 4.2.3.

A major single factor affecting the reliability and validity of this study is the manual coding of the machine learning training data, which can be thought as parallel to content analysis. There are comparable measures to evaluate algorithm performance but the problem with evaluation of text classification methods is the lack of standardized data collection procedures (Kowsari et al. 2019, 4). Lewandowski et al. (2012) questions the reliability of manual coding and inferring the consumer intent from search queries, and hence the entire approach of automatic query classification because the model performance is largely based on the quality of the underlying training data. In research in general, the used techniques are expected to be reliable and they should lead to results that are replicable so that other researchers using the same techniques at different point in time should end up with the same results. For content analysis this implies that the context and reasoning behind inferences must be explicated to others. (Krippendorff 2004, 18, 24.)

While basically manually annotating the search query classes includes a lot of subjectivity, the reliability of classification can be tested by measuring inter-rater agreement. It tells the level of agreement between different annotators or raters and when agreement is high the data is considered reliable (Salminen et al. 2018b, 80). However, Salminen et al. (2018b) found in their study that the agreement ratings can be low in spite of accurate instructions suggesting that the inter-rater agreement might not be the optimal indicator for data quality in social computing tasks. In this study, the inter-rater agreement was not measured but the manual classification was conducted in a cooperative manner so that the unclear cases were discussed with an expert with experience both in search engine advertising in this particular company and the gift industry. Also,

the framework and the context for the manual classification were clearly explicated, as well as the logic behind inferences. The predefined classification criteria and the framework were modified in the course of manual classification when needed for better reliability.

There was a risk of error especially related to certain terms such as irrelevant or ambiguous queries because the researcher herself had problems deciding the suitable class or deciding which rule or logic to follow. As an example, according to the classification criteria terms including the word "price" belong to the purchase class but often these terms also included a brand name, which implies belonging to the brand loyalty class. It had to be evaluated which classification would make more sense considering the entire data. The irrelevant terms were not manually annotated at all, which may have affected the automatic classification performance depending on their relative numbers. However, the chosen classifiers should allow some errors in the classification of the training samples (Mitchell 1997, 54). Perhaps it would have been worthwhile to add one class for the terms that are irrelevant in relation to the purchase funnel of the research company, for example query such as "Anne Berner husband". Another aspect relates to the assumption that the purchase funnel is a hierarchical process. As Barry & Howard (1990, 108) stated, it is difficult to define when one stage ends and another begins, or does not begin. Here it was also assumed that all users are searching to make a purchase.

Altogether, defining the class for a single search query in isolation is challenging as noted by other researchers (Im et al. 2019). By examining isolated search queries it cannot be defined accurately, for example whether the consumer is ready to purchase or still examining different options. It is only known for sure that based on prior research certain types of queries are more typical in the end of the funnel. One possibility could be to follow the entire user search session as Im et al. (2019) or a user survey in addition to ask users what their search intent was as Su et al. (2018) did in their study.

After all, while manually executed content analysis often suffers from low replicability, more objective results can be obtained from machine coding and the classification can be repeated consistently (Okazaki et al. 2014, 469). Therefore automatic classification is more reliable and consistent than conducted by human. However, human intelligence and reasoning skills are needed in the beginning to create the foundation for the automatic classification. If the foundation is poorly made, so are the results provided by machine learning applications questionable. The machine learning method of this study is generalizable to other similar classification tasks but the training data can only be used for the purposes of the focal company. Also the classification criteria are specific to this company and its product type, although the criteria can be leveraged at some level by other businesses.

Although machine learning classifier performance can be measured with several different methods, certain level of uncertainty and limitation are associated with them as

well. The metrics of precision, recall and F-score, which are used in this study, are widely adopted for multiclass classification tasks (Salminen et al. 2018a; Habernal et al. 2014). There are always some trade-offs to consider when evaluating classifier performance and it depends on the task which metric is the most relevant. In some cases it is more important that the classifier is extremely accurate (high precision) and sometimes it is acceptable to have some misclassifications as long as the class is not missing any samples (high recall). This could be for example the case in cancer screening. The uncertainty of these metrics derives from the matter already discussed previously, the quality of training data. The test set for measuring the algorithm performance is separated from the manually classified training data. Therefore the performance metrics might only give an idea of the classification accuracy of the whole data.

As the search queries are classified, it can be further investigated which one of the classes is most valuable or profitable. Also the advertising costs of the different groups can be examined and compared through click costs. As an implication for search engine marketing research this approach offers a way to test which type of ad targeting is more efficient – the traditionally used campaign structure based on product categories or campaign structure constructed according to the consumer purchase journey. It could also be explored how the top of funnel performance effects bottom of funnel performance.

7 SUMMARY

Understanding customer journeys is essential for marketers in order to guide consumers through the purchase funnel and respond more effectively to the consumers' information needs in different stages of their buying journeys. While targeting advertising according to the consumer purchase journey stages is considered important, manual efforts to classify all the keywords or search terms are not feasible. One search advertising account can include tens of thousands of keywords and consumer search queries are continuously evolving. The purpose of the study was to present a supervised machine learning approach to automatically classify consumer search queries into customer journey stages in digital environment.

The main research question was how the searches of search engine users can be automatically classified to different stages of digital purchase funnel based on search query terms using machine learning. Since one of the cornerstones in supervised machine-learning applications is the creation of training data, the study aimed to discover how to perform the initial manual data labeling by examining factors that affect the quality and validity of the training data and therefore the success of automatic classification. Hence, the first sub question was "What are the key considerations for the classification of text-based supervised machine learning training data?" The second sub question refers to the performance of different machine learning algorithms applied to the task and is "What are the key differences in different machine learning technique results for collected sample data?"

There are several different models describing customer purchase journeys but they all refer to the purchase process or path that consumers go through when making a purchase decision consisting of different stages. Generally these models include three fundamental stages: pre purchase, purchase, and post purchase. Pre purchase contains all the phases that consumer goes through before the purchase such as need recognition, information or alternative search and consideration. Purchase stage comprises the purchase decision and the purchase event itself. Post purchase stage covers the entire customer experience and interactions with the brand such as using the product, interaction with customer service, brand loyalty and customer engagement like word of mouth.

Internet and social media have had a significant effect on consumer decision process and there is a lot of discussion on whether the traditional purchase funnel is valid in digital environment. Others think the funnel should be abandoned entirely, while others think the web has just transformed the funnel. Altogether journeys have become more complex consisting of multiple touch points and mixing offline and online channels. A framework consisting of five purchase process stages, including the post-purchase stage of loyalty, was chosen for this study. However, during the manual coding for machine learning classification it was noticed that with that many classes the distinctness of the

classes was inferior and the classes needed to be reduced. As The final purchase funnel class structure used in labeling the training data for machine learning included the stages of awareness, research, purchase and brand loyalty. Also, other elements related to the query or user intent, such as query length and specificity, were utilized in the query classification to the purchase funnel stages. Usually the user queries are broader and shorter in the upper funnel stages and become longer and narrow down to more specific towards the purchase.

Automated solutions and machine learning applications are now a popular approach in the field of marketing, too, because of the vast amount of data available for decision-making. Machine learning has been around for ages but now it is becoming more and more applicable because of the technical advancements in the data processing capacity of computers. Machine learning is based on artificial intelligence and it can be described as designing systems that can predict outcomes by learning from input data, and it enables to spot patterns in large amount of data and make predictions.

The main types of machine learning are supervised and unsupervised learning. The difference between supervised and unsupervised learning is that in the former the results, for instance the class labels, are known beforehand while in the latter the result is unknown. Classification is an example of supervised learning where the resulting categories are determined in beforehand and the purpose is to learn from labeled training data, which consists of a set of samples. By learning from the training data machine can then make predictions about previously unseen data.

The training data should be a representative sample of the whole data and include features that are informative and discriminating enough for the learning algorithm will use these features in learning the data labels. Also, the data sample distribution between the classes should be balanced. There are several machine learning algorithms to choose from and each has its benefits and disadvantages. Different models perform differently on different data with different characteristics and usually several algorithms are tested to find the best performing one to each particular task or problem. Altogether, the quality of data available for the algorithm is more important than which algorithm is used.

The machine learning process includes the main stages of pre-processing, learning and evaluation. Pre-processing consists of data extraction and cleaning and labeling the training data set. The data pre-processing along with the training data creation are crucial and most time consuming steps in the whole process. In the learning stage, the algorithm is applied to the training data and optimized. After this the final model is applied to the test data and results evaluated. When the results seem good the final machine-learning model can be applied to the unseen data.

The main steps of the empirical part of the study included data collection, determining the classification criteria and categories, human coding, programming and training of automated classification algorithms and evaluation of the classification results. Clas-

sification was conducted following criteria that was based on prior literature on customer purchase funnel and query classification while considering the machine learning requirements. This included constructing the classes so that they were distinct i.e. well separable. The generally expected idea was followed that the search query specificity reflects the user intent specificity and the more specific the user intent the closer to the purchase the user is. The manual coding involved using inductive content analysis whereas machine learning can be considered as automated content analysis.

The research data consisted of 190 245 search queries gathered from the Google AdWords account of a Finnish e-commerce company called Elämyslahjat over a period of five years. They were classified to stages of awareness, research, purchase and brand loyalty, which included three subclasses: retail brand loyalty, supplier brand loyalty and competitor brand loyalty. For the automatic classification a multiclass classifier was developed and two different algorithms, Random Forest and XGBoost were applied. Automatic classification was executed in two stages: first, Classifier A performed classification by estimating the probability of a sample belonging to each of the main classes and assigned the class with the highest probability to each sample. After that, all the samples in the brand-loyalty class were classified again by Classifier B to the subclasses. This two-step approach was used to avoid any confusion between the classes that are hierarchical by nature.

The metrics of precision, recall and F-score were used to evaluate the performance of the algorithms and according to them, the algorithms were able to find distinct classes and classify search queries accurately. The performance of XGBoost was slightly better overall, especially in the minority classes, albeit no significant differences were discovered between the two classifiers. The performance metrics show good values, and according to these results a large-scale automatic classification of search terms into purchase funnel stages was successful with the chosen classifiers.

The distribution of search queries did not quite follow the traditional purchase funnel shape, since the research class had the largest number of queries. Although, it might not even be possible to capture the consumer's entire purchase journey through search queries because of the multiple touch points and the complexity of journeys. The reliability of automatic classification depends to a great extent on the quality of training data and there are several factors affecting the quality. Manually annotating the search query classes includes a lot of subjectivity and defining the class for a single search query in isolation is challenging. Also, the class boundaries are not strict. In addition, classifying text data includes special linguistic characteristics that might complicate inferring the right class. These include, for example irrelevant or ambiguous queries.

To conclude, this method and these categories related to consumer purchase journey stages can be utilized in targeting search engine marketing by defining the campaign structure based on purchase journey stage. Afterwards by following the campaign met-

rics marketers can find the most profitable groups, identify good campaign keywords and set up optimal bids for the keywords. The research has implications for businesses that are looking for new, more efficient ways to process web metrics and to utilize and analyze online user data. The end goal is to gain better understanding of customer behavior in different stages of the online purchase process. This is to target both marketing and resources more effectively and segment audiences according to the purchase process stages, not only for search engine advertising but also for search engine optimization and content marketing purposes.

REFERENCES

Agarwal, Ashish – Hosanagar, Kartik – Smith, Michael D. (2011) Location, Location, Location: An Analysis of Profitability of Position in Online Advertising Markets. *Journal of Marketing Research*, Vol. 48 (6), 1057-1073.

Ashkan A. – Clarke C.L.A. (2009) Characterizing Commercial Intent. *CIKM'09*, Hong Kong, China, November 2–6, 2009, 67–76.

Barry, To E. and Howard, D. J. (1990) A review and critique of the hierarchy of effects in advertising. *Internatioal Journal of Advertising*, Vol. 9, 98–111.

Beleites, C. – Neugebauer, U. – Bocklitz, T. – Krafft, C. – Popp, J. (2013) Sample size planning for classification models. *Analytica Chimica Acta*, Vol. 760 (Special Issue: Chemometrics in Analytical Chemistry 2012), 25–33.

Bell, Jason (2015) *Machine Learning: Hands-On for Developers and Technical Professionals*. John Wiley & Sons, Inc., Indianapolis, Indiana.

Ben Gomes (2017) Google blog: "Our latest quality improvements for Search". <https://www.blog.google/products/search/our-latest-quality-improvements-search/>, retrieved 3.4.2019.

Bishop, Christopher M. (2006) *Pattern Recognition and Machine Learning*. Springer Science+Business Media, LLC., New York, USA

Breiman, L. (2001) Random Forests, *Machine Learning*, Vol. 45 (1), 5–32.

Broder, Andrei (2002) A taxonomy of Web search. *ACM SIGIR Forum*, Vol. 36 (2), 3–10.

Bronnenberg, Bart J. – Kim, Jun B. – Mela, Carl F. (2016) Zooming In on Choice: How Do Consumers Search for Cameras Online? *Marketing Science*, Vol. 35 (5), 693-829.

Brownlee, Jason (2017) How Much Training Data is Required for Machine Learning? <https://machinelearningmastery.com/much-training-data-required-machine-learning/>, retrieved 18.4.2019.

Caigny, Arno, de – Coussement, Kristof – Bock, Koen W., de (2018) A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, Vol. 269, 760–772.

Canhoto, A. I. – Padmanabhan, Y. (2014) "We (Don't) Know How You Feel" – A Comparative Study of Automated vs. Manual Analysis of Social Media Conversations. *Conference paper*, July 2014, 1–9.

Chawla, Nitesh V. – Bowyer, Kevin W – Hall, Lawrence O. – Kegelmeyer, W. Philip (2002) SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, Vol. 16, 321–357.

Chen, Chao – Liaw, Andy – Breiman, Leo (2004) *Using Random Forest to Learn Imbalanced Data*. Department of Statistics, UC Berkeley, 1–12.

Chen, Tianqi – Guestrin, Carlos (2016) Xgboost: A scalable tree boosting system. *Proceedings of KDD '16*, San Francisco, CA, USA, August 13-17, 2016, 1–10.

Court, D. – Elzinga, D. – Mulder, S. – Vetvik, O. J. (2009) *The consumer decision journey*. McKinsey Quarterly, June 2009, 1–11. <https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/the-consumer-decision-journey>, retrieved 1.5.2019.

Coussement, K. – Poel, D., van den (2007), Improving customer complaint management by automatic email classification using linguistic style features as predictors. *Decision Support Systems*, Vol. 44, 870-882.

Coussement, K. – Poel, D., van den (2008), Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications*, Vol. 34, 313–327.

Dai, Honghua – Nie, Zaiqing – Wang, Lee – Zhao, Lingzhi – Wen, Ji-Rong – Li, Ying (2006) Detecting Online Commercial Intention (OCI). *WWW 2006*, Edinburgh, Scotland, May 23–26, 2006, 829–837.

Darley, William K. – Blankson, Charles – Luethge, Denise J. (2010) Toward an Integrated Framework for Online Consumer Behavior and Decision Making Process: A Review. *Psychology & Marketing*, Vol. 27 (2), 94–116.

Debaere, Steven – Coussement, Kristof – Ruyck, Tom, de (2018) Multi-label classification of member participation in online innovation communities. *European Journal of Operational Research*, Vol. 270, 761–774.

Elamyslahjat. <https://www.elamyslahjat.fi/tietoa-meista>, retrieved 2.4.2019.

Elämyslahjat. <https://www.elamyslahjat.fi/yritys>, retrieved 2.4.2019.

Engel, James F. – Kollat, David T. – Blackwell, Roger D. (1968) *Consumer Behavior*. Holt, Rinehart and Winston, Inc., New York.

Ghose, Anindya – Yang, Sha (2008) An Empirical Analysis of Sponsored Search Performance in Search Engine Advertising. *WSDM'08*, Palo Alto, California, USA, February 11–12, 2008, 241–250.

Gómez-Ríos, Anabel – Luengo, Julián – Herrera, Francisco (2017) A Study on the Noise Label Influence in Boosting Algorithms: AdaBoost, GBM and XGBoost. *Conference paper, International Conference on Hybrid Artificial Intelligence Systems 2017*. Springer International Publishing, 268–280.

Google Help Center. <https://support.google.com/google-ads/answer/6227565?hl=fi&ref_topic=6231194>, retrieved 2.4.2019.

Gounaris, Spiros – Stathakopoulos, Vlasis (2004) Antecedents and consequences of brand loyalty: An empirical study, *The Journal of Brand Management*, Vol. 11 (4), 283–306.

Habernal, Ivan – Ptácek, Tomáš – Steinberger, Josef (2014) Supervised sentiment analysis in Czech social media. *Information Processing and Management*, Vol. 50, 693–707.

Hadrien (2015a) Rapid Immersion: SEM keywords (Part 1) – The Purchase Funnel, Reef Digital. <http://reefdigital.com.au/blog/videos/sem-keywords-part-1-the-purchase-funnel/>, retrieved 13.9.2019.

Hadrien (2015b) Rapid Immersion: SEM keywords (Part 2) – Cost-Efficiency vs Volume. Reef Digital. <https://reefdigital.com.au/blog/sem-keywords-part-2-cost-efficiency-vs-volume/>, retrieved 13.9.2019.

Hafernik, Carolyn Theresa – Jansen, Bernard J. (2013) Understanding the Specificity of Web Search Queries. *CHI 2013: Changing Perspectives*, Paris, France, 1827–1832.

Hall, A. – Conway, T. – Betts, P. – Parker, C. (2016) From economic man to connected consumers. *Proceedings of the 4th International Conference on Contemporary Marketing Issues (ICCMI) 2016*, Heraklion, Greece, June 22-24, 2016, 53–58.

Haven, Brian – Bernoff, Josh – Glass, Sarah (2007) *Marketing's New Key Metric: Engagement*, Forrester Research, Inc., Cambridge, USA.

He, Haibo – Garcia, Edwardo A. (2009) Learning from Imbalanced Data. *IEEE Transactions on Knowledge And Data Engineering*, Vol. 21 (9), 1263–1284.

Hoban, P.R. – Bucklin, R.E. (2015), Effects of internet display advertising in the purchase funnel: Model-based insights from a randomized field experiment, *Journal of Marketing Research*, Vol. 52 (3), 375–393.

Hotchkiss, Gord. (2004) *Into the Mind of the Searcher*. Enquiro Search Solutions Inc., 1–30.

Howard, John A. – Sheth, Jagdish N. (1969) The Theory of Buyer Behavior. *Journal of the American Statistical Association*, 467–487.

Hsieh, Hsiu-Fang – Shannon, Sarah E. (2005) Three Approaches to Qualitative Content Analysis. *Qualitative Health Research*, Vol. 15 (9), 1277-1288.

Hudson, Simon – Hudson, Rupert (2013). Engaging with consumers using social media: a case study of music festivals. *International Journal of Event and Festival Management*, Vol. 4 (3), 206–223.

IAB Europe (2018) AdEx Benchmark 2017 (digital ad spend in Europe), 28 June 2018, <https://www.iabeurope.eu/research-thought-leadership/iab-europe-report-adex-benchmark-2017-digital-ad-spend-in-europe/>, retreived 3.4.2019.

Im, Il – Dunn, Brian Kimball – Lee, Dong Il – Galletta, Dennis F. – Jeong, Seok-Oh (2019) Predicting the intent of sponsored search users: An exploratory user session- level analysis. *Decision Support Systems*, Vol. 121, 25–36.

Im, Il – Jun, Jongkun – Oh, Wonseok – Jeong, Seok-Oh (2016) Deal-seeking versus brand-seeking: search behaviors and purchase propensities in sponsored search platforms. *MIS Quarterly*, Vol. 40 (1), 187–203.

Jabeen, Hafsa (2018) Stemming and Lemmatization in Python. DataCamp. <https://www.datacamp.com/community/tutorials/stemming-lemmatization-python>, retrieved 9.6.2019.

Jansen, Bernard J. – Booth, Danielle L. – Spink, Amanda (2008a) Determining the informational, navigational, and transactional intent of Web queries. *Information Processing and Management*, Vol. 44 (3), 1251–1266.

Jansen, B. J. – Hudson, K. – Hunter, L. – Liu, F. – Murphy, J. (2008b) The Google online marketing challenge: Classroom learning with real clients, real money, and real advertising campaigns. *Journal of Interactive Advertising*, Vol. 9 (1), 1–9.

Jansen, B.J. – Mullen, T. (2008) Sponsored search: an overview of the concept, history, and technology. *Int. J. Electronic Business*, Vol. 6 (2), 114–131.

Jansen, Bernard J. – Resnick, Marc (2006) An Examination of Searcher's Perceptions of Nonsponsored and Sponsored Links During Ecommerce Web Searching. *Journal of the American Society for Information Science And Technology*, Vol. 57 (14), 1949–1961.

Jansen, Bernard J. – Schuster, Simone (2011) Bidding on the buying funnel for sponsored search and keyword advertising. *Journal of Electronic Commerce Research*, Vol. 12 (1), 1–19.

Jansen, Jim (2011), *Understanding Sponsored Search: Core Elements of Keyword Advertising*. Cambridge University Press, New York, NY.

Jerath, K. – Ma, L. – Park, Y.-H. (2014) Consumer Click Behavior at a Search Engine: The Role of Keyword Popularity. *Journal of Marketing Research*, Vol. 51 (4), 480–486.

Kathuria, Ashish – Jansen, Bernard J. – Hafernik, Carolyn – Spink, Amanda (2010) Classifying the user intent of web queries using k-means clustering. *Internet Research*, Vol. 20 (5), 563–581.

Kettunen, Kimmo – Kunttu, Tuomas – Järvelin, Kalervo (2005) To stem or lemmatize a highly inflectional language in a probabilistic IR environment? *Journal of Documentation*, Vol. 61 (4), 476–496.

Kotler, Philip – Keller, Kevin Lane (2012) *Marketing management*. 14th ed. Pearson Education, Inc., publishing as Prentice Hall, New Jersey.

Kotsiantis, Sotiris – Kanellopoulos, Dimitris – Pintelas, Panayiotis (2006) Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, Vol. 30, 1–12.

Kowsari, Kamran – Meimandi, Kiana Jafari – Mojtaba, Heidarysafa – Mendu, Sanjana – Barnes, Laura – Brown, Donald (2019) Text Classification Algorithms: A Survey. *Information 2019*, 10, 150, 1–68.

KPMG (2017) The path to purchase journey. <https://home.kpmg/xx/en/home/insights/2017/01/the-path-to-purchase-journey.html>, retreived 15.11.2019.

Krippendorff, Klaus. (2004) *Content Analysis: An Introduction to Its Methodology*. Second edition. Sage Publications, Inc., Thousand Oaks, CA.

Lahiri, Aditya (2018) Dealing With Class Imbalanced Datasets For Classification, Towards Data Science. <https://towardsdatascience.com/dealing-with-class-imbalanced-datasets-for-classification-2cc6fad99fd9>, retreived 13.5.2019.

Lavidge, Robert J. – Steiner, Gary A. (1961) A Model for Predictive Measurements of Advertising Effectiveness, *Journal of Marketing*, Vol. 25 (6), 59–62.

Lemon, Katherine N. – Verhoef, Peter C. (2016) Understanding Customer Experience Throughout the Customer Journey. *Journal of Marketing: AMA/MSI Special Issue*, Vol. 80, 69–96.

Lewandowski, Dirk – Drechsler, Jessica – Mach, Sonja, von (2012) Deriving Query Intents From Web Search Engine Queries. *Journal of The American Society for Information Science And Technology*, Vol. 63 (9), 1773–1788.

Li, Hongshuang – Viswanathan, P. K. Kannan, Siva – Pani, Abhishek (2016) Attribution Strategies and Return on Keyword Investment in Paid Search Advertising. *Marketing Science*, Vol. 35 (6), 831–848.

Mayo, Matthew (2018) Text Data Preprocessing: A Walkthrough in Python. KDnuggets. <https://www.kdnuggets.com/2018/03/text-data-preprocessing-walkthrough-python.html?fbclid=IwAR3_YEWmxnEZgujvJ8exYy4XXSd8BqnQg8FdxOh5QM5rG047v1AjPcuKJX4>, retrieved 15.6.2019.

Misra, S. – Pinker, E. – Rimm-Kaufman, A. (2006) *Empirical Study of Search Engine Advertising Effectiveness*. Working paper. <http://digital.mit.edu/wise2006/papers/4A-2_PinkeretalWISE2006.pdf>, retrieved 2.4.2019.

Mitchell, Tom M. (1997) *Machine Learning*. McGraw-Hill Science/Engineering/Math.

Navas-Loro, María – Rodríguez-Doncel, Víctor – Santana-Pérez, Idafen – Fernández-Izquierdo, Alba – Sánchez, Alberto (2018) MAS: A Corpus of Tweets for Marketing in Spanish. *A. Gangemi et al. (Eds.): ESWC 2018 Satellite Events*, 363–375.

Neely, Pam (2019a) 10 Trends That Will Shape Paid Search In 2019. <https://www.acquisio.com/blog/agency/10-trends-that-will-shape-paid-search-in-2019/>, retrieved 3.4.2019.

Neely, Pam (2019b) What Just Happened to Google's Exact Match? <https://www.acquisio.com/blog/agency/google-ads-exact-match/?fbclid=IwAR1k-5hl4Ytl3pXzs8Sh2HB7KTKhemMSix6lG6DrI5PdyE39Rl2bWUvabz0>, retrieved 3.4.2019.

Netzer, Oded – Feldman, Ronen – Goldenberg Jacob – Fresko, Moshe. (2012) Mine Your Own Business: Market-Structure Surveillance Through Text Mining. *Marketing Science*, Vol. 31 (3), 521–543.

Nielsen, Didrik (2016) *Tree Boosting With XGBoost – Why Does XGBoost Win "Every" Machine Learning Competition?* Master's thesis. Norwegian University of Science and Technology, Department of Mathematical Sciences.

Noble, Steven – Cooperstein, David M. – Kemp, Mary Beth – Madigan, Corinne J. (2010) *It's Time To Bury The Marketing Funnel – Marketers Must Embrace The Customer Life Cycle*. Forrester Research, Inc., 1–13.

Okazaki, Shintaro – Díaz-Martín, Ana M. – Rozano, Mercedes – Menéndez-Benito, Héctor D. (2014) How to mine brand Tweets: Procedural guidelines and pretest. *International Journal of Market Research*, Vol. 56 (4), 467–488.

Owusu, Richard A. – Mutshinda, Crispin M. – Antai, Imoh – Dadzie, Kofi Q. – Winston, Evelyn M. (2016) Which UGC features drive web purchase intent? A spike-and-slab Bayesian Variable Selection Approach. *Internet Research*, Vol. 26 (1), 22–37.

Petrik, Andrei (2014) How to Map Keywords to Each Stage of the Buying Cycle. <https://www.searchenginepeople.com/blog/925-buying-cycle-keyword-mapping.html>, retreived 8.2.2019.

Phan, Nina – Bailey, Peter – Wilkinson, Ross (2007). Understanding the Relationship of Information Need Specificity to Search Query Length. *SIGIR 2007 Proceedings*, Amsterdam, The Netherlands, July 23–27, 2007, 709-710.

Punj, Girish (2012) Consumer Decision Making on the Web: A Theoretical Analysis and Research Guidelines. *Psychology and Marketing*, Vol. 29 (10), 791–803.

Qian, Vicky (2017) Clustering Vs. Classification: How To Speed Up Your Keyword Research. <https://ipullrank.com/clustering-vs-classification-speed-keyword-research/>, retreived 8.2.2019.

Raehsler, Lisa (2019) The 8 Best PPC Ad Networks. <https://www.searchenginejournal.com/ppc-guide/best-ppc-ad-networks/>, retrieved 3.4.2019.

Rangaswamy, Arvind – Giles, C. Lee – Seres, Silvija (2009) A Strategic Perspective on Search Engines: Thought Candies for Practitioners and Researchers. *Journal of Interactive Marketing,* Vol. 23, 49–60.

Raschka Sebastian (2015), *Python Machine Learning – Unlock deeper insights into machine learning with this vital guide to cutting-edge predictive analytics*. Packt Publishing Ltd., Birmingham.

Rasmussen, C.E. – Williams, C.K.I (2006), *Gaussian Processes for Machine Learning*. the MIT press, Cambridge.

Raudys, Saunas J. – Jain, Anil K. (1991) Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 13 (3), 252–264.

Reinstein, Ilan (2017) XGBoost, a Top Machine Learning Method on Kaggle, Explained. KDnuggets. <https://www.kdnuggets.com/2017/10/xgboost-top-machine-learning-method-kaggle-explained.html>, retrived 4.6.2019.

Rocca, Baptiste. (2019) Handling imbalanced datasets in machine learning. Towards data science. <https://towardsdatascience.com/handling-imbalanced-datasets-in-machine-learning-7a0e84220f28>, retrieved 27.4.2019.

Rose, Daniel E. – Levinson, Danny (2004) Understanding User Goals in Web Search. *WWW 2004*, New York, USA, May 17–22, 2004, 13–19.

Rutz, Oliver J. – Bucklin, Randolph E. (2011) From Generic to Branded: A Model of Spillover in Paid Search Advertising. *Journal of Marketing Research*, Vol. 48 (1), 87-102.

Salisbury, J. G. T. (2001) Using Neural Networks to Assess Corporate Image. In: *Applications of Computer Content Analysis*. West, M. 65–85. Greenwood Publishing Group. Westport.

Salminen, Joni – Almerekhi, Hind – Milenković, Milica – Jung, Soon-gyo – An, Jisun – Kwak, Haewoon – Jansen, Bernard J. (2018a) Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. *Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018), 330–339.*

Salminen, Joni – Al-Merekhi, Hind A. – Dey, Partha – Jansen, Bernard J. (2018b) Inter-rater Agreement for Social Computing Studies. *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 80–87.

Salminen, Joni – Yoganathan, Vignesh – Corporan, Juan – Jansen, Bernard J. – Jung, Soon-Gyo (2019) Machine learning approach to auto-tagging online content for content marketing efficiency: A comparative analysis between methods and content type. *Journal of Business Research*, Vol. 101, 203–217.

Schönleber, David (2018) A “short” introduction to model selection - An overview over hyperparameter selection & algorithm selection with big and small data. <https://towardsdatascience.com/a-short-introduction-to-model-selection-bb1bb9c73376>, retrieved 9.6.2019.

Skiera, Bernd – Eckert, Jochen – Hinz, Oliver (2010) An analysis of the importance of the long tail in search engine marketing. *Electronic Commerce Research and Applications,* Vol. 9, 488–494.

Smith, Alan D. – Rupp, William T. (2003) Strategic online customer decision making: leveraging the transformational power of the Internet. Online Information *Review*, Vol. 27 (6), 418–432.

Song, Ruihua – Luo, Zhenxiao – Nie, Jian-Yun – Yu, Yong – Hon, Hsiao-Wuen (2009) Identification of ambiguous queries in web search. *Information Processing and Management,* Vol. 45, 216–229.

Soulo, Tim (2018) Long-Tail Keywords: The ‘Secret’ to Getting TONS of Search Traffic. <https://ahrefs.com/blog/long-tail-keywords/>, retrieved 15.4.2019.

Spink, Amanda – Wolfram, Dietmar – Jansen, B. J. – Saracevic, Tefko (2001) Searching the Web: The Public and Their Queries. *Journal of the American Society for Information Science And Technology*, Vol. 52 (3), 226–234.

Stankevich, Alina (2017) Explaining the Consumer Decision-Making Process: Critical Literature Review. *Journal of International Business Research and Marketing*, Vol. 2 (6), 7-14.

Statcounter (2019) Search Engine Market Share Worldwide - March 2019. <http://gs.statcounter.com/search-engine-market-share>, retrieved 3.4.2019.

Statista (2019) Share of individuals using the internet to find information about goods and services in Great Britain from 2007 to 2019. <https://www.statista.com/statistics/286209/internet-use-finding-information-about-goods-and-services-in-great-britain/>, retrieved 29.11.2019.

Su, Ning – He, Jiyin – Liu, Yiqun – Zhang, Min – Ma, Shaoping (2018) User Intent, Behaviour, and Perceived Satisfaction in Product Search. *Technical presentation in WSDM’18*, Marina Del Rey, CA, USA, February 5-9, 2018, 547-555.

Swan, Greg (2018) Google Shopping – How to Target Every Part of the Funnel. tinuiti. <https://tinuiti.com/blog/shopping-feed/google-shopping-funnel/>, retrieved 29.10.2019.

Tang, Chuanyi – Guo, Lin. (2015) Digging for gold with a simple tool: Validating text mining in studying electronic word-of-mouth (eWOM) communication. *Marketing Letters*, Vol. 26 (1), 67–80.

Taniguchi, Hidetaka – Sato, Hiroshi – Shirakawa, Tomohiro (2018) A machine learning model with human cognitive biases capable of learning from small and biased datasets. *Scientific Reports*, 1–13.

Tax, Stephen S. – McCutcheon, David – Wilkinson, Ian F. (2013) The service delivery network (SDN): a customer-centric perspective of the customer journey, *Journal of Service Research*, Vol. 16 (4), 454–470.

Tellis, Gerard J. (1988) Advertising exposure, loyalty, and brand purchase: A two-stage model of choice, *Journal of Marketing Research*, Vol. 25 (2), 134–144.

Vanarsdall, Jillian. (2016) What is the Buyer's Journey? <https://ppgwebsolutions.com/what-is-the-buyers-journey/>, retrieved 10.2.2019.

Vázquez, Silvia – Muñoz-García, Óscar – Campanella, Inés – Poch, Marc – Fisas, Beatriz – Bel, Nuria – Andreu, Gloria (2014) A classification of user-generated content into consumer decision journey stages. *Neural Networks*, Vol. 58, 68–81.

Wijaya, B.S. (2012) The development of hierarchy of effects model in advertising. *International Research Journal of Business Studies*, Vol. 5 (1), 73–85.

Wolny, Julia, dr – Charoensuksai, Nipawan (2014) Mapping customer journeys in multichannel decision-making. *Journal of Direct, Data and Digital Marketing Practice,* Vol. 15 (4), 317–326.

Zhang, Qianyun – Hill, Shawndra – Rothschild, David (2018) Post Purchase Search Engine Marketing. *WWW'18 Companion*, Lyon, France, April 23–27, 2018, 663-670.