Journal of Social Sciences of Mus Alparslan University

anemon

Derginin ana sayfası: http://dergipark.gov.tr/anemon

*Araştırma Makalesi ● Research Article*

Special Issue on *International Conference on Applied Economics and Finance (ICOAEF' 19), 9-11 April, 2019, Kyrenia, T.R.N.C.*

# The Relationship among Personality, Interest, and Life Satisfaction of Facebook Users

## *Facebook Kullanıcılarının Kişilik, İlgi Alanı ve Yaşam Memnuniyeti Analizi*

Vesile Evrim [a,*], Yahya Nissoul [b]

[a] Assist. Prof. Dr., European University of Lefke, Faculty of Engineering, Computer Engineering, Lefke, Northern Cyprus, TR-10, Mersin, Turkey
ORCID: 0000-0001-7733-5229

[b] European University of Lefke, School of Applied Sciences, Management Information Systems, Lefke, Northern Cyprus, TR-10, Mersin, Turkey
ORCID: 0000-0001-7017-5162

ARTICLE INFO

Article history:
Received 03 September 2019
Received in revised form 04 October 2019
Accepted 01 November 2019

Keywords:
Life Satisfaction
Personality Traits
Facebook Likes
Association Rules

ABSTRACT

Today, the data collected by e-commerce systems are beyond the traditional statistical and demographic information of customers. The data provided by comments, likes, tags, photos and more in social media, enable marketing researchers to better evaluate the behavior of the users. Therefore to analyze user behavior and characteristics, in this paper, 3 balanced subset of myPersonality Facebook dataset is tested by Apriori algorithm. As a result the relationship among personality traits, intelligence quotient , satisfaction with life scale and the assigned 12 interest categories of users are analyzed.

MAKALE BİLGİSİ

*Makale Geçmişi:*
Başvuru tarihi: 03 Eylül 2019
Düzeltme tarihi: 24 Ekim 2019
Kabul tarihi: 01 Kasım 2019

Anahtar Kelimeler:
Yaşam Memnuniyeti
Kişilik Özellikleri
Facebook Beğenileri
İlişkilendirme Kuralları

ÖZ

Günümüzde, e-ticaret sistemleri tarafından toplanan veriler, müşterilerin geleneksel istatistiksel ve demografik bilgilerinin ötesindedir. Sağlanan yorumlar, beğeniler, etiketler, fotoğraflar vb. sayesinde bir uygulamanın kullanıcıları hakkında daha fazla bilgi sahibi olmak mümkün olmaktadır. Kullanıcı davranışlarını ve özelliklerini analiz etmek amacıyla, bu çalışmada, myPersonality veri setinin kişilik, Zeka Katsayısı,Yaşam Memnuniyeti Ölçeği puanlarına sahip kullanıcıları içeren 3 altkümesi, kullanıcıların beğendiği kayıtlar ile birleştirilerek, on iki ilgi kategorisine ayrılımış ve Apriori algoritması kullanarak test edilmiştir. Sonuç olarak, türetilmiş ilişkilendirme kuralları sayesinde, Facebook kullanıcılarının kişisel özellikleri ile ilgi kategorileri arasındaki ilişkiler elde edilmiştir.

## 1. Introduction

The use of web technologies and applications enable users to share their personal information through blogs, and social networking sites such as Facebook, Instagram, and Twitter. As users of the sites provide more information about their thoughts, interests, goals and demographic information, companies are attracted to learn more about the users for better marketing strategies. As companies changing their business model to provide personalized services to users (Brusilovski, Alfred and Wolfgang, 2007), knowledge discovery field helps extraction of the hidden personalized information of users from text.

Personality can be defined as a set of characteristics effecting person's behavior, feelings (Schacter, Gilbert and Wegner, 2010) in different situations which plays a major role in decision making (Digman, 1989). In Psychology, Big Five

* Sorumlu yazar/*Corresponding author*.
e-posta: vevrim@eul.edu.tr


http://dx.doi.org/10.18506/anemon.615115

(Costa and McCrae, 2008), a widely accepted model, uses five traits to describe the personality: Agreeableness (A), Consciousness (C), Extraversion(E), Neuroticism(N), and Openness(O). In most of the cases as the traits of the personality is not explicitly provided by users, the trait information need to be extracted from the other available information of the user.

One of the focuses of personality detection from social media is the use of language. To extract the relationship among the personality traits and the word usage, Linguistic Inquiry Word Count (LIWC) (Pennebaker et al., 2001), a well known tool, is commonly used. Although the area of language usage initially studied in specific domains (Pennebaker and King, 1999), in time, it is extended to be used in many domain independent areas such as blogs (Yarkoni, 2010) and social networking sites (Schwartz, 2013).

Golbeck et al. (2011) analyze the personality traits on Facebook user data. They experimented 167 users status updates and social features (e.g friends, demographic information) and by using m5Sup/Rules, and Gaussian Processes, they able to predict the personality traits within the margin of 11% of actual values.

In 2013, Computational Personality Recognition (Shared Task) workshop is organized to enable researchers to test their methodology on the subset of Mypersonality dataset. The provided subset includes 9900 status updates and social networking information of 250 users (Celli et al., 2013). From the participants Farnadi et al. (2013) used LIWC, social network features, time related features, and other features (e.g., number of status updates) of each user with several machine learning algorithms and concluded that different features emphasis on different personality traits. Again several other research works proved that the different features of social media such an emoticons, slang word usage have different effects on determination on different personality traits (Alam, Stepanov and Riccardi, 2013; Markovikj et al., 2013).

Yoram Bachrach et al. (2012) tested the correlation between the personality traits and the number of friends, photos, event group membership and photo tags. Their results show that combination of several features help to achieve good prediction of traits. Their experiments proved that extraversion tend to be the easiest and agreeableness the hardest trait to predict.

As personality is studied, the relationship between personality and Intelligent Quotient (IQ) which refers to cognitive intelligence is also discussed. Although IQ and Personality seems to have commonality, since personality is measured by questionnaires and IQ is measured by ability test, in the past they are considers being a part of different domains. However, recent research proves that it is also possible to measure personality traits through an ability test and therefore IQ can be considered as a part of personality. As Steinberg (2011) experiment on the topic, from the traits, IQ is most strongly related to openness-to-experience and less related to the traits constituted by sensitivity and beauty. While extracting personality traits from social media becomes an interesting topic to work, the effect of social media on different age groups are investigated (Zhan et al., 2016). Zhan also conclude that social media usage can have a positive effect on life satisfaction of people. Schimmack et al (2002) studied the effect of personality on life satisfaction and predicted that extraversion and neuroticism on life satisfaction is mediated by hedonic (pleasant and unpleasant) balance.

As the different personality characteristics are extracted from social media text, the knowledge gained from experiments can be used in recommender systems to enable companies to do better marketing. Hu and Pu (2011) used Pearson's correlation to find the users with similar personality traits and they analyzed the music preferences of similar user. As an extension, they combined their results with rating based collaborative filtering which they obtained significant improvement compared to standard collaborative algorithms. Rochana (2012) analyzed the hotel reviews to extract personality traits and used Nearest Neighbour algorithm to enhance personalized recommendations.

Cantador et al., (2013) used 5 personality traits on a subset of Mypersonality dataset. They used subset of several entertainment domains (movies, TV shows, music and book) on 16 genres of each domain and by using association rules; they observe the relationship between the personality types and different domains.

Although extracting personalities in social media is studied in different domains, in this paper we analyze the relationship among the personality traits, Satisfaction with Life Scale (SWL), IQ characteristics of users with 12 diverse interest groups (Community, Food, Game, Health, Music, Shop, Animal, Home, Religion, Travel, and Politics) on a Mypersonality dataset. As a result, In addition to extracting valuable association rules between the user characteristics and the interest types, we conclude about the relationship among the various user characteristics of the dataset.

## 2. Association Rules

Association rules are rule based machine learning method used to discover the interesting relations in big dataset. The method is mainly used for market research to find the related items from the basket of a user data. By the help of association rules, marketers aim to place items physically closer to help customers to remember what they need while reducing their search time. In online shopping although there is no physical placement of the products, the learned relationships are used for recommendation. Similarly the rules extracted thorough our experiments is expected to be a guide in recommendation process.

### 2.1. Performance Measures

A rule is composed of antecedent/s on the left hand side and the consequent/s on the right hand side which is known as itemset (e.g., $X \rightarrow Y$). In order to select the interesting and significant rules created by association rules, several performance measures are used.

Support is used to identify the frequency (significance) of itemset in the dataset and its value ranges [0,1] (Agrawal,

Imielinski and Swami, 1993). Assuming that T represent the set of transactions in the dataset, and t is the subset of transactions contain the itemset X, Support(X) = |X⊆ t| / |T|. Therefore the support of a rule X→ Y in T is calculated as follows in equation (1)

$$\text{Support } (X \rightarrow Y) = \text{Support } (X \cup Y) \quad (1)$$

Confidence measure in equation (2) is used to find out how often the rule has been true (Brin et al., 1997a). Another words for X→ Y, confidence is the conditional probability of finding the consequent (Y) given the antecedents (X). Confidence value 1 means consequent and antecedent are always appears together.

$$\text{Confidence } (X \rightarrow Y) = \frac{\text{Support } (X \cup Y)}{\text{Support } (X)} \quad (2)$$

Although the general concept is, greater the support and confidence are the better the itemset, this sometimes might not reflect the real case. Such as an itemset which appears in many transactions (high support value) or has a confidence 1, does not necessarily semantically make sense and might be misleading. Therefore other performance measures are also introduced.

Lift measure as presented in equation (3) is used to measure how independent the antecedent and consequent to each other (Piatetsky-Shapiro, 1991). While high lift value shows the dependence on the occurrence of the antecedent and consequent, lift value=1 shows independence and the values between (0, 1) indicates the negative dependence.

$$\text{Lift } (X \rightarrow Y) = \frac{\text{Support } (X \cup Y)}{\text{Support } (X) \times \text{Support } (Y)} \quad (3)$$

Conviction is used to measure the dependence of consequent on the antecedent as in equation (4) (Brin et al., 1997a). Higher the conviction value, consequent is more depend on the antecedent. Conviction value 1 means the items are independent and the values between (0, 1) mean negative dependence.

$$\text{Conviction } (X \rightarrow Y) = \frac{1 - \text{Support } (Y)}{1 - \text{Confidence } (X \rightarrow Y)} \quad (4)$$

Therefore in the scope of our study, we take into consideration support, confidence and conviction measurements in selection of the most relevant association rules generated from the datasets.

## 3. Dataset

In this research myPersonality (Kosinski and Stillwell, 2011) dataset which is collected through the provided IPIP (Goldberg, 2006) test of the Facebook users is used. The collected personality data of the test is measured by Five-factor model of personality and summarized as follows:

(i) Openness (O): This trait refers to people with high imagination, intellectualism, adventurism and curiosity.
(ii) Conscientiousness (C): this trait refers to planned, self-disciplined, task-oriented and efficient people.
(iii) Extraversion (E): This trait refers to outgoing, energetic people who enjoy the company of others.
(iv) Agreeableness (A): This trait refers to friendly and companionate people who have high mortality.
(v) Neuroticism (N): This trait refers to negative, depressive and angry people.

From the users that took the IPIP test, as a return to learn their personality, over 3 million users donated their data for research. Besides the personality data, the dataset includes demographic information, friend relations, user likes, status updates, in addition to SWL, IQ, medication information and many more data.

The aim of this study is to find out the relationship between the Big5 personality traits, IQ, SWL and user interest. Therefore, the data related to these features are mined from the Mypersonality database. The database has roughly 101.000 users' life satisfaction score, 7200 users' IQ test results, 3.100.000 users Big5 personality traits and 46.200.000 "user likes". In total 1631 users had an input for all of the mentioned features with total of 634673 likes in 267 different categories. To be able to have meaningful results relating to interest, the subset of likes (83 categories) are combined under 12 main categories as presented in Table 1 and the resultant dataset included 1511 users' 8619 likes in 12 categories.

Since we would like to analyze users' characteristic with their interests, despite a user might have many likes in a particular interest category, the user's characteristics for that category is considered only once. For example if a user has 2000 likes in a music category, that particular user's SWL, IQ, openness, extraversion, neuroticism, agreeableness, consciousness, scores are considered only once for music category. Therefore the values represented in Table 1 are not the total number of likes in the dataset but the number of people with at least one like in a particular category.

Although the number of likes a user has in a particular category is omitted, having lots of likes in general can be behavioral characteristics of a person. Therefore, count of likes (like_count) a user has regardless of the category is considered as one of the characteristics of a user. As a result, SWL, IQ, openness, extraversion, neuroticism, agreeableness, consciousness, like_count characteristics of a user is analyzed as relate to user interests.

As an overall statistics of 1511 users, among the five personality traits, users score the highest on Openness with a mean value 4.1/5, and the lowest on neuroticism with a mean value 2.8/5, indicating that the dataset tend to have more open and less neurotic people. On the other hand, SWL and IQ results of the users tend to be in standard average.

**Table 1**. The Number of Users in Each Interest Category of Original Dataset

| Animal | Community | Food | Game | Health | Home | Music | Politics | Religion | Shop | Sports | Travel | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 252 | 1267 | 1021 | 1016 | 557 | 176 | 1427 | 532 | 391 | 579 | 846 | 555 | 8619 |

## 4. Experimental Setup

In order to generate association rules from the obtained set, an Apriori algorithm is used on WEKA (Waikato Environment for Knowledge Analysis) (Garner, 1995) open source tool. Getting meaningful results from Apriori algorithm requires defining the minimum support threshold. Therefore, in cases of significantly unbalanced classes of data, since the frequency of the itemset of a minority class is low, high minimum support value, prevents to obtain the rules related to the minority class. On the other hand, low minimum support value causes to have so many rules which might not be interesting.

As the dataset used in this research is combined by unbalanced classes, to prevent the bias towards majority classes, initially the dataset is balanced. Balancing of a dataset is mainly achieved by either removing instances from the majority classes which is called undersampling, or by adding similar instances to the minority class which is known as oversampling. In this project three WEKA filters; SMOTE, Spread Subsampling, and Resampling (Witten and Frank, 2000) are used to balance the data in three different ways. Therefore in the following sections the association rules generated by the three dataset are going to be discussed.

For oversampling, SMOTE (Chawla et al., 2002) synthetic oversampling technique which uses the Nearest Neighbour method to create new samples is used. Therefore the 12 classes are overbalanced to itemset with range [1327-1427] for each set. Spread Subsampling is used to undersample the classes into the minority class level 176 items per each. Resampling filter with bias is used for both oversampling and undersample of a dataset so that each class of the dataset is balanced to 718 items.

Once the three balanced itemsets are constructed, each dataset it passed through discretization process. In this step the numeric values of dataset is transferred into nominal attributes. The 8 feature and their range of values for the datasets are: SWL [0-7], IQ [70-140], like_count[1-5000], Openness [0-5], Neuroticism [0-5], Agreeableness [0-5], Extraversion [0-5], Consciousness [0-5]. Discretization process gives us an opportunity to sub-classify each feature to different bins. Therefore each 8 categories of features are subdivided into 3 bins and "category" feature which is provided as a label class preserved as it is (12 bins). The reason of sub-classifying each feature into 3 bins is to be able to find highest and lowest intervals in the sets to extract the valuable association. For example openness trait is divided into interval of (inf-3], (3-4] and (4, 5] which can be used to observe the less and more open users.

Upon completion of pre-processing steps, the data is fed to Apriori alogorithm for generation of the association rules which is then used to find the characteristics and interest of social network users.

## 5. Experiments and Results

In this study, through the experiments, the following association rules are extracted:

(i) The rules indicating the relationship between the 8 personal characteristics (SWL, IQ, openness, extraversion, neuroticism, agreeableness, consciousness, like_count) of users and 12 interest categories(Animals, Community, Food, Health, Home, Music, Politics , Travel , Games, Shopping, Sports, Religion).

(ii) The top rules generated by the dataset independent than the specific interest category.

All the experiments are run on three balanced datasets and the top results of each are combined together. In the experiments, only the valuable rules are analyzed. Therefore the rules which does not have all the features from the highest and lowest bins such as the rule "E='(2-4]' ⋀ A='(4-5]' ⋀ N='(2-4]' ⇒ C='(3-4]' " is not taken into consideration.

### 5.1. Association Rules Relating to User Characteristics and Category Preferences

In order to find the associations from the eight personality characteristics to the twelve interest categories (class labels), "class association rule" parameter in Apriori algorithm is set to true. In the experiments, the threshold for minimum support value is set to 2% and the minimum confidence value is set to 5%. Although these thresholds seems to be relatively low, in theory since there are 12 classes, there is no possibility for any interest category to be generated with more than 8% minimum support value. The number of rules to be generated is set to 300. From the generated set of rules, only the valuable result of the top 5 rules for each category of interest is analyzed.

Table 2 shows the top 5 valuable 1-itemset rules generated by the three balanced sets. The columns of the table correspond to antecedent, and the rows represent the consequent (interest categories). In the table, "-" indicates that no important association rules generated between the antecedent and consequent. The reverse relationship from the antecedent the consequent is denoted by negative values. The degree of positivity and negativity is increased based on the generation of a rule from more than one dataset. For example "3" indicates that antecedent positively imply the consequent and the rule is produced by all the datasets.

Among the produced rules of three datasets, no contradiction is observed. As many of the users in the dataset tend to be open, openness feature by itself does not seems to be have an effect on determination of the interest categories. The top rules generated by 3 datasets have the

**Table 2**. Top 1- itemset Valuable Rules of Three Balanced Datasets

| | SWL | IQ | Like_Count | Openness | Neuroticism | Agreeableness | Extravert | Consciousness |
|---|---|---|---|---|---|---|---|---|
| Commu-ty | - | - | -2 | - | 1 | - | - | - |
| Food | - | - | -2 | - | - | - | 1 | - |
| Game | - | 1 | -2 | - | - | - | 1 | - |
| Health | - | - | - | - | - | - | - | - |
| Music | - | - | -3 | - | 1 | - | - | - |
| Shopping | - | - | - | - | - | - | - | 1 |
| Sport | - | 2 | -2 | - | -1 | - | 1 | - |
| A-mal | -2 | - | 2 | - | - | - | - | - |
| Home | - | - | 2 | - | - | - | - | - |
| Religion | - | - | 1 | - | - | - | - | - |
| Travel | - | - | 1 | - | -1 | - | - | - |
| Politics | - | - | - | - | -1 | - | 1 | 1 |

biggest overlap with the rules implied by like_count feature. The like_count feature seems to be one of the top indicators of interest categories, despite it is not the most differentiating feature.

The remaining top 5 results of the datasets are obtained to be 2-itemset. From the 12 interest categories only 8 categories are implied by valuable 2-itemset rules. For the 4 categories (Game, Shop, Sport and Religion) no valuable 2-itemset rules are generated, therefore are discarded from the top list provided in Table 3.

**Table 3.** Top 2-itemset Valuable Rules of Three Balanced Datasets

| Association Rules | Confidence |
|---|---|
| O='(4-inf)' ∧ A='(4-inf) ⇒ ANIMAL | 0.12 |
| O='(4-inf)' ∧ SWL='(-inf-3]' ⇒ ANIMAL | 0.13 |
| O '(4-inf)' ∧ Like_Count '(-inf-142]' ⇒ FOOD | 0.11 |
| O='(4-inf)' ∧ E='(4-inf)' ⇒ HEALTH | 0.10 |
| O='(4-inf)' ∧ Like_Count='(340-inf)'⇒ HEALTH | 0.10 |
| O='(4-inf)' ∧ A='(4-inf)' ⇒ HEALTH | 0.10 |
| A '(4-inf)' ∧ Like_Count '(407-inf)' ⇒ HOME | 0.14 |
| O='(4-inf)' ∧ Like_Count='(340-inf)' ⇒HOME | 0.13 |
| IQ='(116-inf)' ∧ Like_Count='(-inf138]'⇒MUSIC | 0.15 |
| O='(4-inf)' ∧ Like_Count='(-inf-138]' ⇒ MUSIC | 0.14 |
| O='(4-inf)' ∧ N='(-inf-2]' ⇒ POLITICS | 0.12 |
| O '(4-inf)' ∧ IQ '(116-inf)' ⇒ POLITICS | 0.10 |
| O='(4-inf)' ∧ Like_Count='(340-inf)'⇒ TRAVEL | 0.10 |

Although openness is not found to be a direct determinant (1-itemset) of interest categories, it appear to be a positive effect on 2-itemset of the 8 categories. As expected, the high ranking of like_count is carried out to the 2-itemset rules. Therefore from the 1-itemset and 2-itemset rules of Apriori algorithm, the following conclusion about the interest categories (IC) can be deducted:

(i) IC1: Users who tend to have more likes are interested in A-mal, Home, Religion, and Travel
(ii) IC 2: Users who have less likes are more interested in Community, Food, Game, Music and Sports
(iii) IC 3: Users who has low satisfaction from life are tend to be interested in Animals
(iv) IC 4: Users with high IQ are more into Games and Sports
(v) IC 5: Agreeable users tent to be interested with Home Health and Animals more.
(vi) IC 6: Users who are less neurotic and more intelligent are more interested in Politics
(vii) IC 7: Open people more likely to get interested with Animals, Community, Food, Health, Home, Music, Politics and Travel
(viii) IC 8: Neurotic people tend to be interested in Community and Music; contrary less neurotic people are more into Sports Travel and Politics

## 5.2. Category Independent Association Rule Generation

Since the datasets includes user interest information in 12 categories, besides analyzing the relationship of personal characteristics with interest categories, it is interesting to observe the relationship among the various features regardless the interest category. Therefore, the top 10 results of the three dataset are selected to extract the main relationships among the all features. In the experiments of Spread subsampled and Resampled datasets, mi-mum support threshold is set to 0.07, and in Oversampled dataset mi-mum support threshold is set to 0.2. The confidence value for all three datasets is set to 0.7.

The association rules produced as a result of setting support and confidence values sometimes can be misleading, and do not show the dependence of the consequent to the antecedent. Therefore to observe the dependence of consequent we also run the tests with mi-mum support value of 0.1 and minimum conviction threshold value of 1.5. From the generated new rule set again the top 10 rules with the highest conviction values are selected.

**Table 4.** The Top Valuable Association Rules of Spread Subsampled and Resampled Datasets

| Association Rules | Confidence | Lift | Conviction | Support | Algorithm |
|---|---|---|---|---|---|
| E='(4-inf)' ∧ IQ='(116-inf)' ⇒ O='(4-inf)' | 0.80 | 1.36 | 2.04 | 0.09 | Both |
| C='(4-inf)' ∧ E='(4-inf)' ⇒ O='(4-inf)' | 0.75 | 1.26 | 1.57 | 0.07 | SS |
| E='(4-inf)' ∧ N='(-inf-2] ⇒ A='(4-inf)' | 0.73 | 1.63 | 2.02 | 0.11 | RS |
| E='(4-inf)' ⇒ O='(4-inf)' | 0.70 | 1.18 | 1.35 | 0.21 | SS |
| SWL='(5-inf)' ∧ N='(-inf-2]' ⇒ A='(4-inf)' | 0.67 | 1.51 | 1.68 | 0.10 | Both |
| C='(4-inf)' ⇒ O='(4-inf)' | 0.65 | 1.1 | 1.16 | 0.17 | SS |
| N='(-inf-2]' ⇒ A='(4-inf)' | 0.64 | 1.43 | 1.52 | 0.19 | Both |
| SWL='(5-inf)' ∧ A='(4-inf)' ⇒ N='(-inf-2]' | 0.64 | 2.2 | 1.94 | 0.10 | Both |
| E='(4-inf)' ∧ A='(4-inf)' ⇒ N='(-inf-2]' | 0.61 | 2.09 | 1.83 | 0.11 | RS |
| A='(4-inf)' ∧ N='(-inf-2]' ⇒ E='(4-inf)' | 0.58 | 1.9 | 1.65 | 0.11 | RS |
| A='(4-inf)' ∧ N='(-inf-2]' ⇒ SWL='(5-inf)' | 0.57 | 1.8 | 1.58 | 0.10 | Both |
| N='(-inf-2]' ⇒ SWL='(5-inf)' | 0.56 | 1.75 | 1.53 | 0.16 | SS |
| O='(4-inf)' ∧ N='(-inf-2]' ⇒ E='(4-inf)' | 0.56 | 1.83 | 1.57 | 0.11 | Both |
| N='(4-inf)' ⇒ SWL='(-inf-3]' | 0.55 | 2 | 1.6 | 0.10 | Both |

Therefore from each set, 20 rules which has the top confidence and top conviction values are selected. However the top 20 rules generated by the oversampled dataset did not produce any valuable rules. Table 4 presents the top valuable association rules of the two datasets with performance measure of confidence, lift, conviction and support. The valuable rules generated by Resampled(RS) and Spread Subsampled(SS) sets seems to be similar as half of the selected rules are produced by both datasets. Non-common top rules seem to be in different itemsets as 1-itemsets and 2-itemsets.

As the conviction value emphasizes the dependence of consequent on antecedent, as expected, 1-itemset rules has lower dependence to the antecedent than the 2-itemsets, therefore relatively their conviction is lower. Similarly no interesting rules are inferred for IQ, consciousness and like_count features indicating that they are not implied from the extreme characteristics of users. As explained in the previous sections, since interest categories have low support and confidence threshold, they are not observed in the top results of these experiments.

Based on the generated rules, the following general conclusions (GC) can be deducted about the characteristics of social network users:

(i) GC1: Users who are extraverts with high IQ tend to be open.
(ii) GC2: Although being extravert and conscious are the indicators of openness, both extravert and conscious users tend to be more open.
(iii) GC3: There is a strong relationship between high life satisfaction, agreeableness and being less neurotic. Therefore it is observed that when any two of the features come together the third one can be implied.
(iv) GC4: There is a strong relationship between extraversions, agreeableness and being less neurotic. Therefore it is observed that when any two of the features come together the third one can be implied.
(v) GC5: High neuroticism observed to be one of the strongest indicators of low life satisfaction.
(vi) GC6: It is also a strong conclusion that open and less neurotic users are extraverts.

## 6. Conclusion and Future Work

Today competitions among traders make it necessary to find the new ways of satisfying user needs. Therefore extracting personal information becomes a need for recommender systems to feed online advertising. In this study we analyzed the association between the personal characteristics of users and 12 interest categories on myPersonality dataset. From the rules, it is observed that Extraversion and Consciousness tend to be least determinant features of 12 categories, on the other hand, user likes_count tend to be one of the most frequent feature of the top association rules. The users interested in Game, Shop, Sport and Religion categories are tend to have an average personal characteristics compared to the other 8 categories.

In the study the relationship among the personality characteristics of a dataset is also observed. In the top generated rules, a strong relationship is found among extraversion, agreeableness, high life satisfaction and less neuroticism. However the most valuable features, SWL, IQ, Like_Count of interest categories did not produce valuable relations with personality traits.

The above rules are generated by run-ng Apriori algorithm on three balanced dataset (oversampled, resampled, spread sampled) of the initial set. It is observed that the top generated rules of the 3 datasets did not produce any

contradicting rules but despite had considerable overlaps. When considering the interest categories it is observed that oversampled dataset is biased towards less like_count while spread sample set is biased for openness. In general the results of the 3 balanced sets produced the almost same rules in different ranking order.

In this paper the interest categories are analyzed with 8 personal characteristics of a person. Thorough the experiments we observed that there are many more personal characteristics of a user (e.g., illnesses, childhood) which can be used as an indicator of different interest categories. Therefore, in the availability of a data, the user interest can be analyzed in a bigger set with grouped set of interests.

## References


Agrawal, R., Imielinski, T., Swami, A. (1993, May). Mi-ng associations between sets of items in massive databases. *In Proc. 1993 ACM-SIGMOD Int. Conf* (pp. 207-216).

Alam, F., Stepanov, Evgeny A.; Rıccardı, Giusepp. (2013, June). Personality traits recog-tion on social network-facebook. *In Seventh International AAAI Conference on Weblogs and Social Media*.

Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., & Stillwell, D. (2012, June). Personality and patterns of Facebook usage. *In Proceedings of the 4th annual ACM web science conference* (pp. 24-32). ACM.

Brin, S., Motwa-, R., Ullman, J. D., & Tsur, S.S. (1997a). Dynamic itemset counting and implication rules for market basket data. *Acm Sigmod Record*, 26(2), 255-264.

Brin, S., Motwa-, R., Ullman, J. D., & Tsur, S.( 1997b) Dynamic itemset counting and implication rules for market basket data. *In SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data*, pages 255-264, Tucson, Arizona, USA,

Brusilovski, P., Alfred K., and Wolfgang N. (2007). The adaptive web: methods and strategies of web personalization (Vol. 4321). *Springer Science & Business Media*.

Cantador, I. Fernández-Tobías, I., Bellogín, A. (2013). Relating personality types with user preferences in multiple entertainment domains. *In CEUR workshop proceedings.*

Celli, F., Pianesi, F., Stillwell, D., & Kosinski, M. (2013, June). Workshop on computational personality recog-tion: Shared task. *In Seventh International AAAI Conference on Weblogs and Social Media*.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling tech-que. *Journal of artificial intelligence research*, 16, 321-357.

Costa, P.T., McCrae, Robert. R. (2008). The revised neo personality inventory (neo-pi-r*)." The SAGE handbook of personality theory and assessment* 2, 2(2), 179-198.

Digman, J.M. (1989). Five robust trait dimensions: Development, stability, and utility. *Journal of personality*, 57(2), 195-214.

Farnadi, G., Zoghbi, S., Moens, M. F., & De Cock, M. (2013, June). Recog-sing personality traits using Facebook status updates. *In Seventh International AAAI Conference on Weblogs and Social Media.*

Garner, S. R. (1995, April). Weka: The waikato environment for knowledge analysis. In Proceedings of the *New Zealand computer science research students conference* (pp. 57-64).

Golbeck, J., Robles, C. T. (2011, May). Predicting personality with social media. *In CHI'11 extended abstracts on human factors in computing systems* (pp. 253-262). ACM.

Goldberg, L.R., Johnson, J.A., Eber, H.W., Hogan, R., Ashton, M.C., Clo-nger, C.R., & Gough, H.G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in personality*, 40(1), 84-96.

HU, R., PU, P. (2011, October). Enhancing collaborative filtering systems with personality information. In *Proceedings of the fifth ACM conference on Recommender systems* (pp. 197-204). ACM.

Kaufman, S.B., Quilty, L. C., Grazioplene, R.G., Hirsh, J. B., Gray, J. R., Peterson, J. B., & DeYoung, C. G. (2016). Openness to experience and intellect differentially predict creative achievement in the arts and sciences. *Journal of personality*, 84(2), 248-258.

Kosinski, M., & Stillwell, D. J. (2011). myPersonality research wiki. myPersonality project. Unpublished manuscript.

Markovikj, D., Gievska, S., Kosinski, M., & Stillwell, D. J. (2013, June). Mi-ng facebook data for predictive personality modeling. *In Seventh International AAAI Conference on Weblogs and Social Media.*

Pennebaker, J. W., Francıs, M. E., Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001)

Pennebaker, J.W.; King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, 77(6), 1296.

Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. *Knowledge discovery in databases*, 229-238.

Roshchina, A. (2012). TWIN: Personality-based Recommender System. *Institute of Technology Tallaght, Dublin.*

Schacter, D. L., Gilbert, D. T., & Wegner, D. M. (2010). Implicit memory and explicit memory. *Psychology*, 238.

Schimmack, U., Radhakrishnan, P., Oishi, S., Dzokoto, V., & Ahadi, S. (2002). Culture, personality, and subjective well-being: Integrating process models of life satisfaction. *Journal of personality and social psychology*, 82(4), 582.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., & Ungar, L. H. (2013). Personality, gender, and age in the language of social

media: The open-vocabulary approach. *PloS one*, 8(9), e73791.

Sternberg, R.J., Kaufman, Scott B. (2011). *The Cambridge handbook of intelligence. Cambridge U-versity Press.*

Witten I.H., Frank E.( 2000). Data Mi-ng: Practical Machine Lear-ng Tool and Tech-que with Java Implementation. *Morgan Kaufmann*; 2000.

Yarko-, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*, 44(3), 363-373.

Zhan L., Sun Y., Wang, N., & Zhang, X. (2016). Understanding the influence of social media on people's life satisfaction through two competing explanatory mecha-sms. *Aslib Journal of Information Management*, 68(3), 347-361.