Date of acceptance Grade

Assessor

Analysis of audience engagement and expressed opinion with Data Mining

Péter Ivanics

Helsinki February 24, 2018

UNIVERSITY OF HELSINKI
Department of Computer Science

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Studieprogram — Study Programme
Faculty of Science	Computer Science

Tekijä — Författare — Author
Péter Ivanics

Työn nimi — Arbetets titel — Title

Analysis of audience engagement and expressed opinion with Data Mining

Ohjaajat — Handledare — Supervisors

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
	February 24, 2018	71 pages + 5 appendices

Tiivistelmä — Referat — Abstract

The rapid digitalization of services in the recent years has created many challenges for software companies. As the amount of data collected increases, firms often lack the resources and expertise needed to process it. Data differ according to each business's portfolio, but in general they include demographic information about users and other details collected during software usage. Conducting thorough analysis of such datasets should be important to software companies, because it contributes to the discovery of previously unseen trends and hidden patterns. This in turn leads to the development of service packages.

The objective of this thesis is to study how user data is understood in related studies and to identify what techniques are utilized for its analysis. As the concept of user data is not common in the research community, its definition and scope for the purpose of this study is introduced. A practical objective of this thesis is to apply data mining techniques to the analysis of audience engagement in the Choicely voting platform.

This study uses exploratory data analysis to get a high-level overview of the data at Choicely. Computer vision is utilized to enhance image content with additional information and to prepare vote transactions for association analysis. Association analysis is used to extract frequent itemsets, whose lift values are studied in the second part of this thesis. Finally, co-clustering is used to identify the underlying structural patterns of votes by demographic groups and of content extracted from the participants' images. The research introduces these applied methods through a single contest in the platform. It is noted what type of content appears most engaging to users and what differences demographic groups show in their usage of the software.

Results show that beauty, sport and entertainment contests have a long history and presence in the platform. The results also show that there are many flaws and biases in the data. For example, it turns out that many users do not fill in demographic information on their profiles, which limits the accuracy of the findings. The capabilities and limitations of the chosen techniques manifest themselves when patterns are extracted from the data.

By implementing the chosen methods, it becomes possible to analyze different usage practices across genders and age groups. The developed methods can also point out content that engages certain user groups more than others. The proposed significance testing approach allows to pinpoint statistically relevant difference between groups. However, it is challenging to compare votes over multiple contests due to the different configurations of voting rules. The chosen co-clustering algorithm has reached its limitations and a different approach is required to achieve computational analysis of the data's structural patterns.

Avainsanat — Nyckelord — Keywords
data mining, user data, association analysis

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Contents

Appendices

1 Introduction

1.1 Background and motivation

Many software businesses collect enormous amount of data generated by end-users [WP17, Bla14, IN07]. End-user data incorporates essential information about how users interact with the system of discussion as well as tells about the users themselves [JHL15a, HMK+14, JHL+16, HLJ+16, AHTB16]. For instance, such data can explain the preferences of users, what kind of content they seek, how frequently they use the software, how they interact with the system and eachother [YKS15, OPLC+13].

Depending on the portfolio of the business, analysis of end-user data can reveal various interesting findings. For instance, the banking industry uses Big Data tools are often to analyze demographic characteristics to maintain old and establish new client relationships [WP17, Bla14]. Extracting the information from the data is challenging: most businesses collect more data than humans are able to process and analyze [IN07, WR10]. As a result, significant part of the knowledge might remain unrevealed and hence business-critical information remains unseen [WP17, IN07, WR10, TSK05].

With the continuous growth of mobile devices in numbers, the amount of data has increased significantly. Due to the wide availability and commonness of smartphones and tablets, anybody can easily generate rich data [JHL+16]. Complementary to the popularity of mobile devices, social media sites have grown a lot in the recent years [HMK+14, OPLC+13, BSG14], providing users many ways of interaction. Due to the combination of these two trends, users leave digital footprints around the Internet as location, media, numerical or textual data [YKS15].

People often express their opinion by sharing, liking or commenting on data over social networks. Study shows, that algorithms can predict users' personality traits using such data more reliably than other humans would do [YKS15]. These facts further increase the call for research in the field of user data analysis. By studying this area, researchers can understand the society, human behavior, preferences and the public's opinion better.

Data analysis tools and methods already exist to facilitate processing user data. However, these applications are often not utilized, because companies rather focus on developing their service package over understanding the information gathered

previously [Bla14, IN07]. The utilization of data analysis and Data Mining tools make the automatic detection of patterns and interesting relationships possible in the large datasets [TSK05, Fri97], which can be give a competitive advantage for businesses in the market. Studies show that the development in Data Mining and Knowledge Discovery in Databases is trending in academic research field as well [Bla14, Zar02].

The computational methods of today's world enable us not only to achieve previously impossible tasks, but also to create tools that may outperform humans [YKS15]. Because of this trait, computational techniques can facilitate our everyday life as well as help us to learn more about the tools and data we work with. On top of thats, data analysis can be performed automatically, which saves time and resources. As a result, there is a growing need for all businesses to introduce data analysis processes in their daily activities.

The goal of this research is to study the potential use of user data in business development and scientific research with the application of Data Mining methods. More specifically, one of the goals of this thesis work is to understand what kind of use cases have been reported in the relevant literature in this area. In order to do that, the research aims to study how content on the Internet is observed by the users and how users interact with various software solutions. Secondly, it is studied what kind of feedback is received from the engaged audience about the presented content in terms of like activities and comments. Thirdly, the tendencies among the demographic characteristics between users are studied in relation to the content which gets them engaged online.

To address this reserach area from a practical perspective, the Choicely voting/audience engagement platform is taken as a case study. Computer Vision, Exploratory Data Analysis, Association Analysis and Co-Clustering are utilized to understand the online content and to reveal yet unseen details about the tendencies of user preferences in terms of votes casted on contest participants in the platform. This is done with the aim of improving the quality of service provided by the company, so that customers of the firm can analyze the data they have collected.

Motivating aspects behind conducting this research are the stimulating challenges of studying user data and the wide range of possibilities in the information that may lie around in databases. Previous research has proven the relevance of statistical analysis on Big Data, such as like activities [JHL15a, JHL+16, OPLC+13, GRZ+16, JHL15b, YKS15], user comments [JHL+16], image tags [JHL+16], image content

[HMK+14, BSG14] and movie ratings [SWE04, KCC12] by revealing interesting findings about user behavior. Moreover, as service providers often get access to user demographics-related data through social network sites in the present time, new possibilities become available to seek correlation between user segments. Studies conducted in this field are also interesting from human behavior point of view, which is another motivation towards studying this area.

The motivation behind this research from the case company's perspective is mainly oriented towards enhancing the already existing service package that is provided to customers. At the beginning of this research project, the company did not utilize any advanced data analysis tools. This thesis work is motivated in the direction to establish the basis of a Data Mining framework at the company and hence increase the business value of the firm. As the amount of data collected by the company is too large to analyze manually, the application of Data Mining techniques assists the firm and its customers to

- understand the composition of their active user base better,
- have more advanced, quantifiable means on the collected data,
- gain an understanding on the users' behavior,
- increase business value by revealing previously unknown patterns.

1.2 Research questions, scope and objectives

In this chapter, the research questions, scope and objectives are presented and explained. The three main research questions are as follows:

RQ1: **How is user data understood and utilized in previous research?** The aim of this research question is to discover the conceptual understanding and potential development areas of user data based on related studies. The scope of the concept is defined and it is discussed, how other researchers have studied user data and related areas in the past.

RQ2: **What kind of content is more engaging for users and draws the most attention in the Choicely platform?** The aim of this research question is to understand which kind of voting contests tend to engage a larger audience. Secondly, the question targets to analyze the attributes of those contestants, who received high number of votes. In other words, the aim is to find common features of contestants, who are highly rated by public opinion.

RQ3: **What are the behavioral characteristics of users by gender, age and location?** This research question targets to answer the question how users tend to use the platform. This covers the analysis of what kind of content users seek, how the votes are spent and what similarities can be observed in the data. To gain more specific understanding, the users will be grouped by their demographic data.

Initially, the study presents an overview on the field of user data analysis in different areas of research as the theoretical part of the study. Particularly, the study addresses the possibilities and challenges of analyses performed on user data in software applications, where users express their opinion via the usage of the system. For example, one can think of like activities on social media, reviewing movies online or voting on their favorite moment of a football match.

Building on top of the theoretical framework, the study is oriented towards applying techniques for Data Mining purposes at Choicely. Being a large field of science, the scope is focused only on a subset of applicable Data Mining techniques and methods, that are utilized to answer the research questions of this study. Specific topics from the field of Data Science are chosen to obtain the answers to the research questions, such as exploring data, data visualization, association analysis and pattern mining.

One of the objectives of this thesis work is to develop advanced data analysis tools in order to assist the case company and its customers to gain a better understanding on the data at hand. In order to do that, the available data is presented and analyzed, the most interesting questions are stated and the research gaps with more influential business value is identified. Afterwards, data analysis methods are discussed which are capable to retrieve such information from the given data set. Finally, the behavior of different demographic user groups is studied in the Choicely platform in terms of their activities and what kind of differences can be identified among the different user groups.

This research is focused on the challenge of dealing with user data from Computer Science perspective. Although human behavior and psychology experts can benefit from conducting studies in this area, the focus is not shifted to that field of sciences in this work.

1.3 Introduction to the Choicely voting platform

Choicely[1] is a voting platform developed by the Finnish Choicely Ltd (formerly known as Lovented Ltd) since 2014. The software provides the possibility for users to engage in interesting audiovisual polls/contests by voting on their favorite contender. The platform has already hosted numerous contests in various fields, such as beauty pageants, public polls, design contests, talent shows, sport events and many others. The customer base of the firm consist of mainly Finnish broadcasters, publishers, advertisers and beauty pageant organizers. On top of the core customer base in Finland, the recent years have brought numerous users and customers from all around the world.

The platform is able to host any kind of visual contests or poll. There can be arbitrary number of participants in the contests. Participants must have at least one image and a name attached to them in order to enter the contest. Participants can have also a video or multiple images, however the scope for this research is limited to study votes casted on single-image contestants. The participants are added by the contest's organizer.

Figure 1 displays an example of how a contests may look like on a mobile device, when the user is browsing the participants. On the left side of the figure the contest participants are shown in the contestant list (in this case snowboarders), the middle

[1] `http://choicely.com/`

image shows one of the contestants in a full-screen view, while the right side of the figure displays the results. Users can cast votes by pressing on the star under the participants' images.

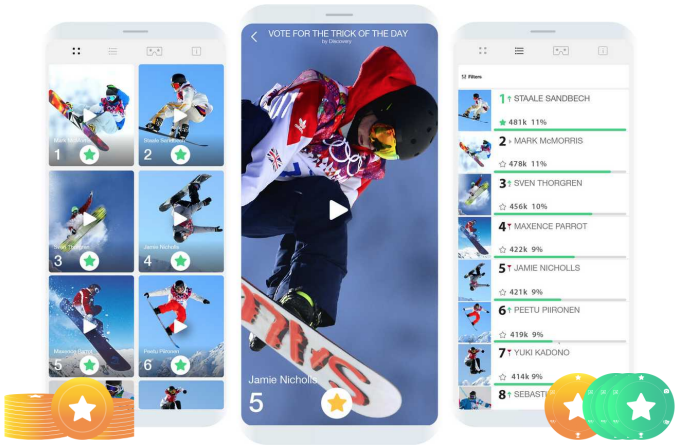


Figure 1: The visual appearance of a contest in Choicely.

Users can use the software through multiple interfaces. Naturally, the company's webpage provides a convenient way to create, browse and vote in contests. Choicely also has free mobile applications available on Android and iOS devices, that can be installed through the Google Play[2] and the iOS App Store[3] to the users' devices.

The company also offers a web widget, which can be embedded as a framework in any webpage easily. The widget is often used by Choicely's customers, because it provides a convenient way to embed rich content in their own web pages, which users are already familiar with. Contests cannot be created through the widget, but users can cast votes the same way as they would on any other platforms. Figure 2 displays the front-end interfaces which the users interact with.

[2] `https://play.google.com/store/apps/details?id=com.choicely.android`
[3] `https://itunes.apple.com/fi/app/choicely/id1158798364`

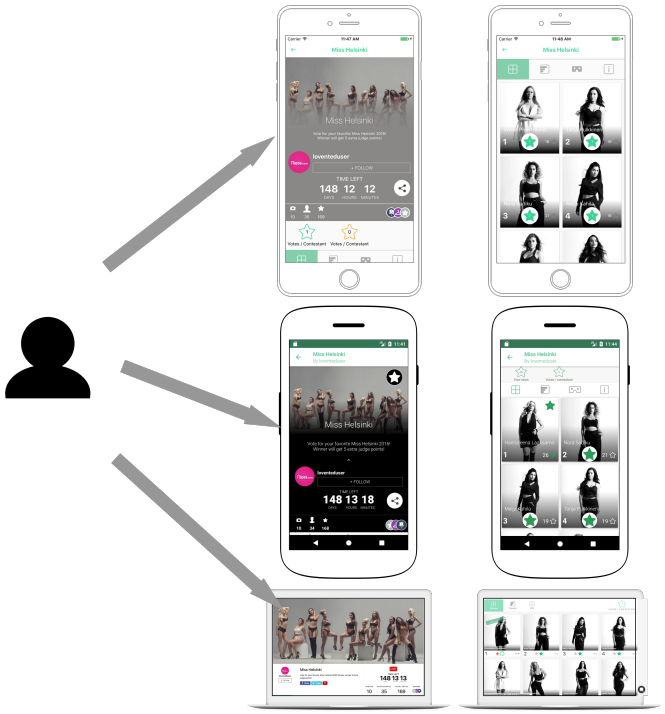


Figure 2: Users can vote in contests through three interfaces: iOS devices, Android devices and web.

Users can create contests and cast votes in contest on any of these platforms. Users may vote in arbitrary number of contests. Users may vote and participate in their own contests if they like. The voting rules can limit the number of votes that users can cast in a contest as well as on individual participants.

Contests have meta data which can facilitate data analysis. Each contest belongs to at least one but up to three of the following categories: "animals", "beauty", "danger", "design", "entertainment", "fashion", "food", "games", "humor", "sports", "travel", "other". The category labels are aligned by the contest's organizer upon creation. Contests also have a starting and an ending time, between which users can vote. While votes arrive into a contest, the vote count value tells the number of total votes received, while the number of unique voters tells how many individual users have voted in the contest. Both of these values can be retrieved for individual participants as well.

Information about contestants is limited to the contest they contend in, their names, short descriptions, an image and an optional video. As a result, currently it is not possible to characterize contestants based on the available meta data. In other

words, one of the limitations of the platform is to extract information concerining the contest participants and their traits. Every user has a profile, where they have the possibility to provide some demographical data about themselves.

There is currently no way of knowing any tendencies about what kind of participants are more engaging to the audience. Furthermore, yet there is no possibility to know if there is a similarity or difference in the behavior of demographic groups in terms of what kind of content they spend votes on.

The goal of this thesis work from the company's perspective is to address these challenges by developing a data analysis framework that addresses these gaps. In order to perform that, a way to supplement missing piece of information from the participants images is needed, which is another aim of this work. Through the results of this study, the customers of the firm can get access to a tool, which assists them in retrospectively analyzing the engagement of their audience.

1.4 Thesis structure

This chapter presents the structure of this thesis work. The next chapter explains the methodology that is chosen to address the research questions stated above. Chapter 3 presents an overview on the theoretical background and on the related work in the field of audience engagement and user data analysis. Chapter 4 presents the Choicely voting platform in-depth and explains how the data analysis methods are applied at the company from a practical perspective. This section presents also the important bits and pieces that are relevant for the technical part of the study, such as the voting mechanishm and the data structure. Chapter 5 summarizes the results of the project at the case company. The results of are listed under subchapters by each method that was utilized. Chapter 6 opens up the discussion on the results, connects the studied case to the reviewed literature, evaluates the obtained findings and derives implications from the results. This chapter is also dedicated to evaluate the chosen methods, their relevance and applicability to the problem. Finally, Chapter 7 concludes the study and points out directions for further research and development.

2 Methodology

The research consists of a theoretical and a practical part. First, an exploratory literature review is carried out to establish the basis, the relevance and the background of this study. As the research continues with the empirical part, the exploratory literature review forms the basis of understanding the background and assists to answer the research questions and the challenges at the case company.

The sources for the literature review include scientific journals, articles as well as textbooks related to the topic of this study. The two former sources are retrieved through three online digital libraries: the ACM digital library, IEEE Xplore and Google Scholar. After finding the first papers, the snowballing technique was used to retrieve further papers in the field. Various keywords were used to obtain related literature in the research area. Keywords included, but were not limited to "data mining", "social media", "user data", "user behavior", "demographic characteristics" and "digital footprints".

The search for the retrieved papers was performed between May 2017 and October 2017. The retrieved papers were then critically analyzed and the most interesting points from and across the studies are summarized in the corresponding chapter. The emphasis is placed on the research goals towards which other studies utilize demographic and user data as well as the computational methods chosen for addressing those goals.

This part of the research forms a sound basis on the understanding and possible utilization of user data. Furthermore, it establishes a common ground for the practical part of the study and helps to address the research gaps and questions. By reviewing related papers it also becomes clear, what kind of challenges are faced and techniques are utilized by other professionals in similar studies.

In the practical part of the study, the data of the Choicely voting platform's databases is analyzed. The analysis is fundamentally focused on two topics: the content uploaded by contest organizers and the users' behavior while using the platform.

The data for this research is provided by the case company. The chosen techniques are applied and the analysis is performed on historical data, which was gathered through contests and votes in the past by Choicely and its customers. The structure and the properties of the data at hand are explained in the Chapter 4.

To grasp on the currently available data, Exploratory Data Analysis (hereinafter EDA) is utilized as the first step of the research. The term of EDA originates from

Tukey [Tuk77], who intially proposed an informal study of data repositories. On top of introducing the term, Tukey's work proposes various approaches and best practices for data exploration. Over the years, the EDA concept has become widely accepted and utilized as a tool for understanding and getting an overview on data.

The underlying reasoning behind performing EDA in this study is its wide usage among researchers to explore the dataset at hand. This way preliminary hypotheses can be made about the data as well as erroneous data can be removed before the beginning of the real data analysis. EDA in many cases also helps to choose the appropriate preprocessing and data analysis techniques for the rest of the study [Tuk77, TSK05].

Accordingly, comprehensive overview on the data related to the most important entities in the Choicely platform is performed by calculating basic statistical measures and visualizing the data. Such entities are the user profiles, the contests, contest participant and the votes performed by users on the contest participants. The findings of the EDA are then utilized to determine how to prune the data such that the acquired sample is representative, does not contain redundant nor faulty data. The results of the EDA are explained in the next chapter.

Based on the results from the EDA, a subset of contests is chosen for more careful analysis. The subset of the contests is limited to those, which have engaged an adequate number of users and gathered a high enough number of votes and unique voters (users who have voted at least once in the contest). After applying the pruning rules, some part of the data is preprocessed so that the analysis can be performed easily. Preprocessing in this case means joining the datasets together along the identifiers of contestants, images or user profiles, respectively.

To address the lack of meta data on the contestants images, Computer Vision is utilized. Being a widely used technique, researchers have successfully utilized this technique in similar studies for various purposes [HMK+14, FNAC15, HLJ+16, BSG14]. One of the areas where the technique was used with great success by Farseev et al. [FNAC15] is the extraction of concepts from images. In this study, the field of Computer Vision is applied through the application of Google Vision to identify the content that is on the contestants' images in the Choicely platform.

Previous studies have looked into how user behavior can be modelled and how similarities between groups' preferences can be measured. Several related studies have utilized different methods, such as Text/Natural Language Processing [OPLC+13, FNAC15, JHL+16, KCC12, HLJ+16], Supervised Machine Learning [WP17, SWE04,

KCC12, FNAC15, HLJ+16, JHL15b, BSG14] or Unsupervised Machine Learning [SWE04, HMK+14, JHL15b] to study similar areas, such as like activities, gender differences or clustering of users.

Similarly to Ottoni et al. [OPLC+13], this study utilizes Association Analysis on the combined demographic, vote and the image label data in the chosen contests. Association Analysis was originally formulated and studied by Agrawal et al. as Associations [AIS93a, AIS93b], which has grown to a large extend in the numerous research fields. Research areas in this field also include Frequent Itemset Generation and Association Rule Mining, originally introduced by Agrawal et al [AIS93b]. This study utilizes these techniques on the given data to identify behaviors and preferences of the different demographic groups. The paragraphs to follow introduce this technique based on the textbook written by of Tan, Steinbach and Kumar [TSK05].

In particular, frequent itemsets are generated from the computer-vision recognized image labels in and users' voting data in the Choicely platform. In other words, the contestants' images in the platform are facilitated with the list of labels. The list of labels is gathered for every vote transaction and is handled as a container (similarly to a "market basket" or "market data" [BMUT97, BMS97, RC10]) out of which frequently voted items are extracted from.

More formally, the support count of an itemset can be defined as

$$\sigma(X) = |\{t_i | X \subseteq t_i, t_i \in T\}| \tag{1}$$

where $T = t_1, t_2, ..., t_N$ is the set of transactions, X is the itemset of discussion and $|\cdot|$ denotes the number of elements in an itemset [TSK05]. A further essential constraint on the itemsets is that they have to contain at least one item [TSK05]. In simple language, this number tells how many transactions contained the itemset X in the given population. The itemset that contains k unique items is often referred to as a *k-itemset*.

Deriving from this, the supports for each itemset can be calculated, that is the percentage of transactions in which the itemset is present. In other words, the support tells the probability of itemset X to be present in a randomly chosen transaction. Formally, this can be stated as

$$S(X) = \frac{\sigma(X)}{N} \tag{2}$$

where N is the number of transactions. In this thesis work, vote transactions are

filtered by demographic groups G and the support $S(X|G)$ is calculated. For instance, by calculating $S(X|gender = male)$ one can obtain the itemset supports for males in a given list of vote transactions. Note that other researches might use the term confidence for studying association rules, but the present study consistenly uses the term of support with this meaning. The extraction of the itemset supports is achieved using the Apriori algorithm, which addresses the complexity growth of itemsets during their generation [TSK05].

The calculated support values can be studied further such that the tendencies in voting among demographic groups can be compared (RQ3). To answer RQ3, the next task is to seek significant differences or interesting similarities in the itemset support values of demographic groups. There are two challenges with this goal:

1. the number of potential itemsets can be large, and
2. there is not a clear way of knowing which of the resulting itemsets will be interesting to look at from the viewpoint of a data scientist.

The first challenge is addressed by the Apriori Principle, originally introduced by Agrawal et al. [AS94]. The principle states that any subset of an infrequent itemset is inevitably infrequent subset as well [AS94, TSK05]. This means that if the algorithm encounters an itemset under the desired minimal support threshold, its subsets can be safely dropped from any further analysis. For this reason, Association Rule Discovery algorithms often introduce a *minsup* threshold [AIS93a, AIS93b, TSK05], which controls the value below which itemsets are considered as infrequent. As a consequence, less popular itemsets can be filtered out and the more interesting itemsets can be pinpointed.

Lift is another common measure in Association Analysis, which fits the purposes of this study very well. In his textbook Tuffery [Tuf11] provides a great introduction to this measure and argues, why it complements association rule mining. In his terms lift also tells the improvement introduced by a rule [Tuf11], which can assist researchers to find out which rules are more (or less) useful for the analysis. The lift is hence a measure on the dependency between condition and the consequence of the association rule. Obtaining the lift value is a simple division between the support of the rule and the result's probability. Formally, it can be stated as

$$L(X|G) = \frac{S(X|G)}{S(X)} \tag{3}$$

The value of lift varies on the range of 0.0 to positive infinity. If the value of the lift is 1.0, the condition makes no impact on the result of the rule, hence is typically not considered very interesting. In case the value is less than 1.0, the rule is said to be useless, while over 1.0 is usually called as useful rule [Tuf11].

The aim by calculating lift values in this research is to identify how different age groups interact with the itemsets. For instance, if the itemset X is more appealing to males than how often it appears in the complete dataset, its lift value $L(X|gender=male) > 1.0$. Similarly, if the opposite can be said for females, the $L(X|gender=female) < 1.0$. In other words, the more engaging itemsets have higher lift values which can suggest more attraction from the audience and hence lead to further, more careful analysis.

Tackling the second challenge stated above is a bit less obvious. The itemset supports can be calculated for all demographic groups in a single contest or a list of contests (e.g. every contest in certain category or contests hosted by the same organizer). The frequent itemsets (which exceed the *minsup* threshold) are organized into the rows of a matrix during the analysis, with each the columns representing a demographic group which is being analyzed. The cells in the matrix contain the support values of the itemsets for the given group of discussion.

Knowing which itemsets or demographic groups show similarities to eachother could be extracted from such matrix. After collecting the image labels and extracting the itemset supports, one may want to identify patterns or the underlying structure in the data, so that interesting findings can be extracted from it. To address this problem, the Co-Clustering (often called also as biclustering or two-mode clustering [VMBDB04]) is used, which was first introduced by Hartigan [Har72] on the historical election voting data between 1900-1968. The main goals in his study are to partition the given dataset such that the similarities between states and over years are extracted from the dataset [Har72].

The novelty of the Co-Clustering approach lies in identifying patterns simultaneously on columns and rows of a (large) matrix of numbers [Har72, VMBDB04]. Like other clustering techniques, this is a Unsupervised Machine Learning technique as well, because the structure of the data is not known beforehand. The key advantage of Co-Clustering compared to other methods on such data is that the clusters can be interpreted directly after its completion [Har72]. In other words, the output directly tells the boundaries of the clusters and matches items to the clusters they belong to.

The reason behind choosing the Co-Clustering approach is its relevance and applicability to this problem. In their structured study, Van Mechelen, Bock and De Boeck explain the mathematical details behind the different approaches in depth [VMBDB04], which can vary from simple division of rows and columns of the data matrix to overlapping and even nested clusters. The researchers conclude, that the large number of models already had an impact on medical field, as it provides tools for DNA sequence analysis or classifying syndromes based on patients' symptoms [VMBDB04]. The widespread suitability of the approach is proven by other reserachers in various fields, such as modelling text documents and words [Dhi01], gene expression data, time course data and sensor network data [Cho08].

Van Mechelen, Bock and De Boeck also argues [VMBDB04], that despite its robustness, the two-mode clustering has not reached as many applications as its "simpler" version, one-way clustering. This is another motivation towards experimenting with this method and evaluating its applicability to the problem studied in this work. For simiplicity, this research utilizes data partitioning [VMBDB04], which can be seen as the simplest Co-Clustering method. Data partitioning in this respect means the identification of areas in the data, where items semantically belong together.

Co-Clustering in the present research is performed on the support values of itemsets by demographic groups. That is, the input matrix lists the itemsets constructed from the labels of the contest participants' images, while its columns contain the demographic groups, for which the supports have been calculated. The cells of the matrix containt the actual support values, which range from 0.0 to 1.0 inclusive.

The aim of applying this method is to identify which demographic groups as well as itemsets show similarities to eachother. By using the Co-Clustering method, the rows and columns of the constructed matrix can be analyzed, its rows and columns are clustered at the same time. This way, the grouping of the similar entities (rows, columns and cells) can be achieved, which contributes to identifying patterns and revealing the underlying structure of the data.

The combination of these methods is unique as none of the studied researches utilizes them in such combination with these kind of goals in mind. The data available at Choicely contains a great specimen of user data. The potential application of these methods are interesting for the case company, as their platform currently does not include any tool with this purpose, while the need from the customers' side continously rises.

3 Audience engagement and user data analysis

Careful analysis on Big Data can enhance business domain understanding for any company. Accordingly, Big Data has became a widely studied and interesting topic all around the field Information Technology in the recent years, both in scientific and industrial research [IN07, TSK05]. Significant amount of the data is somehow related to the users of the software which is responsible for the data collection. Data in this setting often incorporates essential information about the users, such as their demographic data, preferences or how they interact with the software [JHL15a, HMK+14, JHL+16, HLJ+16, AHTB16].

While exploring the background and related literature of this study it was identified, that the concept of user data is not commonly used among researchers. Most of the reviewed studies simply use the terminology that is related to the domain of the study, for instance like activities, user reviews or ratings. The studies often utilize demographic data (typically gender and age), however the concept of demographics and usage patterns based on digital footprints are often used separately.

To set a common ground and understanding for the rest of this thesis work, the concept of user data is defined. User data in this study covers two major kind of data that is related to users. On one hand it covers data, which is willingly uploaded and shared by users, for instance their location information, gender, age or other demographic attributes. On the other hand, user data consists of data which generated through interactions while using the software solution of discussion. User data in this study is the combination of these two aspects, which is demonstrated on Figure 3.

The first part of user data typically consists of the demographic characteristics of the users and the content which they have generated [HLJ+16]. Traditionally demographic data is often collected via surveys and questionnaires by researchers. This kind of data often can be found on social media profiles where social demographic information, such as gender, age and nationality are shared. As a result, this kind of data can be obtained relatively easily in vast amounts, if users are eager to share with the software service provider.

Secondly, data gathered during the usage of a software is an important source of information for scientific studies. Researchers use the term digital footprint [YKS15] to describe this kind of data, which composes the second type of user data in this research. Such digital footprints can vary a lot depending on the software of discus-

sion, but typically consists of like activities, user comments, photos, posts on social media, ratings in movie databases or browsing history.

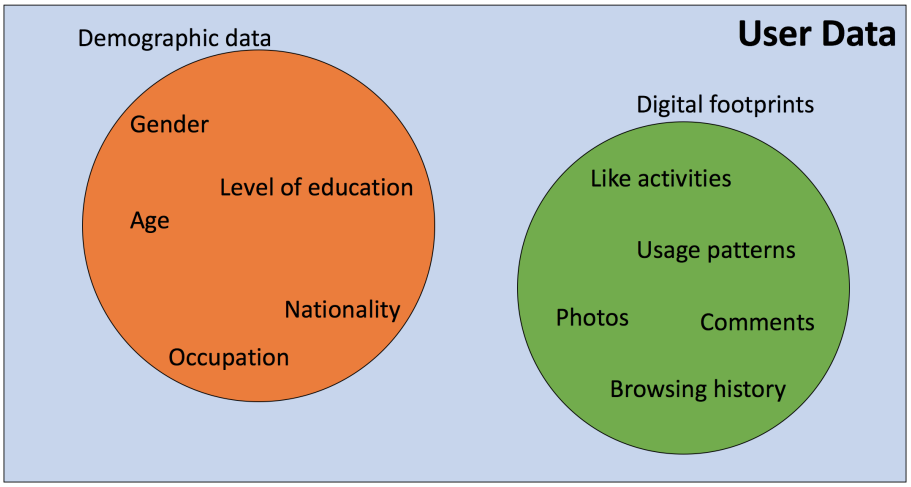


Figure 3: User data in this research is the combination of users' demographical characteristics and digital footprints (the data items mentioned are only examples, there could be other data items in both sets).

3.1 Related research

Previous research have proven the relevance and applicability of Data Mining methods on user data with various goals. These goals include, but are not limited to studying like activities [JHL15a, JHL+16, OPLC+13, GRZ+16, JHL15b], user comments [JHL+16], tags under images [JHL+16], image content [HMK+14, BSG14], movie ratings [SWE04, KCC12] and web usage mining [SCDT00]. All of these studies have revealed interesting findings in their domain based on the data at hand. Moreover, as service providers often get access to user demographics-related data through social network sites in the present time, new possibilities become available to seek correlation between user segments. The paragraphs to follow explain some of the use cases where user data was utilized in order to adress RQ1 of this study.

The study conducted by Wang and Petrounias [WP17] shows that mobile banking in China is more popular among middle-aged males, while the younger generation has not adopted to the new trends yet. By utilizing Big Data analytics, the group of citizens and products for the upcoming marketing campaigns were revealed [WP17], which greatly enhances the marketing activities of financial organizations. Social

diversity was also studied by other researchers in the context of software development growth [AHTB16]. In their study, Aué et al. [AHTB16] have clearly identified correlation between the success of open source projects and the contributors' gender and cultural diversity by utilizing well chosen statistical methods.

Movie databases often contain user reviews on movies, actors and producers of all sort. Such databases are open and are available for the public, and therefore the amount of data has grown huge over the past years. Unsurprisingly, databases like the Internet Movie Database (IMDB) has drawn the interest of researchers [SWE04, KCC12, SKM]. The successful application of statistical methods and data mining techniques have revealed interesting findings. Studies have proven that larger budget for movies does not necessarily result in good ratings by the public, while actors have higher impact on the opinion of the audience [SWE04]. On top of deriving such conclusions, machine learning techniques are emerging to predict future movie rating data, based on prior reviews of users [SWE04] or the analysis of genre and other attributes of movies [KCC12]. Opinion mining is another area of development in this field [SKM], however it is mainly operating on unstructured, textual data via the reviews of movies.

Social Networking Sites (SNSs) are another trending source in discovering the secrets of user data as the number of scientific publications in the topic has increased significantly in the recent years [WARK17]. Various researches have applied advanced Data Mining techniques on Instagram data [JHL15a, BSG14, HMK+14, JHL+16, HLJ+16], more specifically on the tags and comments that are attached to the images. Similarly, like activities and user-generated content is studied by the scientists. It was revealed, that Instagram users can be divided into two groups based on their activities: specialists, who publish and seek content around a certain topic of interest; and generalists, who are interested in all kinds of genres in the social media site [JHL15a]. Data Mining techniques also allowed researchers to conclude, that the teenager users of Instagram tend to be more active, faster to react and more open to communicate with other users on social media [JHL+16, HLJ+16]. Furthermore, it was discovered that media content with human faces are more engaging than other type of media [BSG14]. Finally, rich social media data allowed researchers to analyze behavior and user preferences among genders, age groups and locations [FNAC15].

Like activities performed on social media are widely studied concept [BSG14, JHL15a, JHL+16, OPLC+13]. Comments, hashtags and content that is generated by users

are also widely studied [BSG14, JHL+16, HMK+14, BSG14] and also contribute to this kind of data in this study. However, little research has been conducted in the field on how these usage patterns can be projected onto demographical data, which can greatly contribute to understand user behavior of certain target groups.

Interestingly, most of the SNS-related studies are conducted in the United States of America and Asia [WARK17]. According to Waheed, Anjum and Khawaja, only a few studies were conducted in the European region, which allows to conclude that there is a room for research in the area [WARK17]. However, it is important to highlight, that due to the wide popularity of the international social networks (such as Facebook, Instagram or Pintrest), some part of the data may be derived from users in the European continent. They have also pointed out that some researches focus on sites, that are specific to a particular region or country [WARK17], which means the findings are strongly related to the cultural environment of the user base. Nevertheless, this finding may imply either the lack of interest in the area or the variance in cultural background of users.

Recently Facebook has introduced reactions among their features. Through reactions, users can not only "like" content, but also express other emotions, such as love, joy, amazement, anger or sadness [Hut17, Fra17]. This way users' emotional feelings about the content can be collected easily and efficiently. Shortly after its release, it was identified that the new feature is very popular and generally engages a wider audience than previous likes and comments [Hut17]. Study also shows, that reactions are a great way for publishers to get a feedback on the public's opinion [Fra17].

Social media platforms in enterprise environment are studied similarly to regular social networks [GRZ+16]. Despite the fact that the two types of social media platforms share many features, analysis performed in enterprise environment can provide great insights about how employees interact and cooperate. Among many other findings, studies have shown that blogs posts in an enterprise social media site tend to be more engaging and contribute to form communities inside the organization [GRZ+16]. Such insights are essential for higher management, because they can be used to identify departments that tend to be less interactive or less engaged.

3.2 Tools and methods

Various methods were utilized in previous research to conclude the findings above. The major categories of the techniques and their purposes are explained in the paragraphs to follow. Table 1 below lists the same findings.

The combination of multiple data sources containing users' demographic data, such as Facebook, Instagram, Pintrest or other social media services, is a common practice in researches [FNAC15, OPLC+13]. Researchers have managed to perform complete demographic profiling and concluded that the integration of multiple data sources is indeed a great method of enhancing performance on user data analysis [FNAC15].

Utilizing additional tools, existing datasets can be further enhanced for data analysis purposes. A great specimen for this is computer vision, which is used in numerous studies to identify content of photos [HMK+14, FNAC15]. In two of the reviewed papers [HLJ+16, BSG14], computer vision techniques are utilized to identify faces on and predict ages of people from photos uploaded to social media sites. Through these researches, computer vision was proven to be a powerful tool to facilitate the knowledge by providing researchers further data for their studies.

Text processing techniques are potentially the most common way to discover the insights of user data. The Linguistic Inquiry and Word Count (LIWC) and Latent Dirichlet Allocation (LDA) methods are used by researchers [OPLC+13, FNAC15, JHL+16] for extracting linguistic features of text. In their study, Jang et al. use this technique to identify keywords and perform topic modelling based on users' comments and hashtags on Instagram [JHL+16]. Topic modelling was also applied for analyzing genres of movies by their title and description [KCC12]. Applying natural language processing techniques also allowed researchers to infer age and gender of users [HLJ+16] who have commented on content on Instagram. Utilizing these methods greatly contribute to the understanding of data gathered, which can lead to richer analysis and pattern recognition on the dataset at hand.

Supervised machine learning appears to be a popular technique for prediction tasks performed on user data. For example, decision trees are utilized as prediction tools for the purposes of identifying audience of mobile banking software in China [WP17]. Decision trees were also successfully used for movie rating prediction [SWE04] and classification of movies [KCC12]. Bayesian networks in the same study [KCC12] are also investigated and are found as the most suitable method for predictions in the field of movie rating prediction. Ensemble Modeling, which is another supervised

Table 1: The summary of utilized data mining techniques on user data in other researches.

Technique	Number of studies
Computer vision	3
Text/Natural language processing	5
Unsupervised machine learning	2
Supervised machine learning	5
Frequent Itemset Analysis	1

learning method, is successfully used for user profile learning on social media sites [FNAC15]. Similarly, Support Vector Machines and Logistic Regression was utilized for predicting age group of users based on their social media activities [HLJ+16]. Last, but not least Negative Binomial Regression is put in use to model like activities in researches [JHL15b, BSG14].

Unsupervised machine learning techniques are also present in the literature, however on a smaller scale. As an example, clustering is used by Saraee White and Eccleston [SWE04] for detecting relationships between the ratings given on a movie and the year it was published. The study by Hu et al. [HMK+14] uses the k-means algorithm was used successfully to create five clusters of Instagram users based on the type of content they have uploaded to Instagram. These results share similarities with another study by Jang et al. [JHL15b], where generalist and specialist groups were identified based on their like activities based on natural language processing techniques.

Frequent Itemset Analysis is utilized by Ottoni et al. [OPLC+13] for user portfolio analysis and detecting differences among genders in their data. By utilizing this technique, the researchers have identified significant differences between the two genders' preferences in terms of online content [OPLC+13]. Based on their research it can be concluded, that this technique can be a great tool to perform comparative analyses between demographic groups or pieces of content.

3.3 Summary

The results of this literature review shows, that there is not a common understanding on the term of user data among researchers. While demographic data is very well understood, digital footprints (data gathered during the usage of a software) are less commonly studied for research purposes. In other words, user demographic data and digital footprints are not usually brought under the same hood, are often vaguely defined and sometimes are studied separately. It is clearly identified that the careful utilization of user demographic and digital footprints, software service providers and researchers can retrieve previously unknown information about users' behavioral characteristics.

Access to demographic data in the present time seems to be getting easier for researchers. Social networks already have standard ways of helping users to authenticate themselves while using other services (in other words use their social network credentials to use other services), businesses can get access to rich demographic data of their user base easily.

The examples in the previous sections demonstrate the relevance of Data Mining and Machine Learning methods in the field of user data analysis. The proper application these methods allow us to learn more about the society as well as human behavior, which was not possible in the past. This information is essential for business operations, because it gives insights on the user groups, their preferences and what kind of content keeps the audience engaged. Furthermore, these tools provide researchers novel ways to look at human behavior and differences between demographic groups.

From the chosen set of related studies it seems, that natural language processing and supervised machine learning techniques are the most common approaches for conducting research on user data. The encouraging results achieved by other researchers prove the relevance of computational techniques applied on user data in a wide range of development areas, such as banking, social media, online movie databases, user portfolio analysis to name a few. Despite not being part of this short study, recommender systems could be an interesting topic research area.

In the past these kind of insights were unavailable to researchers and content providers. In the present time, access to this information can facilitate research, business processes, helps determining the future content, analyzing trends and understanding target groups of particular services better. In sum, modern data mining techniques created the potential to study human preferences and behavior from a novel angle.

4 Data mining at Choicely

This chapter explains the Data Mining tasks that are introduced at Choicely through this thesis work. The first subchapter introduces setting in which this research is conducted. The reason behind this is to provide more insights over the introduction presented in Chapter 1.3. Secondly, learning about the system's core functionality related to this research provides a better understanding for the rest of this study. The second subchapter addresses this space and tells about the rules of contests and voting. The third subchapter introduces relevant insights on the company's data related to this research. This subchapter can provide interesting insights to all researchers, who conduct similar studies at companies or wishes to learn how data is typically stored in such research environment. Finally, the last subchapter argues for the relevance of the selected methods and explains how they are applied in this study.

4.1 Research setting

Performing scientific research on such data is interesting for multiple reasons. To begin with, at the time of this research Choicely did not utilize data analysis tools on the collected data. It is in the interest of the company and its customers to better understand what kind of audience was engaged in the past, what kind of content is more (or less) successful and what tendencies in user behavior can be extracted from the data. By doing so, both parties can provide better services and richer content to their targeted customers. Therefore, the introduction of data analysis and visualization tools at the company will greatly enhance business value of the firm, provide deeper understanding on the domain as well as the existing user base.

Secondly, Choicely can be looked at as a social network, because some of the pictures uploaded to the platform is generated by users. Users also have the possibility to express their appreciation or support towards some contest participants by spending votes on them. Similarly to social networking sites, where the "like" feature is often used [JHL15a, BSG14] this phenomena can be looked at as a way of expressing personal opinion.

In comparison to most of the currently available social networks, voting platforms like Choicely are observed by the audience differently. On one hand, social media sites usually list posts or images on a feed, where there is theoretically no relation between the posts that follow eachother. On the other hand, contestants in the

Choicely platform share similarities as they were nominated for the same contest. Accordingly, there must be some similarity among them as all are subjects of the contest's topic, rules and are competing for the best possible result.

This slight difference can make a big change in terms of user behavior. The focus moves from "what kind of content I like" to "which piece of content I like the most in comparison to the rest". Consequently, users will scan through some (or optimally all) of the contestants and make unconscious decisions upon whether to give vote(s) on certain contest participant(s). The users express their favor and support towards a subset of contenders, hence helping them to reach their ultimate goal: winning the contest. This uniqueness compared to other social networking sites offers a great possibility for research. Therefore, the hands-on goals towards the analysis are two-folded:

1. identifying what kind of images among different contests users like, and
2. what kind of content similar group of users like.

One of the challenges in connecting users with topics is that there is no indication on what the content on the participants' pictures is. Contest authors only assign categories to the contests to be created, which does not necessarily describe the entrants. Other researches have successfully utilized computer vision to gather meta data for the uploaded content in image sharing communities [BSG14, HMK+14].

Deriving from the success in previous studies [HMK+14, FNAC15, HLJ+16, BSG14], computer vision is applied in this research to identify labels that appear on the participants' images. For instance, a beauty pageant's entry image may be labeled with meta data, such as "Beauty", "Photo Shoot", "Smile" and "Blonde". Similarly, a design contest entry might have topic labels, such as "Landmark" or "Architecture". Figure 4 displays an example, where the Google Vision API was used to extract labels from images which were used in contests of Choicely.

Combining the identified labels and the vote data can provide information on user behavior. For instance it can be identified which demographic group of users like what kind of content, or the behavioral differences between two or more groups can be compared to eachother. Furthermore, this kind of information could be used to recommend new content in the platform to users which they have not seen before. Last but not least, it could be identified that which traits of participants contribute to more votes and engagement by users or certain user groups.

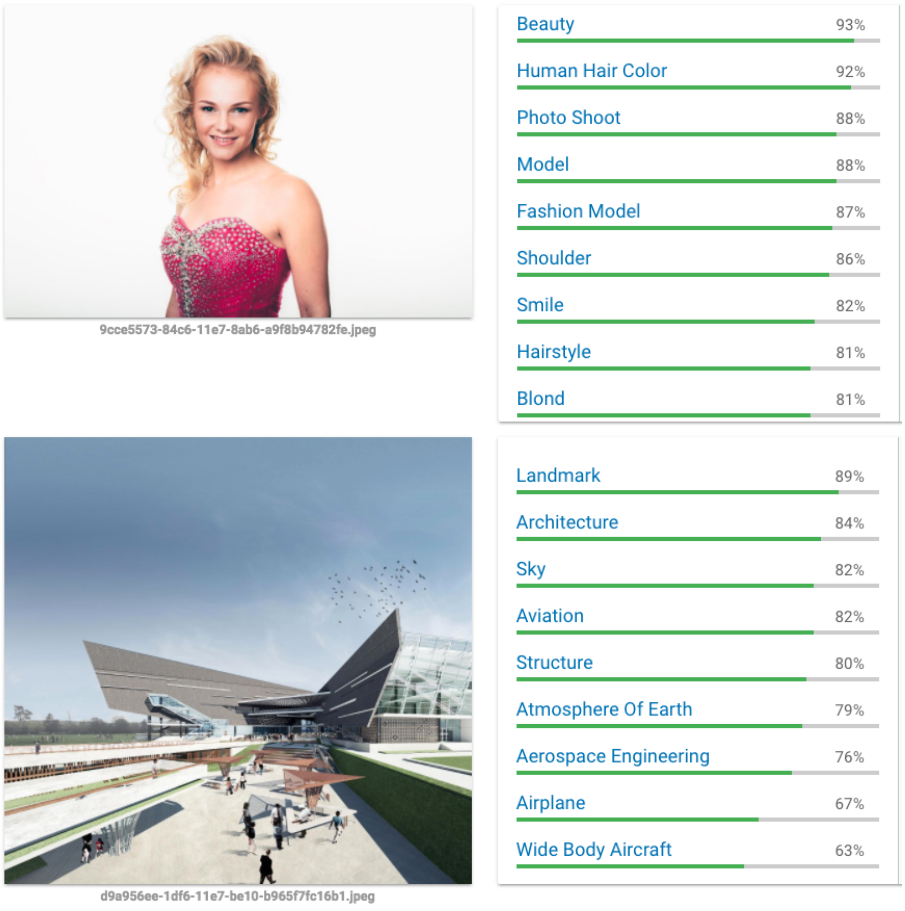


Figure 4: The labels identified by the Google Vision API on two of the contest participant's images.

4.2 The voting mechanism and the user profiles

Various voting options are available for Choicely contests. The author of the contest has the choice of setting a limit on how many votes users can spend on individual participants or the whole contest in overall. For instance, if the maximum votes in the contest is set to 1, users can give exactly a single vote on one and only one participant. Configuration settings allow infinite votes as well. In this case, users may vote on all of the participants as many times as they like. Removing votes is possible, if the author has decided to enable this possibility. Votes cannot be modified after the contest has ended. Each contest has its own voting configuration.

On top of the regular free votes, contest authors may allow users to earn more votes (called "silver votes") by sharing the contest on social media or by watching adverts. Furthermore, contest authors can allow users to purchase more votes (called

"star votes" or "gold votes") with exactly the same restriction settings as explained above. Note, however that the configuration for the three vote types are distinct for every contest. This means that the limitation on free/silver/star votes may differ for individual participants as well as the whole contest. For instance, a contest author may allow users to spend only 5 votes for free, but unlimited number of silver and gold votes in a contest.

In this research, there are no distinctions made between the different vote types. That is, free, silver and star votes are considered as equal. The rest of this study makes no difference between the different vote types, but simply considers them as votes on contest participants, which is a sign of favor shown by the user towards the contest participants. Nevertheless, it would be interesting to separately study if there is any significant difference in how users spend different types of votes.

Each user profile contains the features listed in Table 2. For the purposes of this research, the fields written in italic text in the table below are considered interesting part of user profiles. The three highlighted attributes are all categorical variables, that contain demographic information about users. By filtering the data using these demographic attributes, one can obtain data generated by targeted segments of users and compare those datasets. This way the behavior of the demographic groups can be studied, which provides answer to RQ3. To keep the scope of this research compact, only the country information of the users' location is utilized in this study (the state and city are not used).

Field	**Type**
Full name	Free text
Profile picture	Image
Cover image	Image
Gender	Male/Female/Other/Not chosen
Location	Country, state and city
Birthday	Datetime
Age group	0-17/18-24/25-34/35-44/45-54/55-64/65+/Unknown
Introduction/Bio	Free text

Table 2: The list of fields and their types for each user profile.

The demographic data is filled by the users at the time of signing up for the service. Contests require users to have a user profile in the Choicely platform. Choicely offers authentication through social media (Facebook and Google+) as a convenient option for users to sign up with one click. In this case, the social media platform

provides information about the user's profile to Choicely, which allows the automatic population of the demographic data of the user. Optionally, user profiles in Choicely can be created through a regular sign-up process, where users pick a user name as their identities. In such case, their demographic data is unknown by default and it is up to the users to complete their profiles.

In sum, the user data in Choicely consists of the user profiles, which contain demographic information about the users, namely age group, gender and location. On top of that, contests have a number of participants with arbitrary number of votes that the users have already casted. The latter kind of data can be seen as digital footprints generated by users in the Choicely platform. The combination of these datasets sums up to the user data as introduced in the previous chapter and Figure 3 in this particular case. This is further supported with the meta data of contests and contest participants, which was explained in the previous subchapter.

4.3 Data structure and retrieval

In order to better understand the data, the data structure and architecture of the Choicely platform is explained briefly in this subchapter. Most of the platform's data is stored currently in the Google Cloud Platform[4], more specifically in Google Datastore[5] and Google BigQuery[6]. Google Vision[7] is utilized as a computer vision service to gather the meta data for the contestants' images.

Google Datastore is highly scalable document database which is built on top of NoSQL technology [LLC17b]. By providing flexible storage, performant computing resources, encryption possibilities and high availability, the service can serve wide range of applications and various type of business data of companies [LLC17b]. Google BigQuery is a data warehouse for enterprise purposes, large-scale data storage, processing and analysis [LLC17c]. Most of Choicely's data is stored in Google Datastore, however the computer-vision identified tags of the participants' images (Figure 4) are located in Google BigQuery at the present time.

The data is structured into entities in both Datastore and BigQuery. There is parent-child connection between entities, which are key for the retrieval of some of the unique entities. For instance, contests belong to Users or Brands (as contest

[4] `https://cloud.google.com`
[5] `https://cloud.google.com/datastore`
[6] `https://cloud.google.com/bigquery`
[7] `https://cloud.google.com/vision/`

organizers), which are separate entities in the platform. Certain entities can be identified by multiple parent entities. Entities that have no parent(s) can be identified with their unique identifiers and indexed with some other fields. For example, users can be filtered by gender and age group, or unique identifier.

The data is generated by contest organizers, who create contests in the platform and users, who cast votes on their favorite participant(s) in the contests through one of the interfaces explained in Figure 2. The contest participants' images are identified automatically via the combination of Google Vision[8] and Google Cloud Functions[9] once they are uploaded to the database. Google Vision is another component of the Google Cloud, which provides powerful image analysis solutions for software developers [LLC17a]. In this study, the usage of this component is limited to the classification of the content on the images as explained above. Google Cloud Functions are used to be able to automatically assign the labels to the images so that the manual work can be reduced to the least minimum.

Similarly to SQL join statements, the aforementioned identifiers and the parent-child relationship can be used to join list of various entities together. This property is used to aggregate and prepare data for the purposes of the analyses to be performed. Accordingly, the Contest, User, ContestVote, Vote and ImageLabel entities are joined together via their connections to construct the data structure displayed in Table 3. This data structure establishes the basis of the Association Analysis, which is performed to address RQ3.

Figure 5 displays the architecture of the Choicely platform on a high level. The Users of the system use the Choicely clients on their mobile or personal computer to create new contests or cast votes in the existing ones. The Vote and Contest data is collected and stored to the databases that are hosted by Google Datastore. The data analysis is then performed by the Data Analysis framework, which is used by either a data analyst or an automatic service. This thesis work is devoted to establish the core functionality of this framework through this thesis work. Finally, Knowledge Discovery and Data Visualization is performed on the output of the framework.

[8] `https://cloud.google.com/vision/`

[9] `https://cloud.google.com/functions/`

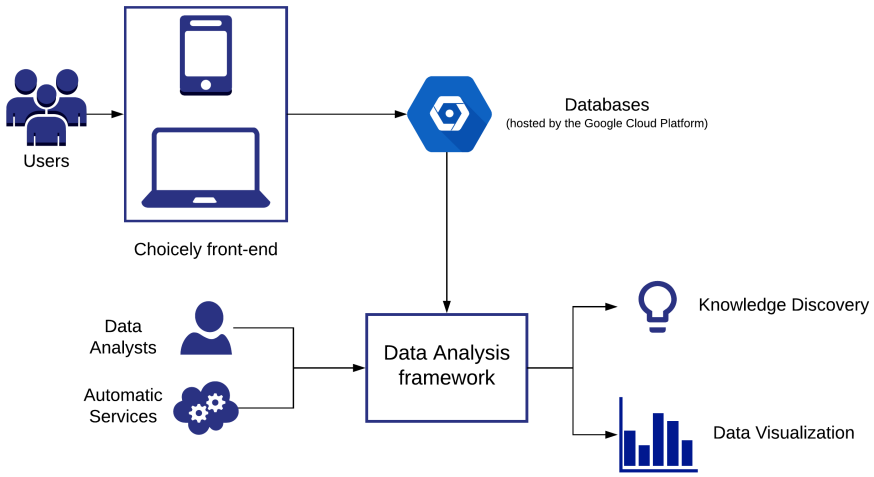


Figure 5: The brief architectural overview of the Choicely platform.

4.4 Applying the chosen methods

To answer RQ2 and RQ3 (presented in Chapter 1.2), the chosen methods presented in Chapter 2 are applied in relation to Choicely's data. The paragraphs to follow elaborate a more in depth, what kind of data is used in case of the Choicely to answer the stated questions. On top of that it is explained, how the data was transformed in order to achieve the results. Data records collected before 1^{st} January, 2018 are used for the analyses.

The EDA fundamentally focused on studying the user engagement in contests and over the contest categories. In other words, the EDA is performed on the data extracted from the Contest entities in Choicely. This choice is done with the aim of addressing the research space of RQ2.

Contests have a few attributes which can help answering RQ2. To gain an understanding on how many users are typically engaged in contests, the number of unique voters is studied in comparison to the number of contests. The number of unique voters means the users, who have voted on at least one participant in a contest. This value can be retrieved for all contests in the platform. Contests can be filtered by their categories which is another approach towards finding answers to RQ2. Similarly to studying the number of unique voters over contests, the same metric is applied to all contest categories. By looking into how many users have voted in which kind of contests, one can get a grasp on the kind of contests that tend to attract more users.

To answer RQ3, Association Analysis using the Apriori algorithm is performed. In order to be able to execute the analysis, some preprocessing has to be done. The list of labels extracted from the participants in the contest are listed for each transaction alongside with the voters' demographic data. In other words, the data has to be transformed such that the input contains rows of the voters' demographic attributes and the computer-vision identified labels from the participants' images. In order to achieve this, the User, Contest, ImageLabel and Vote tables are joined together. The retrieved data is combined as shown in Table 3. One challenge in constructing this data was that these pieces are located in different storage places (Google Datastore and Google BigQuery), which often happens at companies in software development field.

Gender	**Age group**	**Country**	**Labels**
female	25-34	fin	['fashion model', 'hair', 'model', 'beauty', '...]
male	18-24	fin	['dark hair', 'hair', 'model', 'smile', '...]
male	65+	fin	['fashion model', 'hair', 'shoulder', 'beauty', '...]
female	18-24	swe	['fashion model', 'hair', 'model', 'beauty', ...]
male	18-24	hun	['beauty', 'blond', 'human hair color', 'model', ...]

Table 3: The format of the data used for Association Analysis (the records dispalyed in the table are only examples).

After transforming the data, the Apriori algorithm is applied to calculate the itemset supports. This can be done on arbitrary amount of vote transactions as long as they have been transformed to the desired format. However, there are certain cases when it makes sense to perform the analysis. For instance, the analysis can be performed

1. on transactions of a single contest: this way contest organizers get a chance to analyze their engaged audience and seek the tendencies in their behavior,
2. on transactions extracted from the complete dataset: this way one can attempt to find patterns in how users behave on the system-level,
3. on transactions of a targeted group of users: this approach allows to understand the preferences of a desired group of people across multiple contests,
4. on transactions of a single user: this way the behavioral patterns of the individual user can be studied and analyzed.

It would be interesting to take a look into all of these aspects, however the scope of this thesis work limits the analysis to be performed. On top of that, items 3 and 4

might go into "too personal" directions, hence interfere with privacy issues and raise ethical concerns. In order to keep the scope of the analysis on a reasonable level, items 1 and 2 in the above list are taken in the analysis and the rest are discarded.

Using the approach explained above, the support of the itemsets extracted from the labels can be calculated. The support in this case tells the percentage of votes, which were casted on an itemset of discussion. For instance, *S({"beauty", "blonde"}) = 0.6* would mean that 60 % of all votes were casted on participants whose images contained the *{"beauty", "blonde"}* itemset. Likewise, if the list of vote transactions is filtered by one of the demographic attributes, the same can be said for the investigated group of users. The itemsets in this analysis are all extracted by first stating conditions for the value of the demographic group (i.e. gender is male), which puts a condition next to the value. Note that many researches in the field of Data Mining use the term confidence for such conditional associations. In this reserach, the term support is used for this purpose and context consistently.

To ensure that the algorithm's runtime is not too long, the *minsup* value is set to 0.05, which corresponds to 5 % support. This means that all itemsets, that have less support than this value for all demographic groups, are pruned during the progression. On top of reducing runtime, this threshold is set to eliminate less interesting itemsets and keep only the interesting ones on the output.

The result of the Apriori algorithm then can be collected into a table displayed on Table 4. This table shows the itemsets as rows and the targeted group of users as columns. The values in each columns correspond to the support values of the itemsets for the given group of users in the set of transactions that was fed into the algorithm.

X	*S(X|gender=male)*	*S(X|gender=female)*	*S(X|gender=other)*	*S(X|gender=not_specified)*
{"beauty", "blonde"}	0.6	0.8	0.15	0.25
{"cat", "fluffy", "cute" }	0.6	0.1	0.5	0.0
{"sea", "photo shoot", "model"}	0.0	1.0	0.33	0.65
{"building", "architecture"}	1.0	0.0	0.25	0.0
{"building"}	1.0	0.0	0.33	0.0

Table 4: The output format of the Association Analysis comparing genders, where X is the itemset and S is the support (the records dispalyed in the table are only examples).

On top of the support values, lift is calculated and studied for the extracted itemsets. The goal of these calculations are to study attraction towards itemsets in relation to their presence in the contest. Despite the fact that lift is based on the support

values, it helps studying not only the frequent itemsets, but also the improvement gained by the associations. This helps to compare differences between demographic groups and to better understand the attractiveness of itemsets.

The lift values can be collected into a table displayed on Table 5. Similarly as the itemset supports were gathered, this table shows the itemsets as rows and the targeted group of users as columns. The values in each columns correspond to the lift values of the itemsets for the given group of users in the set of transactions that was fed into the algorithm.

X	*L(X|gender=male)*	*L(X|gender=female)*	*L(X|gender=other)*	*L(X|gender=not_specified)*
{"beauty", "blonde"}	2.0	2.5	0.25	0.33
{"cat", "fluffy", "cute" }	1.25	0.05	1.0	0.0
{"sea", "photo shoot", "model"}	0.0	3.0	1.0	2.0
{"building", "architecture"}	4.0	0.0	1.0	0.0
{"building"}	3.0	0.0	1.0	0.0

Table 5: The output format of the Association Analysis comparing genders, where X is the itemset and L is the lift (the records dispalyed in the table are only examples).

In this imaginary example, the lift value *L({"building"}|gender=male) = 2.0* suggests, that the support for "building" among male voters is two times as probable than "building" appearing on a randomly chosen contestant's image from the studied dataset. If the lift value is 1.0, then the probabilty of finding the itemset in the targeted group's vote transactions is the same as on the images in the contest. In other words this means, that the group shows no more and no less attraction towards this content than what is already available in the data. This case may occur for instance, when the same label appears on every image in the contest (its support is 1.0 for all transactions). Hence, this case is not considered interesting. Finally, lift value less than 1.0 suggests less attraction for the itemset, which can also suggest interesting findings.

In order to better understand the results of the Association Analysis, first a single contest (Miss Suomi 2017[10]) is looked at more carefully. By looking at a single case study, the results can be analyzed and interpreted in a smaller scope and hence be understood easier. Secondly, the results of the analysis performed on the system-level are presented and interpreted briefly. This approach is chosen to demonstrate the capabilities and the potential of the technique in both smaller and larger datasets.

[10] `https://choicely.com/contest/164f52c7-9df8-11e7-b3c9-d1a0f88250ad`

The Miss Suomi 2017 contest has been hosted in the Choicely platform late September, 2017. The data used for the analysis in this contest consists of 176 votes from 73 users in overall. There are 10 participants in the contest, who were photographed while dressed in white cocktail dresses on a shore of a lake or sea. Each photo has one and only one participant. Each participant has only one and only one photo in this contest.

Depending on how many individual participants the chosen transactions target, the number of itemsets (hence the number of rows) grows rapidly. Also the number of demographic groups (hence the number of columns) may differ depending on which the property is used from the users' profile. Essentially, the number of categorical groups for genders can go up to 4, for age groups up to 7 (see Table 2), but many more for location-based groups. As a conclusion it can be said that despite the supports are obtained, most likely their number will be too overwhelming to look at and analyze. Furthermore, one might assume that there patterns and structural trends in the support values, for instance some itemsets may share similarities, while others may differ from eachother in terms of their support values.

For this reason, the Co-Clustering method is used to identify if any itemsets or demographic groups show similarities in the voting data. The method is applied on the output of the Association Analysis displayed in Table 4. More precisely, only the matrix of support values is taken as the input of the Co-Clustering method, which is an matrix of real values on the range of *[0.0, 1.0]*.

A simplified example output of the method is displayed in Table 6. As shown, the support values in the cells with the same background color show similarities in their values. For instance, the support values in the green cluster are all 1.0, while the red cluster contains values that are also somewhat similar. This suggests that these itemsets and demographic groups are similar in nature. It can be seen, that this representation adds a layer on top of the output of the previous Association Analysis step (Table 4) by reorganzing and clustering the rows and columns of the data matrix. By applying Co-Clustering, the closely-related itemsets and the groups are clustered together and grouped closeby, which facilitates the interpretation of the results.

X	*S(X/gender=male)*	*S(X/gender=female)*	*S(X/gender=other)*	*S(X/gender=not_specified)*
{"beauty", "blonde"}	1.0	0.00	0.05	0.0
{"cat", "fluffy", "cute" }	0.0	0.75	0.70	0.0
{"sea", "photo shoot", "model"}	0.0	0.66	0.90	1.0
{"building", "architecture"}	0.0	0.00	0.10	1.0
{"building"}	0.5	0.00	0.00	1.0

Table 6: An example output of the Co-Clustering algorithm itemset supports by comparing genders for $k = 4$ clusters. X is the itemset and S is the support. The clusters are highlighted with different colors (the records dispalyed in the table are only examples).

The implementation is done with the help of the bicluster module [sld17] of the popular scikit learn package [PVG+11]. The Spectral Co-Clustering class is used for the analysis, which proposes to pinpoint biclusters with higher values than others according to the documentation [sld17]. The algorithm treats the input matrix as a bipartite graph, in which the rows and columns are nodes and the values in the cells are the weighted edges between them [sld17]. This logic suits the input matrix, as the higher support values in the cells represent stronger attraction between the itemsets and the demographic groups.

It is important to highlight in advance, that the chosen algorithm assigns one column and row exactly to a single cluster. As a result, the yielding clusters do not overlap and cells can not belong to more clusters. The reason behind choosing this particular algorithm lies in its simplicity and the applicability to the problem. This can be seen as a limitation to this particular algoritm, nevertheless the potential viability in this setting of the approach can be discovered and analyzed.

Using this approach, the clusters that are similar are going to be organized close to eachother. The algorithm's input requires a k value, which corresponds to the number of clusters to seek. The cluster labels are attached to each row and column respectively, which suggests the itemsets and demographic groups that belong to the same cluster. In the performed analyses, the k value of the Co-Clustering algorithm is set to 4.

By knowing the clusters, one can conclude if there are tendencies among demographic groups in terms of behavior. For instance, if the age group 65+ and 0-17 would be assigned in the same cluster by the algorithm, that would prove the similar behavior of the users in these two groups. In other words this would mean, that the certain kind of content leads to the engagement of both of these groups, but not necessarily the others. This can be valuable information to marketing professionals

in the situation, when they wish to target a specific group of discussion. The demographic traits could be further combined in theory (e.g. male teenager users from Finland), however this deep analysis is out of the scope of this thesis work.

5 Results

This chapter presents the results achieved via applying the chosen methods introduced in Chapter 2. The results obtained via each method are listed under the following subchapters respectively. This chapter provides answers to RQ2 and RQ3 (Chapter 1).

5.1 Exploratory Data Analysis

By the end of 2017, there is a total number of 573 contests in the platform. To identify what kind of content is more engaging (RQ2), first let us look at the amount of unique voters over the number of contests. Table 7 lists the key figures of the number of unique voters over contests in the platform.

It can be easily seen that most of the contests engage very small amount of users, as the median of the unique voter count for all contests is 3. This means that at least half of the contests have had only 3 users who actually voted for any of the participants. One of the reasons behind this is that the company did not establish a large user base yet. Therefore there are many users who created only one contest but never used the platform on the long run. Many of the contests serve only testing purposes, hence engage only a few users. Such records can create bias in the upcoming analyses, because their data does not represent realistic scenarios.

Measure	**Value**
Mean	447
Standard deviation	2992
Min	0
25th percentile	1
Median (50th percentile)	3
75th percentile	22
Max	54684

Table 7: The basic statistical measures of unique voters over contests.

For this reason, contests with less than 100 unique voters are excluded in the remainder of the analysis, because such observations are not representative. This dataset contains 166808 vote transactions by 145000 users over 81 contests, 1113 contest participants and 432 labels on the images recognized by Google Vision. For the remainder of the EDA, this filtered dataset is used.

Figure 6 displays the same distribution for the filtered set of contests. In this figure contests with more voters are more apparent. The highest number of unique voters is close to 55 000 in one of the contests, the mean value ($\mu = 2525$) and the standard deviation ($\sigma = 6804$). These numbers mean that there is a large variance in the amount of engaged users in contests. From the data it cannot be clearly said which traits make a contest more attractive to users. Presumably the marketing activities done by the contest's organizers have strong impact on the size of the engaged audience. For instance news agencies have an established customer base already, who was supposedly targeted by these contests via the web widget provided by Choicely (Chapter 1.3).

The boxplot on the right side of the figure uses the 95 percentile (around 5200 unique voters), above which the outliers can be seen. It can be also seen that the most of the contests engage 260-2600 unique voters (as described by the first and the third quartiles). There are 6 large contests, from which the biggest have engaged 54684 voters.

The six large contests are worth investigating a bit more closely. Four[11,12,13,14] out of the six large contests were beauty pageants, labeled with the categories of "beauty", "fashion" and "entertainment". The two other contests[15,16] are listed only in the "other" category, which is certainly a mistake. By looking at the latter two contests, it can be easily seen that they would better belong to the "entertainment" and "sports" category.

In all six of the large contests, the contest participants were people: either sportsmen, celebrities or beauty queens/kings. It is interesting that none of the large contests have had objects, places or other intangibles as contest participants, although the platform has seen many of such participants previously.

[11] https://choicely.com/contest/5ca98554-0f7d-11e7-9f0c-6f102a54d68d
[12] https://choicely.com/contest/fb112461-9000-11e6-9e28-87ebd7a21d0d
[13] https://choicely.com/contest/7425566e-8c8e-11e6-b8ce-2147b021362f
[14] https://choicely.com/contest/164f52c7-9df8-11e7-b3c9-d1a0f88250ad
[15] https://choicely.com/contest/50819173-f838-11e6-b171-b949f18a4d21
[16] https://choicely.com/contest/4257ea9c-3e21-11e7-84ec-5f5a9bcfd190

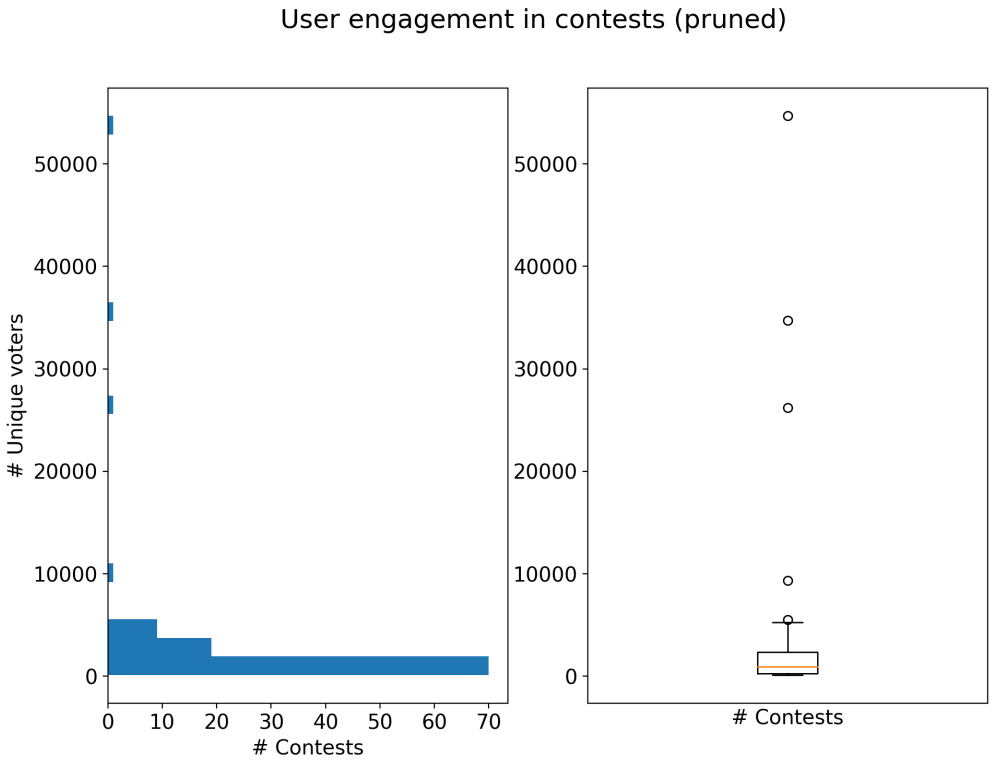


Figure 6: The number of unique voters over contests after filtering out contests with less than 100 unique voters.

In the next step, let us look at the distribution of contest categories in the filtered set of contests. It can be seen from the histogram on Figure 7, that the amount of "beauty", "entertainment", "sport" and "fashion" contests is considerably high compared to the rest of the categories. This finding is well aligned with the case company's profile at the point of conducting this study. As pointed out in Chapter 1.3, most of the company's customer base consits of Finnish broadcasters and advertisers. Hence there is no surprise in the distribution of the categories except for the "other" group.

By manually looking at contests in the "other" category it can be seen, that many (28 out of 34) of these contests are actually a sport-related. To name a few examples, there are contests with titles such as "Best player poll" ("Paras pelejaa aanestys")[17], "Who is the hottest driver?" ("Kuka on kuumin kuski?")[18], "Fastest driver of the

[17] `https://choicely.com/contest/5f8f8470-914d-11e6-bd5b-e571d894172f`
[18] `https://choicely.com/contest/93d5f89c-5676-11e7-8cf7-0759c198269e`

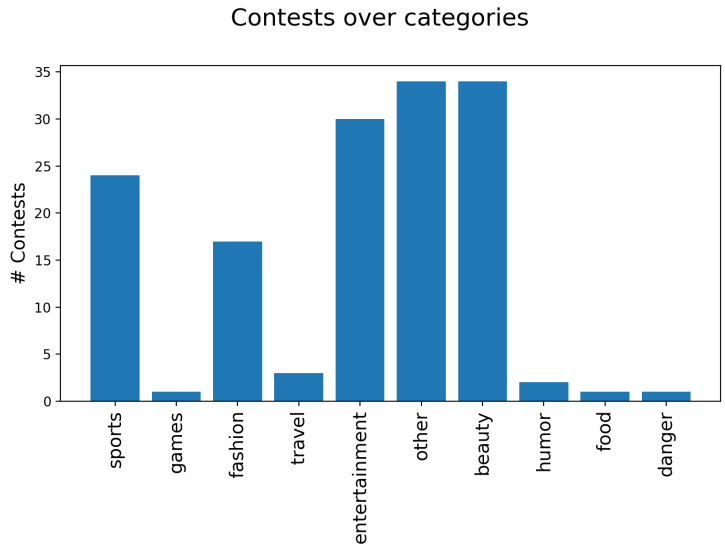


Figure 7: The number of contests in each category.

race" ("Kisan nopein kuski")[19]. This finding is not a suprise knowing the firm's customer base, but it also suggests the popularity of sports contests. If these contests were labeled correctly, sports contest would top the charts (Figure 7) with the highest bin around 50. Another interesting fact is, that in all of the sports contests, participants are athletes, hence the images contain human beings in these contests as well.

Three of the contests were related to the "travel" category, one to "fashion" and "beauty" and one to "entertainment". The error in this case is not as high as with sport contests, however correcting these category labels would facilitate data analysis in the future. For this reason, it is suggested to the company to review such issues, correct them manually and potentially prevent them happening in the future.

To answer the question of most engaging contest categories, the number of unique voters over contest categories is studied. Simple statistical measures (sum, mean, median and standard deviation) are calculated for the number of unique voters for each contest category. Table 8 displays the results.

It can be seen, that "entertainment" and "beauty" contests cover the majority the amount of unique voters ($\approx 66.60\%$ of the total). The values of these categories together are similar with the "fashion" category. In these categories the mean and the

[19] `https://choicely.com/contest/bb7db707-3683-11e7-9f48-cbe34704f83e`

Contest category	Sum of unique voters	Mean of unique voters	Median of unique voters	Standard deviation of unique voters
beauty	135866	3996	1482	9854
other	61455	1807	1240	1831
entertainment	161233	5374	1728	11772
sports	15220	634	299	757
fashion	75872	4742	874	12972
travel	4451	1483	409	1719
humor	1367	683	683	274
food	767	767	767	0
danger	958	958	958	0
games	164	164	164	0

Table 8: The basic statistical measures of unique voters for each contest category.

median of the voters is also considerably high, which suggests their attractiveness. However, the high values in the standard deviation of the unique voters indicate that values are widely spread around the mean. The relevance of this observation is, that not all contests in these categories engage a large audience necessarily. Thus it can be concluded, that these categories tend to appear together and also attract a larger audience compared to the rest in general.

The categories "travel", "humor", "food", "danger" and "games" have hosted only a considerably low number of contests (Figure 7). Due to the small amount of data, it is difficult to derive any relevant results about the attractiveness of these categories at this point.

The above results contribute towards answering RQ2. The results allow to derive the following conclusions:

- many of the contests are not labeled and hence belong only to the "other" category - fixing these labels manually could contribute towards better results,
- the "fashion", "beauty" and "entertainment" categories often appear together in contests,
- contests in the "beauty" and "entertainment" categories appear to be engaging to large audiences,
- contests where participants are human beings appear to be attractive to users,
- the platform has hosted only a few contests in some of the categories and hence it is not possible to derive significant findings about the attractiveness of those contest at the moment.

5.2 Association Analysis

The results of the Association Analysis are presented in this chapter are limited to the 1-itemsets and 2-itemsets in order to keep the findings easily understandable. The following figures and paragraphs display and explain the results. The itemsets in the figures are ordered by the variance of the values over the bins (however the variances are not displayed in the figures). This way the itemset with the highest variance is on the top, while the itemset with the smallest variance is on the bottom of the figure. All of figures in this chapter follow this convention of ordering. First the chosen Miss Suomi 2017 contest is analyzed, followed by the system-level analysis.

Figure 8 displays the 1-itemset supports and lifts for genders. The first finding to note is the *{"beauty"}* and *{"dress"}* itemsets with support and lift values 1.0 for all genders on the bottom of the figures. These values suggest that every single vote transaction in this contest has contained these itemsets. The reason behind this observation is that all of the participants' images had these labels. Therefore, inevitably all of the vote transactions picked up these itemsets and every vote in the contest has full support towards them. Such values do not show any significant meaning and therefore are excluded from the rest of the figures.

It is interesting to investigate, why all of the images have these two labels. As the participants' images were shot in a similar environment, with considerably similar content (e.g. the white dresses, the background, the hair styles, only one model in the image etc.), the Google Vision API assigned the same labels to the images. This gives a proof that computer vision in this case is more targeted towards getting an overall idea about the content of an image, rather than identifying its specific traits and attributes.

Figure 9 provides a good possibility also to compare gender differences. For instance, the support *S(X={"model"}|gender=female)=0.45* is considerably higher than its male counterpart *S(X={"model"}|gender=male) = 0.26*. The lift values in this instance are *L(X={"model"}|gender=female) = 2.25* and *L(X={"model"}|gender = male) = 1.30*. Similar observation can be made for the *{"fashion model"}* itemset, which suggests correlation between these two labels. This suggests two findings: the label "(fashion) model" appears to be more attractive to female voters compared to males and the presence of model in an image makes participants more appealing to all genders, but especially females.

Similarly, the *{"photo shoot"}, {"sea"}, {"gown"}* and *{"wedding"}* itemsets have

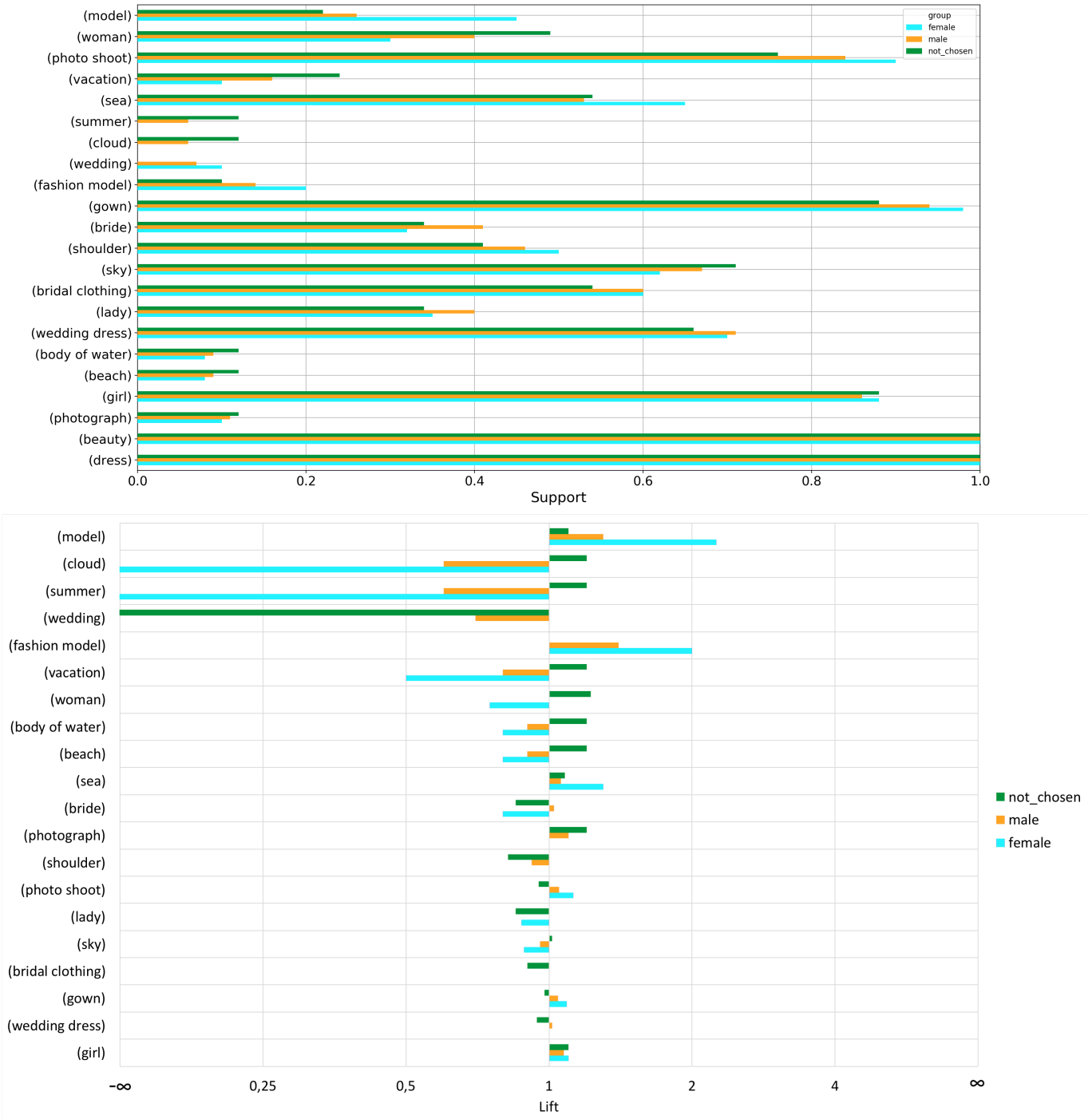


Figure 8: The 1-itemset supports (top) and lifts (bottom) in the Miss Suomi 2017 contest by genders (all itemsets are displayed ordered by the variance of the values).

higher support in case of female voters. This may suggest that females are more likely to vote on images, where the object (the model and the dress) is put more into focus. The lift values however reveal, that the interest for *{"sea"}, {"gown"}* and *{"photo shoot"}* are higher for all groups, which proposes that the number of votes increases when sea is in the background.

Another curious thing to note is the lift values of the *{"summer", "cloud", "beach"}* itemsets, which are lower than 1.0 for both males and females. This observarion may suggest that having these in the background of the image is less attractive compared to the sea as background. Furthermore, *{"wedding"}* has lift value less than 1.0 for genders other than females, which suggests less interest for this topic.

Interestingly, the support and the lift for the *{"cloud"}* and *{"summer"}* itemsets are 0.0 for females, while the values of the male group are slightly higher. This means that no female users have voted on images with these labels (unless they belong to the *not_chosen* gender). This allows the conclusion that males might prefer images with more light and more colorful background in their votes. The lift and the support for *{"bride"}* and *{"lady"}* itemsets for males are also higher, which can mean that the model being a girl has more impact on their votes than her dress or the surroundings. However, as the lift value is not topping the charts for these itemsets, which suggests no major impact in users' behavior.

Next, the 1-itemsets are calculated similarly in the same contest for age groups. Figure 9 displays results. It is interesting to notice that age groups 45-54 and 65+ have considerably higher support than the rest of the groups in case of the *{"shoulder"}, {"bridal clothing"}* and the *{"wedding dress"}* itemsets. The lift values for these itemsets are also higher for these groups which suggests their attractiveness. It seems, that the open-sholder wedding dresses displayed on the pictures raise the engagement of these groups. This may indicate also that dresses become influential in favoring a model in this kind of a contest for these groups. In other words, dress may be a more important aspect than the model herself.

Another interesting observation is that the values for some itemsets, such as *{"body of water"}, {"beach"}, {"cloud"}, {"vacation"}, {"beach"}* show 0.0 support and lift for the groups above age of 45. At the same time, these itemsets show somewhat higher support and lift by the rest of the age groups (0-17, 18-24, 25-34, 35-44). This finding may indicate that activities such as having vacation on a beach in a warm country is more in the interest of young people, which can be valuable information for travel agencies or marketing specialists. With such valuable information at hand,

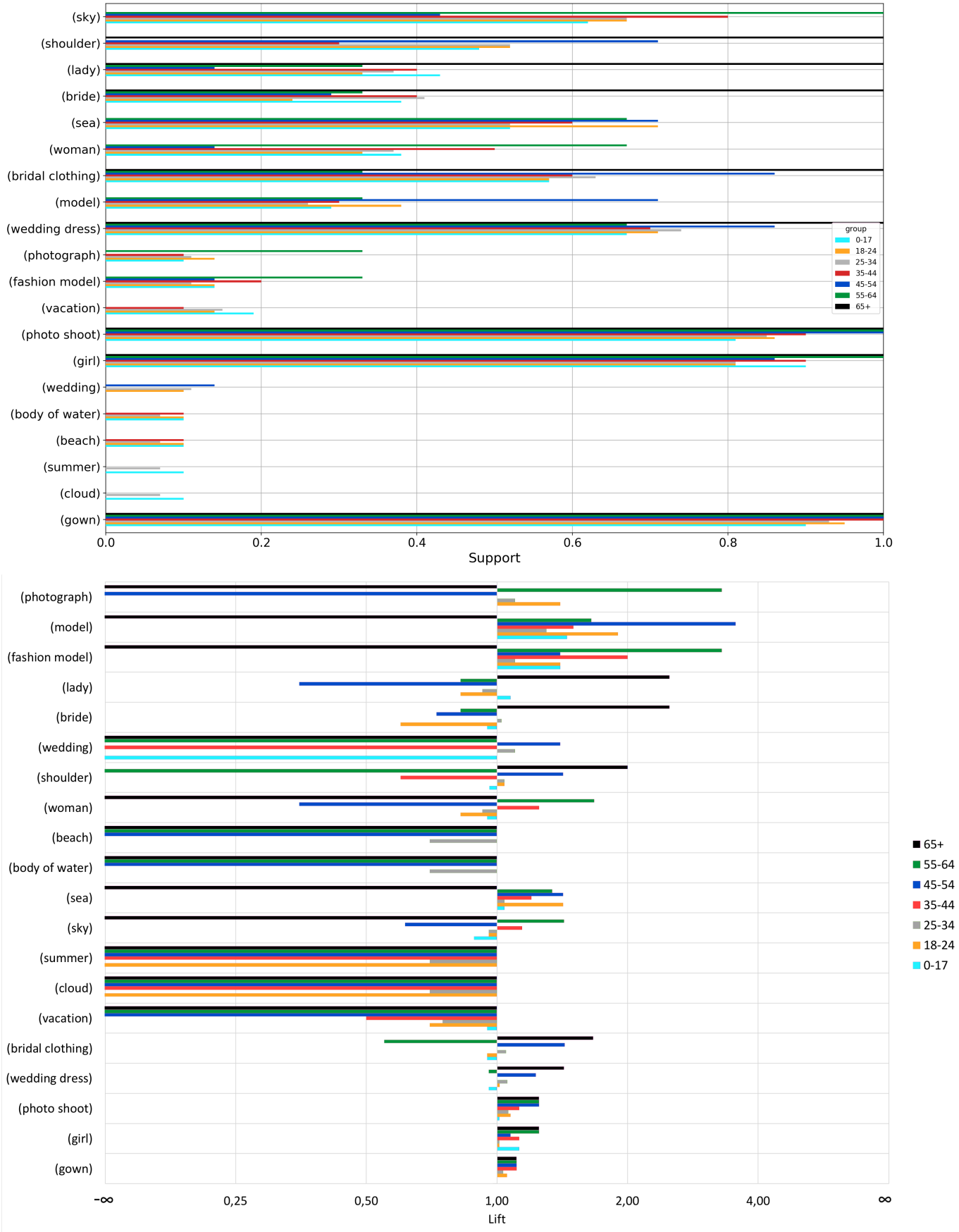


Figure 9: The 1-itemset supports (top) and lifts (bottom) in the Miss Suomi 2017 contest by age groups (only the first 15 itemsets are displayed ordered by the variance of the values).

tourism offices could provide more customized advisory service for customers with different background.

Strong difference can be observed in the lift values of *{"sea"}* for the 65+ age group. The lift value is 0.0 in this case, while all the other groups have values over 1.0. This means that sea did not engage the elderly whatsoever, while triggered the interest the rest of the groups. It can be hence said in general, that sea appears to be an attractive trait in the background for most of the age groups but does not interest voters above 65 years of age. Another reason behind this observation could be that the amount of gathered votes from this age group is low, hence create bias in the results.

Finally, the 1-itemsets supports in the contest are analyzed by the location (country) of the users (Figure 10). The first interesting observation is, that the support and lift of the USA group are 1.0 on some itemsets, while 0.0 on the others and there are no values in between.

By looking at the data the reason becomes obvious: the amount of transactions, where the users' country information is set as USA, Ukraine and Hungary are considerably low (below 5). When looking at the lift values this bias comes even more appealing for users in USA and Ukraine. This may not be a big surprise knowing that this contest was mainly advertised in Finland and not in other countries. As a matter of fact, there were 62 Finnish voters and 21 voters from the Philippines, which are still considerably low compared to the total number of voters.

This finding suggests that many of the voters in this contest did not indicate their country information and hence do not display on Figure 10. In other words, the amount of users, whose location information is filled is too low. Hence, the results of these groups are only the upper and lower extremes, which means bias in the data. It can be also said that this particular contest did not engage users across borders on a broad scale and therefore is not a good case for this kind of analysis.

Encouraging the users in the future to provide more information would allow to derive more relevant results. At the present time, no strong conclusions can be drawn from the location data of the voters in the Miss Suomi 2017 contest. To ensure the relevance of the support and lift calculations, it would be hence a good move to put a lower limit (e.g. 100 observations) on the number of voters, similarly as it was done in the EDA phase for number of unique voters in contests. For all of the above reasons, no more analysis on the country-based support and lift values is performed in the remainder of this study.

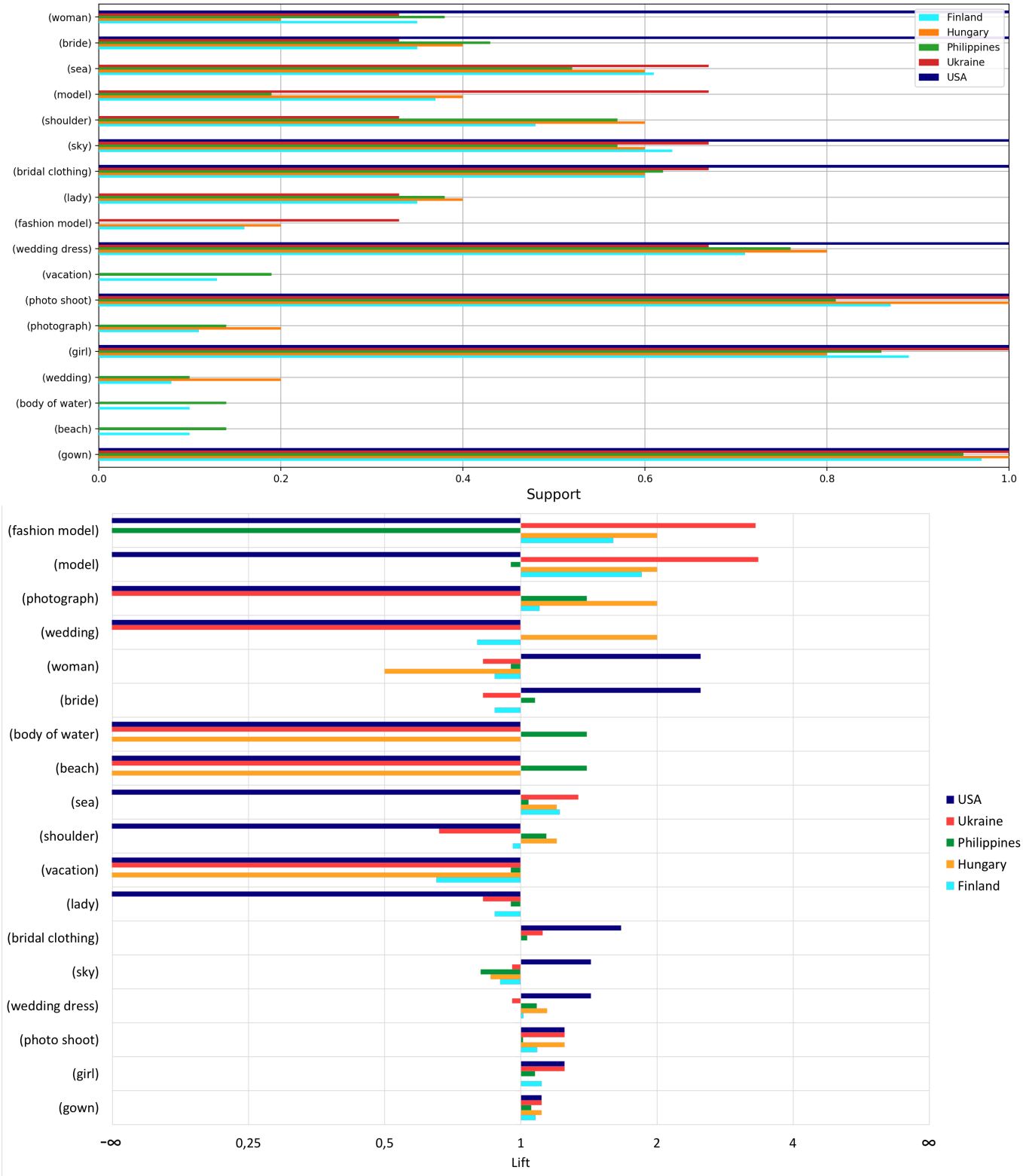


Figure 10: The 1-itemset supports (top) and lifts (bottom) in the Miss Suomi 2017 contest by countries.

The next step of the Association Analysis covers the investigation on k-itemsets, where the k-value is higher than 1. Figure 11 displays the support values for the 2-itemsets by genders. To enhance the readability of the figure, only the first 15 values are shown.

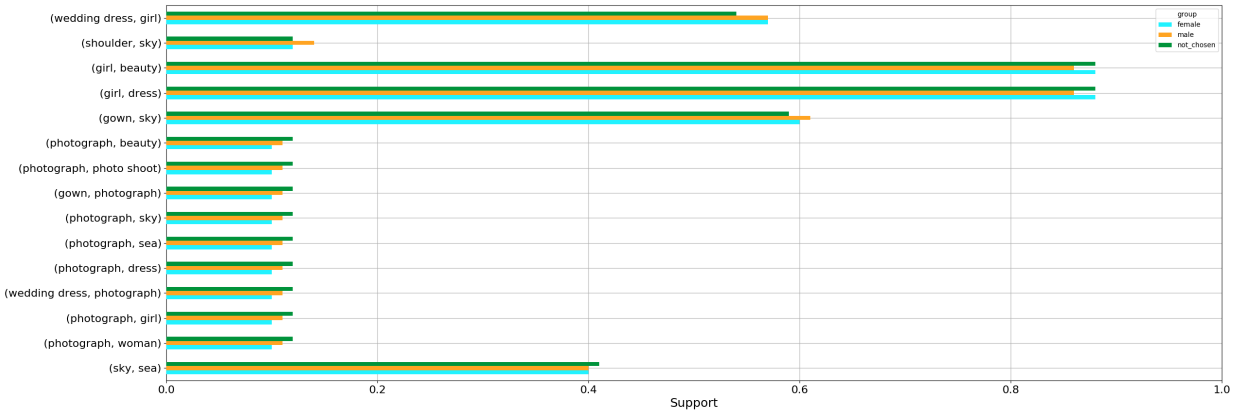


Figure 11: The 2-itemset supports in the Miss Suomi 2017 contest by genders (only the first 15 itemsets are displayed, ordered by the variance of the support values).

An important observation to make is the identicality of the 2-itemsets, where *"photograph"* is present. This is due to the fact that all of the pairing itemset Y (i.e. *{"beauty"}*, *{"sea"}*, *{"photo shoot"}* etc.) are always present on images, where *"photograph"* is present. In other words, the confidence $C(\text{"photograph"}|Y) = 1.0$, where Y is the itemset next to *"photograph"*. When discussing k-itemsets, where $k \geq 2$, this issue may arise and provide identical support values for some combinations. When the complete list of 2-itemsets is plotted, the same observation can be made for many other itemsets, such as *{"bride"}*, *{"shoulder"}*, *{"body of water"}*, etc.

To address this issue, such itemsets are combined and analyzed together. In the above example (Figure 11, itemsets with *{"photograph"}*), the 2-itemsets with identical support values are merged into a single itemset, such that it contains all 1-itemsets respectively, i.e. *{"photograph", "beauty", "photo shoot", "gown", ..., "woman"}* in this case. This way the combined itemset in this case has 10 items in overall. This is done repeatedly for all itemsets, which show the same support values. The lift values are then calculated for the resulting itemsets. For the sake of interestingness, the support values are not displayed here and only the lift values are analyzed. Figure 12 displays the lift values for genders. Similar results for age groups are listed in Appendix 3, however they are not analyzed here.

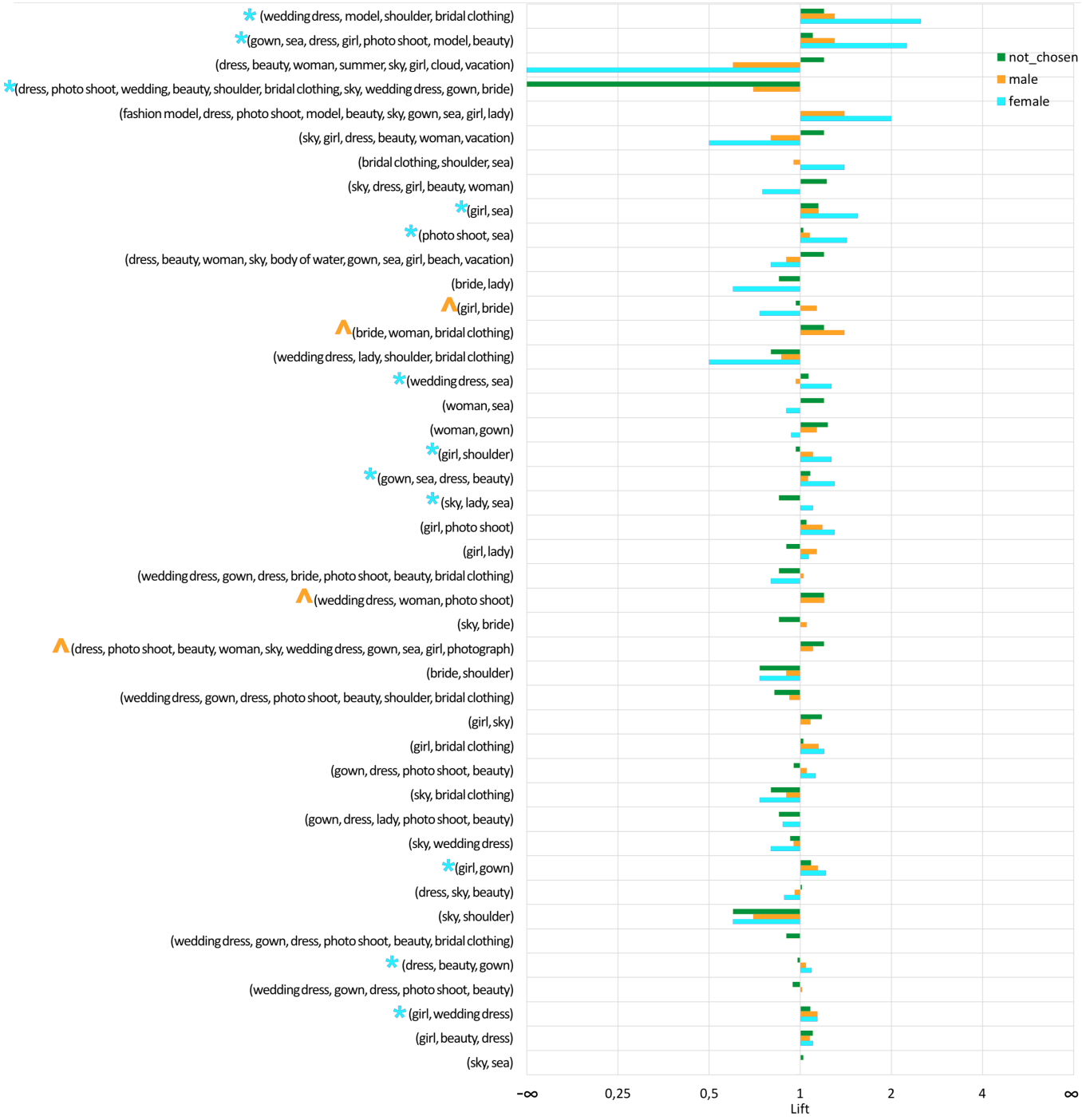


Figure 12: The multi-item itemset lifts in the Miss Suomi 2017 contest by genders (all itemsets are displayed ordered by the variance of the lift values).

The results show that the lift values for females tend to be higher than the other groups in many cases. These suggests high interest towards the topics that appear in the itemsets compared to the rest of the groups. Such itemsets are marked with a blue * mark on the left side of the figure. As the number of items in the itemsets varies, these often share similarities, but they also describe certain topics. For instance, it can be seen that the items in the sets are telling about a lady wearing some sort of wedding/bridal dress. In other words, the items which explain the dress as such are more dominant and appear accross multiple itemsets.

Other itemsets show more attraction from male voters. These are marked with an orange ˆ on the left side of the figure. In contrast, these itemsets describe not the dress as much as the model on the image and the background (sea, sky). It can be hence speculated that female voters might relate to the dress and the model (maybe think about themselves as being the models wearing the dress), while male voters might focus on the setting and the model in the image more.

Another interesting observation is that male voters' lift rarely goes to extremes, but usually is around the value of 1.0 in most cases. In contrast, the lift for the female group goes to extremes in both end for many more itemsets. This may suggest that females follow more of a specialist approach than males, who are more generalists. Another explanation behind this observation can be the fact that the male voters dominate the mean. This assumption is confirmed when the distribution of the genders is looked at: 66 % (n=95) of the voters in this contest were males, while 33 % (n=40) are females and 33 % (n=41) did not provide gender information. Based on these results it seems, that females are strongly engaged towards certain topics in this contest while males are engaged accross multiple topics which appear in the contestants' images.

The final part of the Association Analysis has taken a look at itemset supports and lifts on the system level. That is, all of the vote transactions from contests with 100 unique voters were extracted from the system. This is the same dataset, which was used for the major part of the EDA in Chapter 5.1, containing a total of 166808 votes over 81 contests.

The biases in the data become appearent when the votes are studied on the system level. Most of the itemsets reflect on the great majority of "beauty" and "fashion" category contests, as itemsets such as *{"model"}, {"beauty"}, {"photo shoot"}, {"shoulder"}* etc. From the itemset supports it seems that all of the itemsets were extracted from contests of these two category, as they clearly describe concepts that

appear in such contests. This observation also supports the previous finding, that mainly contests in these categories were hosted in the platform with great success.

While studying the results it was identified, that the *not_ chosen* group has considerably higher support over males and females in many cases. This can be explained by the fact that many users (49.82 %) did not provide their gender information on their profile, which yields in a much higher number of observations (vote transactions) for users whose gender is unknown. 46.4 % (77416 observations) from the the vote transactions were received from users with unknown gender. Likewise, only 27 % (44918 observations) of the votes were received from females and 26.6 % (44407 observations) from males. Similar observations can be made for the age group values, where 67 % of the voters have not filled their age.

It can be therefore seen that the sample size of the *not_ chosen* gender and age group is dominant in the system-wide data. This creates a bias in the support values as the data for these group is more representative than for others. While this problem did not emerge on a single-contest level, when looking at all transactions, it becomes apparent. It could be assumed, that the distribution of the two genders is equal in this group, however there is no proof on this claim. For these reasons, strong claims and concrete findings are harder to derive from this data due to the bias explained above. The system-level data is therefore not analyzed, however some of the extracted figures are displayed in the appendices (Appendix 1 and Appendix 2).

5.3 Statistical significance

An important aspect of evaluating the results is the concern of statistical significance. Calculating the support and lift values in a single contest provides an idea on the engagement measures, however the differences among groups in these measures on their own do not indicate statistical relevance. Stated differently, even if the lift and support values differ for two groups of users, that may not mean that the difference is statistically representative.

For this reason, the paragraphs to follow outline an approach towards evaluating the statistical significance in the comparison of the groups. In other words the following paragraphs present a way which can be used for the hypothesis testing on the behavioral differences of demographic groups in Choicely for itemsets. Some non-trivial challenges are also outlined as part of the discussion on this issue.

The testing approach is introduced thorugh an example from the results of the

Association Analysis. The *{"model"}* itemset is taken as an example. The values to be presented are all obtained using the data gathered in this Miss Suomi 2017 contest. The reason behind choosing this itemset as an example is that it was discussed already in the preceeding paragraphs and the demographic groups have shown interesting differences in this particular case.

First the terms used and the null hypothesis are formulated. Let *f({"model"})* be the relative frequnecy of "model" and *f({"model" | gender = g })* is the relative frequency given that the voter's gender is *g*. Let n be the total number of votes, *n(gender=g)* be the number of votes performed by voters in the g gender. The number of "model" given that gender is male is binomially distributed under the null hypothesis. Using the gathered data, the above values are as follows:

$$
\begin{gathered}
n = 176 \\
n(gender = male) = 95, \\
n(gender = female) = 40, \\
n(gender = not\_chosen) = 41, \\
f(\{"model"\}) = \tfrac{52}{176} = 0.3, \\
f(\{"model"\}|gender = male) = \tfrac{25}{95} = 0.26, \\
f(\{"model"\}|gender = female) = \tfrac{18}{40} = 0.45, \\
f(\{"model"\}|gender = not\_chosen) = \tfrac{9}{41} = 0.21
\end{gathered}
$$

Let the null hypothesis H_0 be that gender g does not differ from others genders in terms of voting behavior, i.e.

$$H_0\text{: } f(\{"model" \mid gender = g \}) = f(\{"model"\}).$$

Let the alternative hypothesis H_1 be that g differs from other genders, i.e.

$$H_1\text{: } f(\{"model" \mid gender = g \}) \neq f(\{"model"\}).$$

Using the values above and the binomial testing method, a two-sided test is performed. The chosen significance level is set to $\alpha = 0.05$. The obtained p-values for all genders are as follows:

$$
\begin{gathered}
p_{male} = 0.50, \\
p_{female} = 0.055, \\
p_{not\_chosen} = 0.30
\end{gathered}
$$

As a conclusion it can be said, that none of the obtained p-values indicate strong enough proof to reject the null hypothesis in every case. The studied groups in this particular case hence do not show statistically significant differences. The female group is however fairly close to the significance region. To provide a visual proof on this observation, Figure 13 displays the p-values over the number of votes on this itemset for females. It can be seen that the p-value at n_{female}=*18* is just above the chosen level of significance.

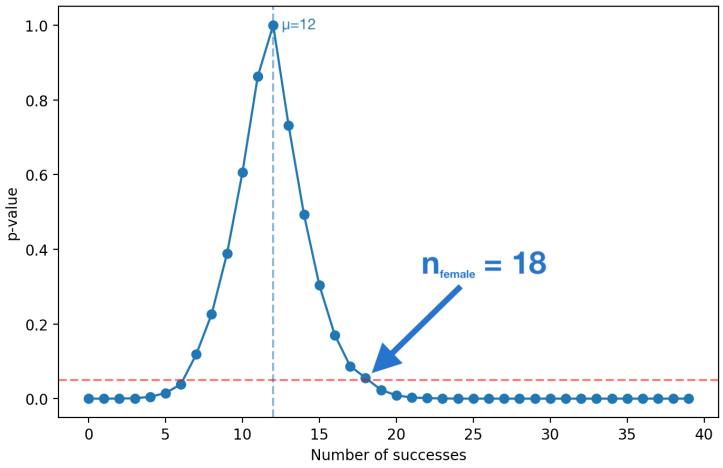


Figure 13: The p-values over the number of successes (votes) for the *{"model"}* itemset in the Miss Suomi 2017 contest for the female group.

The statistical significance evaluation faces further challenges and possibilities. It would be interesting to compare for example results of multiple contests. Due to the fact that contests may have different voting rules, this is rather difficult to compare contests to one another. On top of that, the issue of synonyms in the image labels may rise if votes over multiple contests are analyzed simultaneously. For example, a model can not be only a fashion model/beauty peagant, but also a 3-dimensional model of a building. The same labels therefore may apper over contests with different meaning, which is again not sensible to compare. Last but not least, the sample size is still too small to perform statistically significant analysis over multiple contests.

5.4 Co-Clustering

The chosen Co-Clustering algorithm (introduced in Chapter 2) was first executed on a smaller set of 1-itemsets in the Miss Suomi 2017 contest. The clustering is performed for the supports calculated for 22 itemsets for both genders and age groups. The results are presented in the pragraphs to follow only by age groups. The analysis on the results by genders are not presented because there were no significant findings identified. This is mainly because the amount of data is smaller (*22*3=66* values in the matrix) compared to the results by age groups (*22*8=176* values in the matrix).

Figure 14 displays the matrix of itemset supports after the Co-Clustering for 3 clusters. Choosing the number of clusters was essentialy done by looking at the results and analyzing which k-value produces the most sensible results in this particular case. The colors correspond to the support of the itemset: the lighter values indicate smaller, the darker colors indicate higher support values. The clusters are circulated with different colors and displayed on the side and the bottom of the chart.

As it can be seen on the figure, the clustering algoritm organizes the rows of the matrix based on the patterns their values show. For instance, Cluster 3 contains the itemsets, which have high support values for most of the age groups. This cluster in particular contains the more engaging labels, which supports the findings of the Association Analysis. The itemsets are clustered together, because they have similar support values accross age groups. It can be also seen that these itemsets have similar contextual meaning, for instance lady, girl and bride. Therefore it is not much of a surprise, that their support values are similar and hence are in the same cluster.

Cluster 2 in this case contains only 2 itemsets (*{"wedding"}* and *{"model"}*), which seemed to be the interest of the 45-54 age group. Similarly, Cluster 1 contains a list of itemsets, which show differences for the 55-64 group. These distinctions has became even more apparent when the algorithm was executed with 4 clusters: in this case the 65+ age group was grouped under its own cluster alone. Strong conclusions about these findigns cannot be made, however it seems that the clustering approach has the capability to reliably distinguish between the groups and itemsets, which show differences. The clustering may not be entirely correct and human revision might be needed, for instance the *{"bride"}* itemset looks considerably different from the rest of the rows in Cluster 3.

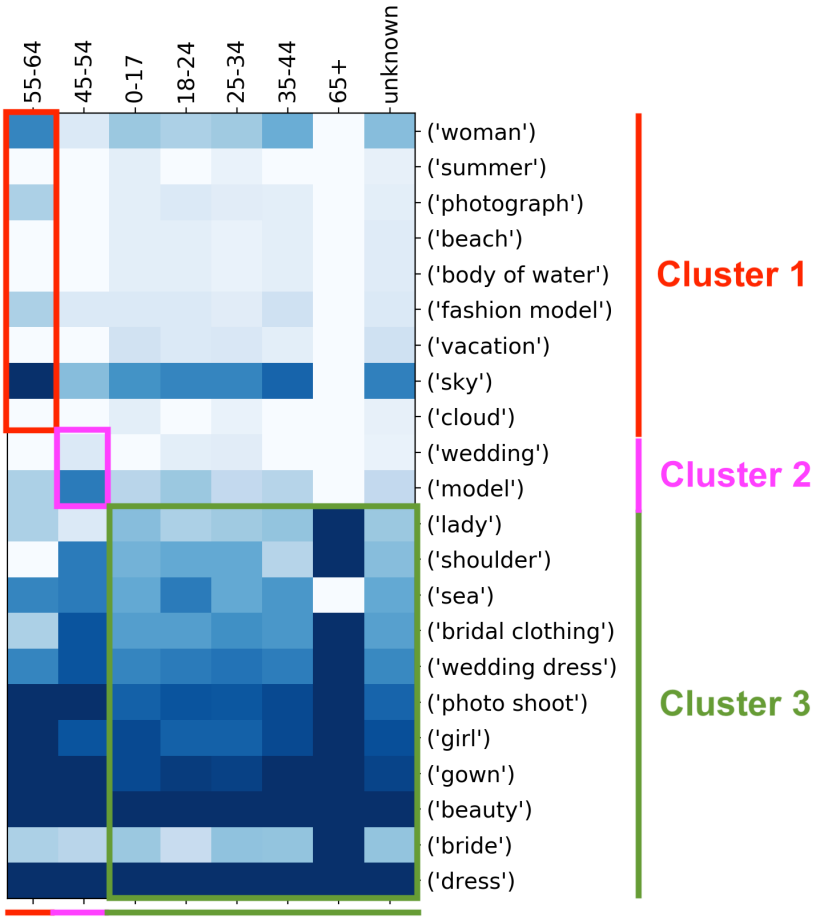


Figure 14: The results of the Co-Clustering in the Miss Suomi 2017 contest for 1-itemsets with 3 clusters by age groups.

Figure 15 displays the results for the multi-item itemsets. These are the same itemsets, which were analyzed during Association Analysis (created by combining Frequent Closed Itemsets, Figure 12). For the same reasons as above, the Co-Clustering is performed on age groups rather than genders. The matrix of support values is clustered to 4 clusters.

Cluster 2 contains most of the itemsets that are supported by all of the groups. The rest of the age groups are also assigned under cluster 2, where itemsets are generally have higher supports for all of the groups. It can be hence said that the itemsets in Cluster 2 are in the interest of generally all of the groups, while itemsets outside this are more specific to the interest of only some groups. This can be useful information to marketers, whose aim is to target one of these groups with certain content. Another useful aspect in this visualization is, that the itemsets with

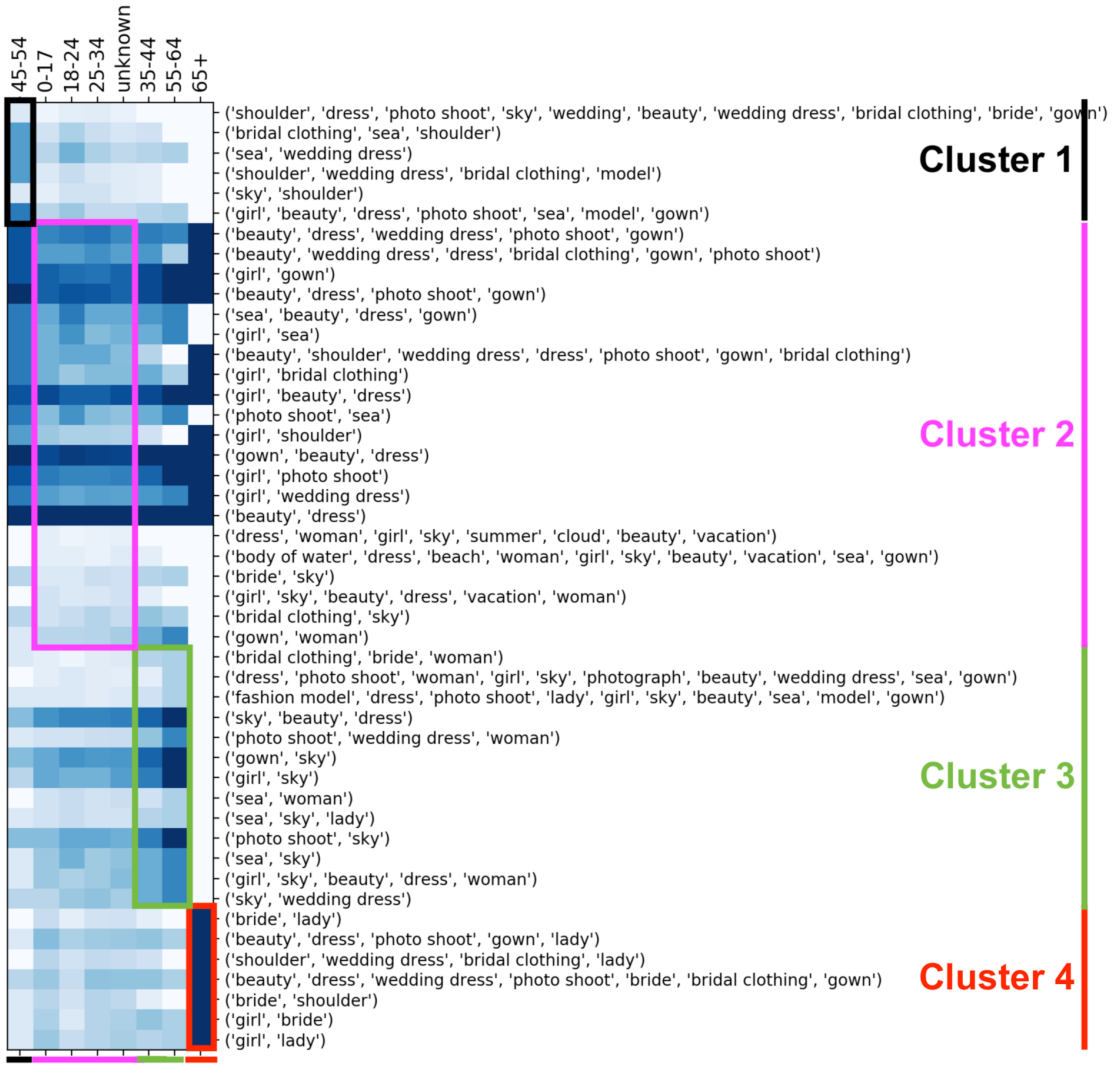


Figure 15: The results of the Co-Clustering in the Miss Suomi 2017 contest for multi-item itemsets with 4 clusters by age groups.

various supports can be distinguished easily based on the color's depth. Identifying itemsets that belong together is considerably easier on the right side of the figure as the Co-Clustering algorithm puts these close to eachother.

One of the most interesting observations in this case is that the age groups 35-44, 45-54 and 65+ are separated from the rest of the groups. This indicates that the interests of these groups are somewhat unique compared to the others. It can be also seen that the support values in groups 35-44 and 55-64 are considerably similar, not only in Cluster 3 but also outside of the clusters. It could be hence concluded, that some demographic groups follow a generalist, some others a specialist approach. In this particular case, users between the age of 0-34 are more specialists towards the itemsets in Cluster 2. Likewise, users from groups 55-64 and 65+ are unique, because some content (in clusters 1, 3 and 4) is engaging only to them.

It is interesting to see that some cells do not belong to any of the clusters according to the algorithm's outcome. This particular co-clustering method assumes a block-diagonal structure, which might not be the desired outcome in this case. Approximately half of the cells were unclustered in both cases above, which may not be the optimal solution in all cases. After seeing the results above it could be said, that this structure might not be the best solution to the problem of identifying the structure in the dataset. Some interpretations can be made intuitively, however a more sophisticated co-clustering approach could reveal more insights of the data.

Finally, the Co-Clustering is executed on the system level. The *minsup* value is set to 0.05 for this analysis and the itemset size is not limited to any numbers. When performing the analysis for genders, the same observation can be made as during the Association Analysis, namely that the data from the *not_chosen* age and gender groups dominates the data, because of which this group has higher support values for most itemsets. Inevitably, this fact also has an impact on the behavior of the Co-Clustering algorithm.

With 4 clusters and 166808 vote transactions, the results are as follows. All three genders (male, female and not chosen) are assigned to their own clusters with a subset of the itemsets in the platform. This findings suggests that the different genders tend to behave differently on the system level and there is a difference in which itemsets engage them.

The majority of the itemsets (977 out of 1626) are assigned to the "not_chosen" group, which is the consequence of the observation stated in the previous paragraph. Interestingly, the number of itemsets for females (148) is considerably lower than for

males (482). This suggests that males have wider range of interests than females, if the data is analyzed on the system-wide level. There are 19 itemsets that are not assigned into the same cluster with any of the demographic groups, which may mean that these are on the same level of interest for all of the groups.

When the system-wide analysis is performed for age groups with the same settings, the results are as follows. Similarly to genders, the users whose age is unknown are grouped under their own specific cluster with most of the itemsets (888 out of 2332). However, it is interesting to observe that age groups 0-17, 18-24, 25-34 and 45-54 are grouped under the same cluster with 314 itemsets. Age group 35-44 has its own cluster with 578 items and the 55-65 and 65+ age groups are again clustered together with 551 itemsets.

It can be seen that there is a clear distinction between the younger and the older generation in the clustering. This finding suggests that the engaging content is different for these groups, while the middle-aged (34-54) users are somehow between them. The algorithm was executed with various number of clusters, but this behavior was always present. This observation further enhances the validity of this finding and suggests further investigation on the reasons in this area.

6 Discussion

Based on the reviewed literature, there is no common understanding on the term of user data among researchers. While demographic data is very well understood, digital footprints (data gathered during the usage of a software) are less commonly studied in the field. It is clearly identified that the careful utilization of user demographic data and digital footprints, software service providers and researchers can retrieve previously unknown information about users' behavioral characteristics.

Users' demographic data and digital footprints are not usually brought under the same hood, are often vaguely defined and sometimes are studied separately. By introducing the concept of user data, a common ground of understanding is set for researchers of the future. As a scientist in this area, it is essential to understand the variety of user data in terms of availability and format. Depending on the software at hand, the data generated by digital footprints varies, while user demographic data is more uniformed.

Access to demographic data in the present time seems to be getting easier for researchers. Social networks already have standard ways of helping users to authenticate themselves while using other services (in other words use their social network credentials to use other services). Taking advantage of this, businesses can get access to rich demographic data of their user base easily. This fact creates a research space for studies, that have not been possible in the past.

The reviewed research papers demonstrate how others have utilized Data Mining and Machine Learning methods to reveal patterns or behavioral characteristics of users. The proper application data analysis methods allows us researchers, to learn more about the society as well as human behavior, which was not possible previously. This information is essential for business operations, because it gives insights on the user groups, their preferences and what kind of content keeps the audience engaged. In conclusion, the interest and relevance of applying computational methods in this area is proven.

Seeing the development by other researchers, user data can be used for various purposes. One is the acceptance of some new content or feature offering, such as the case of online banking. Through data analysis tools it is considerably convenient to see how well users of a software adapt to a new set of tools. By being aware of the required learning curve and awareness, banks could tell which part of their customer base (audience) is more interested in and ready for the usage of a new

toolset. Another angle to look at this application area is the preferences of the audience, not only in terms of content but also set of tools to choose from to perform specific tasks.

It is commonly known that different people like different things. Getting easy access to these individual desires is what keeps up the motivation and engagement towards the usage of a (software) service. The reviewed studies on SNSs as well as the case of Choicely have proven differences between groups of users in this respect. Typically extracting information related to users' preferences from the data is interesting already, but even more importantly helps companies to provide more customized services to users.

The set reviewed studies show, that Natural Language Processing and Supervised Machine Learning techniques are the most common approaches for conducting research on user data. The encouraging results achieved by other researchers prove the relevance of computational techniques applied on user data in a wide range of development areas, such as banking, social media, online movie databases, user portfolio analysis, to name a few. This suggests that computational methods are versatile enough to be applied in various domains. However, as every software and dataset is unique in some sense, there is not a single method that could be consdered as a silver bullet to solve problems. Association Analysis is shown to be a great method for this purpose, however finding statistically significant findings may be challenging depending on the studied data.

In the past audience engagement was hardly possible and knowing the interests demographic user groups was challenging. In the present time, access to this information can facilitate research, business processes, help determining the future content, analyze trends and understanding target groups of particular services better. In sum, modern Data Mining techniques created the potential to study human preferences and behavior from a different angle.

The results from studying the Choicely platform demonstrate an actual realization of the user data concept. The platform incorporates demographic data of its users as well as digital footprints of the software's usage. The nature of the platform is interesting, because users' engagement, voting trends and favor towards contest participants can be extracted from the data. Using the developed data analysis framework, contest organizers can analyze the engagement summary after a contest is over in a new way.

The developed tools can not only pinpoint the content that is more appealing to

users, but also reveal the tendencies shown by different demographic groups. This provides the potential for developing customized services and content for the groups, which then leads to better user experience while interacting with the platform. For example, if younger users show more interest in sports, while elders in traveling, customized content recommendations can be made to them, either in the Choicely platform or in external sites. Furthermore, this knowledge can also contribute to better product design and marketing, because the quantified results can confirm preliminary hypotheses concerning users' personal desires.

The chosen methods for the practical part of this study help the company to understand the data at hand better, as well as to provide better services to their customers. The methods used in this study are relevant tools towards answering the research questions, however their limitations were also demonstrated in the study. Hence it is important to know the pros and cons of these methods and think critically about their applicability and relevance to problems. Therefore, the chosen methods are retrospectively reviewed in the scope of this study in the paragraphs to follow.

EDA has provided great access to statistical measures as well as visualization tools to explore the data. Through this study it can be seen great technique for getting a grasp on what the data at hand looks like and what it contains. This method provides the possibility of laying down initial hypotheses and conceptualizing the "big picture". Looking at the data in general is a good idea not only to explore, but to understand how observations relate to eachother and what the connections between them are. Despite being a useful technique, EDA is limited to this only purpose and cannot answer complex questions, nor it can identify underlying structure or patterns in many cases. Last but not least, EDA can reveal also the potential biases in the data with the choice of proper visualization and statistical measures. In this study EDA for instance pinpointed the fact that many contests are considerably small in terms of voters and hence helped enhancing the results in later stages. The finding that some contests did not have proper meta-data and labeling was also obtained through this research method. Last but not least, the technique also revealed which type of contests in the platform have more data to work with, which led to better and valid results.

By performing Association Analysis some preliminary assumptions in the EDA phase have been confirmed. More importantly, the calculation of the itemset supports have contributed towards answering the research questions in more details. Based on the results it can be concluded, that contests in the "beauty" and "fashion" categories

create higher engagement than others, which may be explained by the history and the customer base of the firm. Nevertheless, it seems that contests where participants are human beings are more attractive to voters, whereas abstract objects or places appear less interesting to the audience.

Itemset supports on their own do not necessarily tell all of the interesting information about the engagement of the audience. The lifts of the itemsets are more interesting to look at and can enhance the provided information with more interesting and meaningful insights.

It was identified, that studying the generated itemsets in a single contest is possible and can assist the organizer to retrospectively analyze the content which has engaged the audience. Using this data, different demographic groups and their preferences can be compared and analyzed. Due to the fact that the biases in the data enlarge on the system level, it is difficult to derive valid results using the current methods and data on a larger scale. It would be hence important to enhance the quality of the data over the contest or to use an even subset of data for such purposes.

This method also contributed to identify, that the labels assigned to images by Computer Vision are very similar to eachother. This led to similar support values in some cases and somewhat biased results. The reason behind this is the fact that chosen Computer Vision tool recognized only the main objects on the images and omitted the details. It would be interesting to enhance the image labels with more specific tags or other meta data (either manually or automatically), that tell unique features of the participants, then perform the same analysis and compare the results. This approach could point out more details concerning which detailed features of participants lead to more attention and engagement from users.

The large number of generated itemsets via this technique also makes the readability of the results challenging. When the generated itemsets grow over certain size (typically 2-3 items), the number of output items is too high to analyze manually. The Co-Clustering technique addresses this challenge well by clustering the matrix of itemset supports and the demographic groups. Through this technique, the itemsets and demographic groups whose structures are similar can be identified and analyzed more carefully together. The results of this approach also can make suggestions for demographic groups, who share similarities in terms of their voting behavior.

One of the challenges with the Co-Clustering method is to choose the number of clusters. For the time being, there was not a single good way of suggesting the best value for this value, because the best match for the number of clusters depends

on the size of the dataset (in terms of itemsets and demographic groups) and the underlying structure of the data. Nevertheless, it seems that trial and error works considerably well in this case, but interpreting the results is often not evident. It would be interesting to develop the solution towards a direction, where this challenge is addressed.

The assignment of the cluster labels in the current implementation could be questioned. The fact of assigning one cluster to every row and colum puts a limitation to the outcome. A large number of cells do not belong to any clusters, because the algorithm assumes a block-diagonal structure, which might not be the case in this kind of dataset. In order to improve on that, it would be a good idea to apply more complex and flexible Co-Clustering algorithms for the analysis of the underlying structure. Such algorithm could consider cell-wise comparisons and cluster assignments, which would be more robust in identifying homogeneity of items inside clusters.

Through the results it can be seen how differences in the behavior of demographic groups are revealed. For instance, through the application of the methods it is speculated that users in the Choicely platform can be seen as specialists and generalists. Furthermore, the data suggests distinctions between the younger end elder generation as well.

Biases in the data limit the conclusions and the viability of the results and raise threads to their validity. To address this issue, there is a need for a more careful evaluation on the statistical significance of the results. The dataset could be enhanced with more reliable and validated data. One could carry out a structured data collection from a chosen sample of users to ensure the validity of the collected data. Another angle on this topic is that the Choicely platform is rather unique in nature compared to social media sites for instance. Therefore the results obtained in this study may be difficult to reproduce in other online software platforms.

Privacy and ethical concerns are interesting for many reasons to discuss through the scope of this study. Undoubtedly, the recent technical evolvement has brought many challenges with respect to human rights and data protection, hence there is a need to lay down a common standard in the European Economic Area (EEA). The General Data Protection Regulation (hereinafter GDPR) [Eur16] is being issued on 25th May 2018, which regulates the business and data processing activities by protecting personal data through its novel standards all over the world. According to the regulations, personal data means any kind of information through which the

natural person (or data subject, who has generated that data) can be identified [Eur16].

The data analysis activities performed in this study are operating on personal and sensitive data, however the results do not reveal any piece of that data. Based on the results of the audience engagement presented in the present research, there is no possible way to identify individual data subjects. The results only show the extraction and statistically significant information about the collected data, but fully hide the individual's personal data. In terms of the GDPR, these activities are called "profiling" [Eur16]. As by Article [Eur16], the data subject has to be clearly informed about the purposes and goals how his or her generated data is being used. Many other important aspects, such as portability, erasure or automated decision making are addressed over multiple articles (12-23) concerging the rights of the data subjects and the responsibilities of the processor [Eur16].

In terms of the case company this means, that the purposes of any data processing tasks should be clearly stated and communicated to users. A standard and convenient channel to forward this information to the platform's users is the company's privacy policy[20], which already contains a list of points in this topic. Nevertheless, as the results and methods of this study are being added to the platform, this list should be extended by clearly expressing the purposes and goals of data collection and analysis.

[20] `https://choicely.com/about/privacy`

7 Conclusions and future work

User data is widely collected and commonly used for research purposes as well increasing business value of services. The willingness to give this data from users' perspective is present, as many social media sites create a reliable and convenient hub for such repositories. As a consequence, businesses and researchers can easily get access to rich demographic and software usage data through the interfaces of social media platforms. Furthermore, these platforms serve the purpose of convenient authentication and integration to other services, which is highly appreciated by the users. On top of that, the software service providers get easy, quick and reliable access to demographic information about their users.

Despite its wide availability, the concept of user data is not commonly used in the literature. Nevertheless, numerous studies were conducted in the past - mainly via social media websites - to analyze engagement and trends on how users behave in an online environment. The research topics mainly resolve around what kind of content is attractive to them or how they communicate with eachother in the online software. One of the challenges in this area is, that every software platform is unique, therefore there is not a single way of studying these aspects. Secondly, the data at hand can differ from platform to platform, nevertheless it is commonly too large to analyze using human resources. For this reason, various Data Mining and Machine Learning techniques were utilized by the scientific community to enhance and analyze the data at hand.

Through this research it was revealed, that while user demographic data is often well established, digital footprints are very domain-dependent in nature. It is concluded that the both sides of user data is necessary to analyze user characteristics and behavior. The combination of computational methods and such data has been already utilized in various areas, such as banking, studying and predicting movie ratings, social media studies and enterprise social networks. This observation proves the relevance and the wide applicability of this research area. Depending on the domain, various tools and methods, such as computer vision, machine learning, association analysis and natural language processing techniques are available, but none of them is a silver bullet for every dataset.

The Choicely voting/audience engagement platform is an excellent specimen for a software service, where user data is utilized. The aim of this research was to study the user data that is collected by the Choicely platform. The study analyzes the

type of content that appears more engaging to users and investigates the voting behaviors of different demographic groups.

The user data at Choicely is two folded. Users from all over the world log in to the platform, typically using one of their social media credentials to authenticate their identity. As a result, their personal profiles are created in the Choicely platform with their demographic data pre-populated, which establishes one half of the data. The second half of user data is generated by users while using the software. In case of the Choicely platform this means vote transactions on the participants of the contests.

Each contest participant can be connected to a single image in the platform. The contents on the images varies and generally speaking, there is no standard of labeling the images with meta data for analysis purposes. As a result, analysis on the vote transactions is difficult, because it is not possible to know the type of content users find engaging.

To tackle this problem, Computer Vision (using Google Vision) was integrated to the platform. This technology can reliably assign meta data to the images concerning their content automatically, whenever they are uploaded to the platform. The vote transactions hence can contain not only on who voted on which participant, but also what the content on that participant's image is. The combination of this data together with the users' demographic data opens up the possibility for a more sophisticated data analysis on the data, in order to study engaging content and users' behavior in the platform.

To enhance the firm's business package, the basis of a data analysis framework were established. Association Analysis is utilized to extract the support values of the itemsets that appear in the vote transactions. The vote transactions can be studied from multiple angles, however this study is limited to standalone contests and system-wide analysis. By studying transactions of a single contests organizers can evaluate, which labels tend to be more engaging to different demographic in the contest. Furthermore, this methods provides room for analyzing differences between groups as well as itemsets that appear in the contest. The system-wide analysis can reveal the same on a larger, global scale from a richer dataset.

To computationally identify patterns and the underlying structure of the itemset supports, the Co-Clustering is used. The Co-Clustering is performed on matrix of itemset supports by demographic groups. This approach provides a reliable way to identify which demographic groups and itemsets tend to behave similarly in the

data. As a result, engaging content that are specific to groups can be pinpointed, differences in the voting behavior between groups of users can be revealed.

The results show that there are certain contest categories, which have hosted many more contests, hence have got more engagement than others. The Exploratory Data Analysis revealed, that there is a high number of very small contests. Deriving from this, only a fraction of the whole dataset is actually ready for complex data analysis, because the rest simply does not have sufficient amount of data to analyze.

The Association Analysis on the itemsets recognized by Computer Vision reveals some interesting insights on the kind of content, which is more engaging to users. One of these findings is, that the contests where participants are actual people appeal to voters more than objects or other abstract concepts. Secondly, objects that appear in the foreground of the images might have more impact on the voting trends than the background or the surroundings, but there is no clear support for this claim. The results also reveal that Computer Vision in this research is useful to identify overall concepts of the images, however it fails to recognize smaller details on the pictures.

The proposed evaluation on the statistical significance establishes the basis of comparing multiple groups of users once a contest is over. Further development on the automatic execution of this method is needed, which could then easily pinpoint differences in users' behavior. On top of that, the statistical comparison of results over multiple contests is an interesting topic for further studies.

Last but not least, the Co-Clustering approach was successfully used to pinpoint similar demographic groups and itemsets in the vote transactions. This technique provides a robust way to analyze behavior of demographic groups of users and pieces of content. However, this method has not proven useful on the system level, because the data in the platform is often incomplete. More particularly, the demographic data is often missing for many of the users and the assigned labels to the images are often share many similarities in contests.

Future work could address on enhancing the quality of the data. The validity of the results in this study could be elevated applying these techniques on a more carefully selected and tested dataset. Future work could also build on top of the results of this study, for example a content recommendation system could be built using the results of Association Analysis and Co-Clustering. It would be also interesting to see the chosen methods performance in similar software platforms, particularly in social media applications.

Due to the fact that user data involves a large set of personal data, there is an emerging need in studying and establishing standards for the research community. Future work could elaborate also more in-depth on the ethical concerns when conducting research with sensitive user data. Furthermore, it would be a interesting to approach this area not from only computer science, but human behavior point of view. Finally, the application of Data Mining and other computational techniques to facilitate studies focusing on psychology and human behavior would be another stimulating direction for researchers. Conducting more studies in this area would further deepen our understanding on human behavior, preferences as well as gender and cultural differences.

References

AHTB16 Aué, J., Haisma, M., Tómasdóttir, K. F. and Bacchelli, A., Social diversity and growth levels of open source software projects on github. *Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM '16, New York, NY, USA, 2016, ACM, pages 41:1–41:6, URL `http://doi.acm.org/10.1145/2961111.2962633`.

AIS93a Agrawal, R., Imielinski, T. and Swami, A., Database mining: a performance perspective. *IEEE Transactions on Knowledge and Data Engineering*, 5,6(1993), pages 914–925.

AIS93b Agrawal, R., Imieliński, T. and Swami, A., Mining association rules between sets of items in large databases. *SIGMOD Rec.*, 22,2(1993), pages 207–216. URL `http://doi.acm.org/10.1145/170036.170072`.

AS94 Agrawal, R. and Srikant, R., Fast algorithms for mining association rules in large databases. *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, San Francisco, CA, USA, 1994, Morgan Kaufmann Publishers Inc., pages 487–499, URL `http://dl.acm.org/citation.cfm?id=645920.672836`.

AS95 Agrawal, R. and Srikant, R., Mining sequential patterns. *Proceedings of the Eleventh International Conference on Data Engineering*, ICDE '95, Washington, DC, USA, 1995, IEEE Computer Society, pages 3–14, URL `http://dl.acm.org/citation.cfm?id=645480.655281`.

Bla14 Blasiak, Kevin, Big data; a management revolution : The emerging role of big data in businesses, 2014. URL `{http://theseus.fi/handle/10024/74701}`.

BM01 Bose, I. and Mahapatra, R. K., Business data mining - a machine learning perspective. *Information & management*, 39,3(2001), pages 211–225.

BMS97 Brin, S., Motwani, R. and Silverstein, C., Beyond market baskets: Generalizing association rules to correlations. *SIGMOD Rec.*, 26,2(1997), pages 265–276. URL `http://doi.acm.org/10.1145/253262.253327`.

BMUT97 Brin, S., Motwani, R., Ullman, J. D. and Tsur, S., Dynamic itemset counting and implication rules for market basket data. ACM Press, 1997, pages 255–264.

BSG14 Bakhshi, S., Shamma, D. A. and Gilbert, E., Faces engage us: Photos with faces attract more likes and comments on instagram. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 2014, pages 965–974.

Cho08 Cho, H., Co-clustering algorithms: Extensions and applications.

Dhi01 Dhillon, I. S., Co-clustering documents and words using bipartite spectral graph partitioning. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, New York, NY, USA, 2001, ACM, pages 269–274, URL `http://doi.acm.org/10.1145/502512.502550`.

Eur16 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*, L119, pages 1–88. URL `http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2016:119:TOC`.

Fey96 Feyyad, U. M., Data mining and knowledge discovery: making sense out of data. *IEEE Expert*, 11,5(1996), pages 20–25.

FNAC15 Farseev, A., Nie, L., Akbari, M. and Chua, T.-S., Harvesting multiple sources for user profile learning: A big data study. *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, ICMR '15, New York, NY, USA, 2015, ACM, pages 235–242, URL `http://doi.acm.org/10.1145/2671188.2749381`.

Fra17 Fractl, How are news publishers' audiences reating on facebook?, 2017. `http://www.frac.tl/research/facebook-reactions`. [29.9.2017]

Fri97 Friedman, J. H., Data mining and statistics: What's the connection. *Proceedings of the 29th Symposium on the Interface Between Computer Science and Statistics*, 1997.

GRZ+16 Guy, I., Ronen, I., Zwerdling, N., Zuyev-Grabovitch, I. and Jacovi, M., What is your organization 'like'?: A study of liking activity in the enterprise. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, New York, NY, USA, 2016, ACM, pages 3025–3037, URL http://doi.acm.org/10.1145/2858036.2858540.

Har72 Hartigan, J. A., Direct Clustering of a Data Matrix. *Journal of the American Statistical Association*, 67,337(1972), pages 123–129. URL http://dx.doi.org/10.2307/2284710.

HLJ+16 Han, K., Lee, S., Jang, J. Y., Jung, Y. and Lee, D., Teens are from mars, adults are from venus: Analyzing and predicting age groups with behavioral characteristics in instagram. *Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, New York, NY, USA, 2016, ACM, pages 35–44, URL http://doi.acm.org/10.1145/2908131.2908160.

HMK+14 Hu, Y., Manikonda, L., Kambhampati, S. et al., What we instagram: A first analysis of instagram photo content and user types. 2014.

Hol99 Holsheimer, M., Data mining by business users: Integrating data mining in business processes. *Tutorial Notes of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, New York, NY, USA, 1999, ACM, pages 266–291, URL http://doi.acm.org/10.1145/312179.312196.

Hut17 Hutchinson, A., Should you use facebook reactions as a voting mechanism on live broadcasts?, 2017. http://www.socialmediatoday.com/social-business/should-you-use-facebook-reactions-voting-mechanism-live-broadcasts. [28.9.2017]

IN07 Inmon, W. H. and Nesavich, A., *Tapping into unstructured data: integrating unstructured data and textual analytics into business intelligence*. Pearson Education, 2007.

JHL15a Jang, J. Y., Han, K. and Lee, D., No reciprocity in "liking" photos: Analyzing like activities in instagram. *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, HT '15, New York, NY, USA, 2015, ACM, pages 273–282, URL http://doi.acm.org/10.1145/2700171.2791043.

JHL15b Jang, J. Y., Han, K. and Lee, D., No reciprocity in liking photos: Analyzing like activities in instagram. *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. ACM, 2015, pages 273–282.

JHL+16 Jang, J. Y., Han, K., Lee, D., Jia, H. and Shih, P. C., Teens engage more with fewer photos: Temporal and comparative analysis on behaviors in instagram. *Proceedings of the 27th ACM Conference on Hypertext and Social Media*, HT '16, New York, NY, USA, 2016, ACM, pages 71–81, URL `http://doi.acm.org/10.1145/2914586.2914602`.

KCC12 Kabinsingha, S., Chindasorn, S. and Chantrapornchai, C., Movie rating approach and application based on data mining.

LLC17a LLC, G., Cloud vision api, 2017. `https://cloud.google.com/vision/`

LLC17b LLC, G., Google cloud datastore overview, 2017. `https://cloud.google.com/datastore/docs/concepts/overview`

LLC17c LLC, G., What is google bigquery?, 2017. `https://cloud.google.com/bigquery/what-is-bigquery`

LRD09 Liu, Y.-H., Ren, Y. and Dew, R., Monetising user generated content using data mining techniques. *Proceedings of the Eighth Australasian Data Mining Conference - Volume 101*, AusDM '09, Darlinghurst, Australia, Australia, 2009, Australian Computer Society, Inc., pages 75–81, URL `http://dl.acm.org/citation.cfm?id=2449360.2449377`.

OPLC+13 Ottoni, R., Pesce, J. P., Las Casas, D. B., Franciscani Jr, G., Meira Jr, W., Kumaraguru, P. and Almeida, V. A., Ladies first: Analyzing gender roles and behaviors in pinterest. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

PVG+11 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E., Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, pages 2825–2830.

RC10 Raeder, T. and Chawla, N. V., Market basket analysis with networks. *Social Network Analysis and Mining*, 1, pages 97–113.

SCDT00 Srivastava, J., Cooley, R., Deshpande, M. and Tan, P.-N., Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explor. Newsl.*, 1,2(2000), pages 12–23. URL `http://doi.acm.org/10.1145/846183.846188`.

Sei16 Seiter, C., Psychology of facebook?, 2016. `https://blog.bufferapp.com/psychology-of-facebook`. [30.9.2017]

SKM Sumathi, T., Karthik, S. and Marikannan, M., Performance analysis of classification methods for opinion mining.

sld17 scikit-learn developers, Biclustering, 2017. `http://scikit-learn.org/stable/modules/biclustering.html`

SWE04 Saraee, M., White, S. and Eccleston, J., A data mining approach to analysis and prediction of movie ratings. *Transactions of the Wessex Institute*, pages 343–352.

TSK05 Tan, P.-N., Steinbach, M. and Kumar, V., *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.

Tuf11 Tufféry, S., *Data mining and statistics for decision making*, volume 2. Wiley Chichester, 2011.

Tuk77 Tukey, J. W., *Exploratory Data Analysis*. Addison-Wesley, 1977.

VMBDB04 Van Mechelen, I., Bock, H.-H. and De Boeck, P., Two-mode clustering methods: A structured overview. *Statistical methods in medical research*, 13,5(2004), pages 363–394.

WARK17 Waheed, H., Anjum, M., Rehman, M. and Khawaja, A., Investigation of user behavior on social networking sites. *PloS one*, 12,2(2017), page e0169693.

WFHP16 Witten, I. H., Frank, E., Hall, M. A. and Pal, C. J., *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

WP17 Wang, S. and Petrounias, I., Big data analysis on demographic characteristics of chinese mobile banking users. *2017 IEEE 19th Conference on Business Informatics (CBI)*, volume 02, July 2017, pages 47–54.

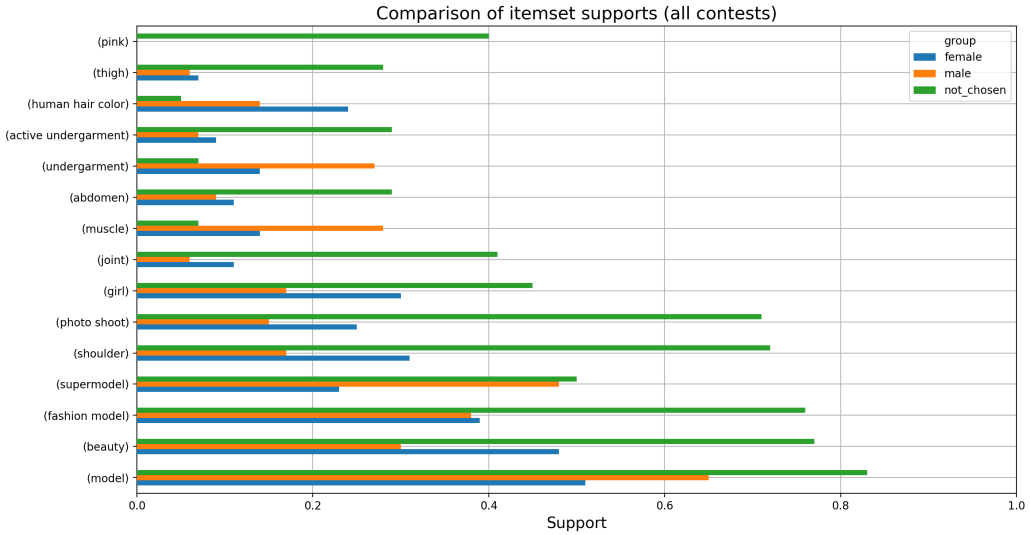
WR10 Wegener, D. and Rüping, S., On integrating data mining into business processes. Springer, 2010.

YKS15 Youyou, W., Kosinski, M. and Stillwell, D., Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112,4(2015), pages 1036–1040.

Zar02 Zarsky, T. Z., Mine your own business: making the case for the implications of the data mining of personal information in the forum of public opinion. *Yale JL & Tech.*, 5, page 1.

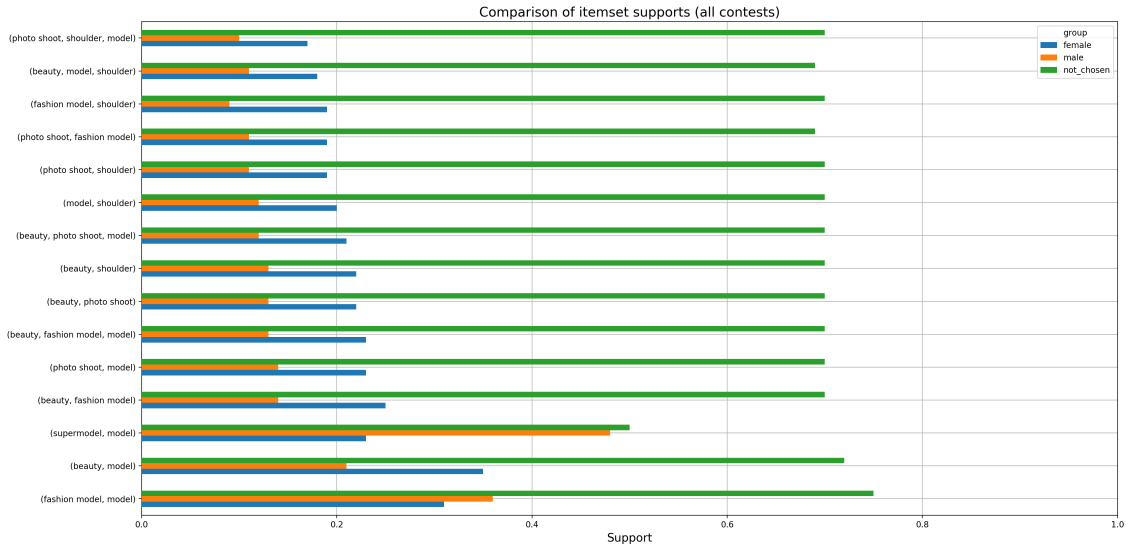
ZEEA11 Zengin, K., Esgi, N., Erginer, E. and Aksoy, M. E., A sample study on applying data mining research techniques in educational science: Developing a more meaning of data. *Procedia - Social and Behavioral Sciences*, 15, pages 4028 – 4032. URL `http://www.sciencedirect.com/science/article/pii/S1877042811009542`.

Appendix 1. Itemset supports over all contests



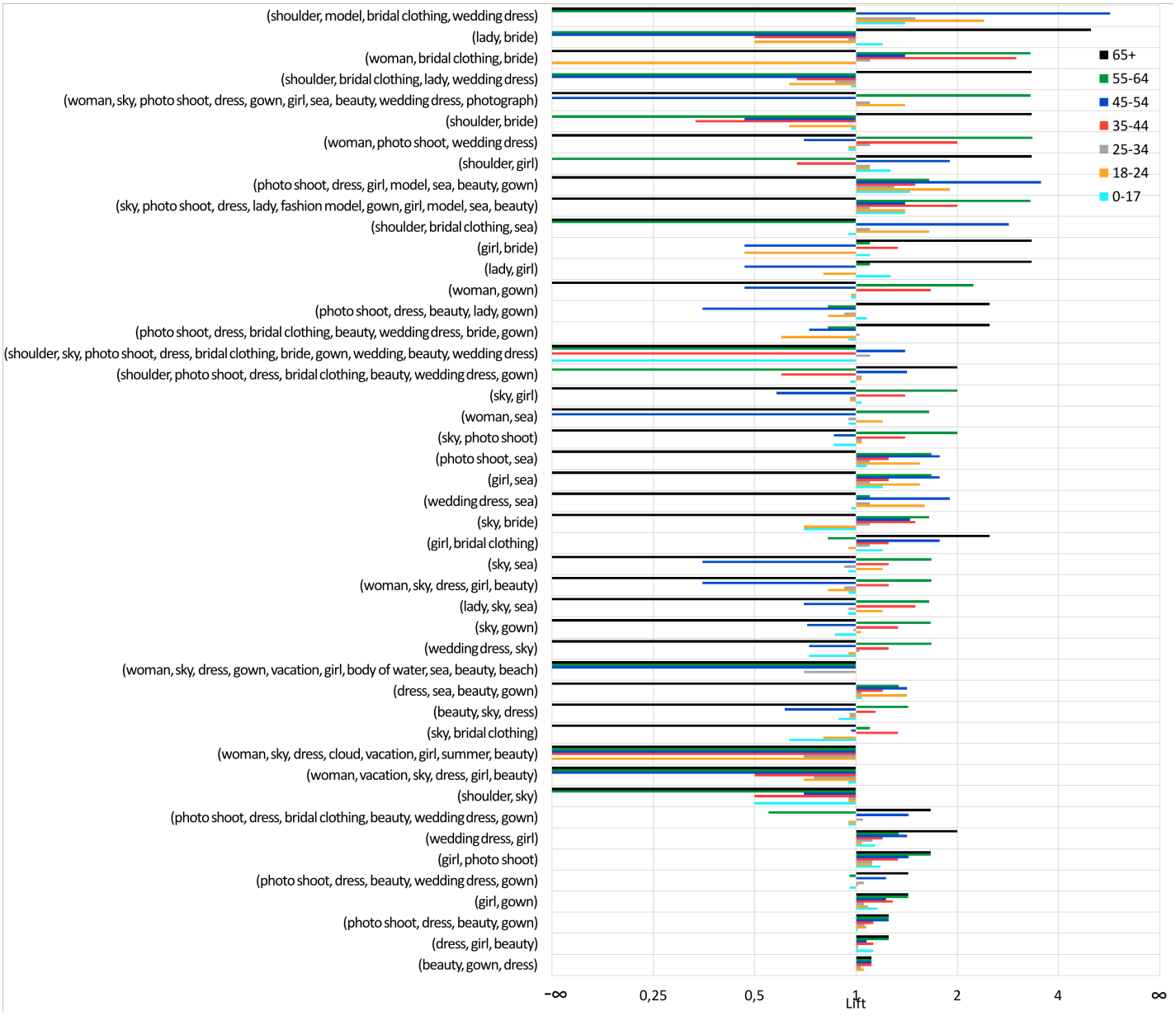
The 1-itemset supports in all contests by genders (only the first 15 itemsets are displayed ordered by sum of the support values).

Appendix 2. The k-itemset supports over all contests



The k-itemset supports in all contests by genders, where $k \geq 2$ (only the first 15 itemsets are displayed ordered by sum of the support values).

Appendix 3. Multi-item itemset lifts in Miss Suomi 2017 by age groups



The multi-item itemset lifts in the Miss Suomi 2017 contest by age groups (all itemsets are displayed ordered by the variance of the lift values).