# Exploring country level gender differences in the context of online dating using classification trees

Slava Kisilevich and Mark Last

Department of Computer and Information Science
Konstanz University
slaks@dbvis.inf.uni-konstanz.de[* * *]
Department of Information System Engineering
Ben-Gurion University of the Negev
mlast@bgu.ac.il[†]

**Abstract.** The key component of Social Networking Sites (SNS), gaining increasing popularity among Internet users, is the user profile, which plays a role of a self-advertisement in the aggregated form. While computer scientists investigate privacy implications of information disclosure, social scientists test or generate social or behavioral hypotheses based on the information provided by users in their profiles. Statistical analysis of the SNS phenomenon often is performed using only a very small sample of information extracted from a particular SNS or by interviewing students from a particular university. In this paper, we apply classification algorithm to a large-scale SNS dataset obtained from more than 10 million public profiles with 50 different attributes extracted from one of the largest dating sites in the Russian segment of the Internet. In particular, we build gender classification models for the residents of the most active countries, and investigate the particular differences between genders in one country and the differences between the same-genders in different countries. The preliminary results are reported in this paper. To the best of our knowledge, this is the first attempt to conduct a large-scale analysis of SNS profiles and compare gender differences on a country level.

**Key words:** Social Networking Sites, Self-disclosure, Gender differences, Classification trees

## 1 Introduction

Rapid technological development of the Internet in recent years and its worldwide availability has changed the way people communicate with each other. Social Networking Sites such as Facebook or MySpace gained huge popularity worldwide, having hundreds of millions of registered users. A major reason

[* * *] http://www.informatik.uni-konstanz.de/arbeitsgruppen/infovis/mitglieder/slava-kisilevich/

[†] http://www.bgu.ac.il/~mlast/

for the increased popularity is based on social interaction, e.g. networking with friends, establishing new friendships, creation of virtual communities of mutual interests, sharing ideas, open discussions, collaboration with others on different topics or even playing games. The key component of SNS is the user profile, in which the person cannot only post personal data, e.g. name, gender, age, address, but also has the opportunity to display other aspects of life, such as personal interests, (hobbies, music, movies, books), political views, and intimate information. Photos and videos are equally important for a self-description. All SNS allow the user to upload at least one photo. Most mainstream SNS also feature video uploading.

Various research communities have realized the potential of analysis of the SNS phenomenon and its implication on society from different perspectives such as law [1], privacy [2–4], social interaction and theories [5–9]. Many hypotheses and social theories (gender and age differences, self-disclosure and self-presentation) have been raised and tested by social scientists using the context of Social Networks. Statistical analysis is the widely used instrument for analysis among social scientists and rely on the sampling rather than on data collected from an entire population segment. The common approach to perform Social Network analysis is to analyze a sample of user profiles or to conduct a survey among students (usually less than 100) of a particular university by presenting descriptive statistics of the sample data and performing significance tests between dependent variables [2, 10, 3, 8]. The major drawback of such approach with respect to Social Networks is that in light of the large population of SNSs, which can vary from tens to hundreds million users, the results of the statistical analysis cannot be generalized for the whole population and theories can hardly be validated using only small samples. Moreover, Social Networks are heterogeneous systems, with people living in different parts of the world. To the best of our knowledge, the state of the art Social Science research of Social Networks does not take into account the spatial characteristics of the population. For example, due to cultural differences, the theory of self-disclosure tested on students from American universities may be not valid if applied on information obtained from students of Chinese universities, even if both groups use the same Social Network. Although, the problem and the importance of space and place in the Social Sciences was already highlighted a decade ago [11], this knowledge gap was not closed until do date. Therefore, in order to improve our understanding of social behavior, to analyze, to find hidden behavioral patterns not visible at smaller scales, and to build new theories of large heterogeneous social systems like Social Networks, other approaches and computational techniques should be applied [12].

In this paper, we answer the following hypothetical question: “Can we find some hidden behavioral patterns from user profiles in the large-scale SNS data beyond mere descriptive statistics.”

We answer this question by applying a classification algorithm to the data obtained from more than 10 million profiles having more than 50 different attributes extracted from one of the largest dating site in the Russian segment of the Internet. Specifically, we build gender classification models for most active countries and investigate what are particular differences between genders in one country and what are the differences between same-genders in different countries. Dating sites can be considered as a special type of social networks where members are engaged in development of romantic relationship. Information revealed in the users' profiles is an important aspect for the assessment of potential communication, for maximizing the chances for online dating for the owner of the profile, and the minimizing of risks of online dating for the viewer of the profile. For this reason, in the broad context, assuming that the goal of the member of the dating site is to find a romantic partner, we investigate patterns of self-presentation that can vary from country to country and differ for both genders.

The preliminary results suggest that the classification model can successfully be used for analysis of gender differences between users of SNS using information extracted from user profiles that usually contain tens of different categorical and numerical attributes.

Comparing gender differences on a country level as well as using data mining approaches in the Social Science context, is to the best of our knowledge, the first attempt to conduct a large-scale analysis of SNS profiles.

## 2 Related Work

Gender differences have been studied long before the Internet became widely available. However, with the technological development of the Internet and proliferation of Social Networks, the research has focused on the analysis of online communities and differences between their members. Many studies were performed in the context of Internet use [13, 14], online relationships [5], ethnic identity [8], blogging [10], self-disclosure and privacy [2–4]. Since we could not find any related work on large-scale analysis of gender differences in social networks, we are going to review some of the recent studies and findings about gender differences in general.

Information revelation, privacy issues and demographic differences between users of Facebook SNS were examined in [2] and [3]. [2] interviewed 294 students and obtained their profiles from Facebook. The goal of the survey was to assess the privacy attitudes, awareness of the members of the SNS to privacy issues, and the amount and type of information the users reveal in their profiles. It was found that there is no difference between males and females with respect to their privacy attitudes and the likelihood of providing certain information. Likewise, there is no difference between genders in information revelation. If some information is provided, it is likely to be complete and accurate. However, female students are less likely to provide their sexual orientation, personal address and cell phone number. [3] interviewed 77 students to investigate different behavioral aspects like information revelation, frequency of Facebook use, personal network size, privacy concerns and privacy protection strategies. Again, there were almost no difference between female and male respondents in the amount and type of the information revealed in their profiles. [4] analyzed about 30 million profiles from five social networks of Runet and conducted a survey among Russian speaking population to cross-check the finding extracted from the profiles and assess privacy concerns of members of Russian social networks. It was shown that there are differences between type of revealed information between females and males and these differences conditioned on the reported country of residence (20 most populous countries were presented). Particularly, males disclose more intimate information regardless of their country of origin. However, the country with the highest difference in the amount of disclosed intimate information was Russia (20.67%) and the lowest was Spain (5.59%). In addition, females from 17 countries revealed more information about having or not having children, economic and marital status, and religion. The only exceptions were females in Russia, Israel and England.

Social capital divide between teenagers and old people, and similarities in the use of the SNS were studied in [7] using profiles from MySpace social network. The results of the analysis indicate, among other criteria, that female teenagers are more involved in the online social interaction than male teenagers. Likewise, statistical tests showed that older women receive more comments than older men. Additionally, linguistic analysis of user messages showed that females include more self-descriptive words in their profiles than males. Friendship connections, age and gender were analyzed in [6] using $15,043$ MySpace profiles. The results showed that female members have more friends and are more likely interested in friendship than males, but males are more likely to be interested in dating and serious relationships. In the study that analyzed emotions expressed in comments [9], it was found that females send and receive more emotional messages than males. However, no difference between genders was found with respect to negative emotions contained in messages.

Online dating communities are typically treated differently because goals of the dating sites are much more limited in terms of connection development and often bear intimate context, which for the most part shifts to the offline context. Issues such as honesty, deception, misrepresentation, credibility assessment, and credibility demonstration, are more important in the dating context than in the context of general purpose social networks. Researchers are particularly interested in the analysis of self-presentation and self-disclosure strategies of the members of dating sites for achieving their goal to successfully find a romantic partner. [5] interviewed 349 members of a large dating site to investigate their goals on the site, how they construct their profiles, what type of information they disclose, how they assess credibility of others and how they form new relationships. The study found that cues presented in the users' profiles are very important for establishing connections. These cues include very well-written profiles, lack of spelling errors and uploaded photos. The last time the user was online considered to be one of the factors of reliability. Most of the respondents reported that they provide accurate information about themselves in the profiles.

## 3 Data

The data used in this paper was collected from one of the biggest dating sites in Runet: *Mamba*[1]. According to the site's own statistics (June 3, 2010), there are $13,198,277$ million registered users and searchable $8,078,130$ profiles. The main features of the service is the user profile and search option that allows searching for people by country, gender, age and other relevant attributes. The friend list is discrete, so other registered users cannot know with whom a user is chatting. The friend list is implicitly created when the user receives a message from another user. There are no means to block unwanted users before they send a message. However, users with the specially paid for VIP-status may get messages only from other VIP users. The user may exclude his/her profile to be searchable, but most of the profiles are searchable and accessible to unregistered users.

The user profile consists of seven sections also called blocks, where every block can be activated or deactivated by the user. Table 1 shows the names of sections and attribute parameters available in every section. We excluded the *About me* section, in which the user can describe himself in an open form, some intimate attributes of the *Sexual preference* section and the option to add multimedia (photos or videos) The attributes are divided into two categories. In the first category, only one value can be selected for the attribute (denoted as “no” in the Single selection column), other attributes contain multiple selections (denotes as “yes” in the Single selection column). Most of the attributes also contain an additional open field that allows the user to provide his/her own answer. The user can extend his/her main profile by filling two surveys. The one survey is provided by MonAmour site[2], owned by Mamba and contains about 100 different questions that estimate the psychological type of the respondent according to four components scaled from 0 to 100: *Spontaneity, Flexibility, Sociability, Emotions.* Another survey is internal and contains 40 open questions like *Education, Favorite Musician, etc.* In addition, the user can provide additional information about himself/herself to assure that he/she is a real person. For this, he/she should send a free SMS to the company and confirm his/her mobile number.

In order to collect the data, we developed a two-pass crawler written in C#. In the first pass the crawler repeatedly scans all searchable users which results in a collection of a basic information about the user such as *user id, profile URL, number of photos in the profile, and country and city of residence.* In the second pass, the crawler downloads the user's profile, checks if it is not blocked by the service provider and extracts all the relevant information, which is described in Table 1 including fields of the internal survey.

In a two month period, between March and June 2010, we extracted information from 13,187,295 millions users, where 1,948,656 million profiles were blocked, leaving us with 11,238,639 million valid profiles.

[1] http://www.mamba.ru/

[2] http://www.monamour.ru/

**Table 1.** Profile sections and attributes

| Section | Attributes | # of options | Single selection | Example |
|---|---|---|---|---|
| Personal | Age | - | yes | 20 |
| | Gender | 2 | yes | Male |
| | Zodiac | 12 | yes | Capricorn |
| Acquaintances | Seek for | 5 | no | Seek for a man of age 16-20 |
| | Aim | 13 | no | Friendship and chatting |
| | Marriage | 5 | yes | Married and live together |
| | Material support | 4 | yes | I am ready to become a sponsor |
| | Kids | 5 | yes | I have kids, we live together |
| Type | Weight | 1 | yes | 70 kg. |
| | Height | 1 | yes | 180 cm. |
| | Figure | 8 | yes | Skinny |
| | Body has | 2 | no | Tattoo, Piercing |
| | Hair on the head | 8 | yes | Light colored |
| | Hair on the face or body | 8 | no | Chest, Hands |
| | Profession | - | - | Open field |
| | Day regimen | 3 | yes | I get up early |
| | Languages | 87 | no | English, German |
| | Economic conditions | 5 | yes | Wealthy |
| | Dwelling | 7 | yes | I live with my parents |
| | Life priorities | 8 | no | Carrier, Wealth, Family |
| Interests | Leisure | 14 | no | Reading, Sport, Party |
| | Interests | 19 | no | Science, Cars, Business |
| | Sports | 12 | no | Fitness, Diving |
| | Music | 11 | no | Rock, Rap |
| | Religion | 7 | no | Christianity, Atheism |
| | Smoking | 5 | no | I rarely smoke |
| | Alcohol | 4 | no | I like to drink |
| | Drugs | 9 | no | I never tried |
| Car | Car | 76 | yes | Nissan |
| Mobile | Mobile | 50 | yes | Ericsson |
| Sexual preferences | Orientation | 4 | yes | Hetero, Bi |
| | Heterosexual experience | 6 | yes | Yes, we lived together |
| | Frequency | 7 | yes | At least once a day |
| | Excitement | 14 | no | Smells, latex, tattoos |

## 4 Methodology

In this section we describe the data mining process that includes data selection, data transformation and model construction.

### 4.1 Data selection

The data preparation and selection is very crucial for the data mining process. If sampled data is not a good representation of the whole dataset, the data mining process will fail to discover the real patterns. Another aspect of data preparation is related to user profiles. As was already discussed in Sections 1 and 2, the ultimate goal of members of the dating site is to find a romantic partner. Since this kind of activity may involve elements of intimacy, persons employ different strategies to balance the desire to reveal information about themselves and stay anonymous (for example, the profile without a photo). Moreover, many people may run several user profiles for different purposes.

In order to minimize the impact of fake profiles on the pattern mining, we employed a four level filtering process. First, the profiles of persons who filled the external survey on the MonAmour site (described in Section 3) were retrieved. Since the respondent should answer about 100 questions, it is unlikely that the person has non-serious intentions on the dating site. Second, we retrieved profiles who filled additional external survey that includes about 40 questions. Next, the users with the status "real" were retrieved and finally, the users who uploaded at least one photo and no more than one hundred photos were extracted. Table 2 shows the demographic statistics by country and gender. It also shows how many profiles were selected for mining and the resulted percentage of females and males in the selected instances. The selected age range was 16 to 50. Due to the large number of profiles in Russia, we extracted no more than $20,000$ profiles for every age value and gender on every filtering step.

### 4.2 Data transformation

Almost all the attributes described in Table 1 were selected for inclusion into the model (except for *Weight* and *Height*). Numerical attributes include age, number of photos and number of words, whose length is more than two, used in the "About me" section. Attributes such as *Figure, Music, Car* or *Body has* whose values are not important for classification but only the fact of their presence or absence, were encoded as binary attributes: if the person provided information about his figure, it was coded as binary *True*, otherwise it was treated as *False*. On the other hand, attributes, whose values are relevant for classification were encoded as multi-valued categorical attributes. For example, the *Marriage* attribute has four explicit options (*I am married, we live together; I am married, we do not live together; I have a fictional marriage; No, I am not married*) and one implicit *no answer*. In this case the four options were coded like *1,2,3,4*, while in the case of implicit answer it was treated as a missing value. Another group of attributes that may take more than one value (when the user chooses more than one answer) was decomposed into separate binary attributes representing distinct answer categories. For example, the user can provide 13 different answers related to his aim on the site (*Aim* attribute). These 13 answers are categorized into six categories: *Friendship, Love, Sex, Sex for Money, Marriage* and *Other*. In this case, if the person provided his answer on the question from the

**Table 2.** Demographic statistics of the 20 most active countries and statistics related to the sampled data

| **Country** | **Total** | **Males %** | **Females %** | **# instances** | **Sampled Males %** | **Sampled Females %** |
|---|---|---|---|---|---|---|
| Russia | 7,999,976 | 35 | 65 | 1,332,563 | 40 | 60 |
| Ukraine | 1,294,260 | 48 | 52 | 813,322 | 17 | 83 |
| Kazakhstan | 473,561 | 43 | 57 | 222,579 | 18 | 82 |
| Belarus | 328,029 | 55 | 45 | 264,131 | 20 | 80 |
| Germany | 129,732 | 57 | 43 | 71,586 | 16 | 84 |
| Azerbaijan | 107,125 | 81 | 19 | 20,183 | 35 | 65 |
| Uzbekistan | 89,709 | 78 | 22 | 26,788 | 27 | 73 |
| Moldova | 84,306 | 59 | 41 | 54,561 | 15 | 85 |
| Armenia | 70,362 | 58 | 42 | 12,308 | 41 | 59 |
| Georgia | 69,805 | 80 | 20 | 18,163 | 26 | 74 |
| Latvia | 54,521 | 41 | 59 | 33,310 | 14 | 86 |
| Estonia | 49,030 | 48 | 52 | 27,991 | 16 | 84 |
| USA | 47,741 | 60 | 40 | 25,702 | 15 | 85 |
| Israel | 43,001 | 63 | 37 | 23,481 | 26 | 74 |
| England | 36,261 | 39 | 61 | 12,525 | 20 | 80 |
| Lithuania | 35,270 | 41 | 59 | 17,243 | 14 | 86 |
| Turkey | 35,230 | 84 | 16 | 9,003 | 26 | 74 |
| Kyrgyzstan | 35,107 | 64 | 36 | 16,263 | 21 | 79 |
| Italy | 18,681 | 58 | 42 | 12,495 | 10 | 90 |
| Spain | 18,619 | 61 | 39 | 9,919 | 12 | 88 |

*Friendship* category, a binary *True* is assigned to that attribute, otherwise *False* is assigned. Two binary attributes that were composed from the *Seek for*, namely *Seek for a man* and *Seek for a woman* were removed since they are found in the majority of profiles, highly correlated with the opposite gender and trivial in terms of gender classification.

### 4.3 Model construction

Our hypothesis is that specific gender differences exist on the country level as well as there are differences between the same-genders in different countries. The differences should be expressed in specificity of attributes and values that describe the gender. In other words, we hypothesize that profiles of females and males living in the same country have unique characteristics, which determine the gender of the owner of the profile. In addition, we hypothesize that, although the main characteristic of the users of the featured dating site is Russian language, cultural differences impact the characteristics of user profiles even for people of the same gender. The data mining process that can capture unique characteristics of the genders is a decision tree learning, which is based on model construction using input variables and prediction of the target class value (gender in our case).

We applied C4.5, a popular decision tree induction algorithm on the sampled data for every country with the *gender* as a binary class attribute, using Weka data mining package [15]. We set the minimum number of instances per leaf to 10 and left all other options in their default state (pruned decision tree, 0.25 pruning confidence factor). Table 3 shows, for every country, the total number of rules generated by the algorithm, the number of rules per gender, the number of frequent rules (the rules that classify more than 100 instances) and the number of rules that cover more than 90% of the sampled data.

**Table 3.** The total number of rules generated by country, gender, the number of rules that classify more than 100 individuals (Frequent Rules), the number of rules that cover 90% of the instances in the sampled dataset

| **Country** | **Rules** | **Male** | **Female** | **Frequent Rules Male** | **Frequent Rules Female** | **90% Rule Coverage Male** | **90% Rule Coverage Female** |
|---|---|---|---|---|---|---|---|
| Russia | 7,462 | 3,747 | 3,715 | 619 | 688 | 1135 | 732 |
| Ukraine | 2,957 | 1,458 | 1,499 | 151 | 313 | 666 | 116 |
| Kazakhstan | 1,075 | 513 | 562 | 47 | 113 | 251 | 68 |
| Belarus | 1,372 | 654 | 718 | 64 | 143 | 340 | 92 |
| Germany | 429 | 200 | 229 | 14 | 49 | 128 | 39 |
| Azerbaijan | 221 | 101 | 120 | 14 | 14 | 50 | 38 |
| Uzbekistan | 191 | 96 | 95 | 15 | 20 | 47 | 19 |
| Moldova | 250 | 119 | 131 | 12 | 25 | 68 | 15 |
| Armenia | 147 | 75 | 72 | 6 | 8 | 39 | 25 |
| Georgia | 151 | 74 | 77 | 9 | 11 | 42 | 13 |
| Latvia | 177 | 79 | 98 | 4 | 23 | 55 | 17 |
| Estonia | 205 | 48 | 91 | 3 | 28 | 62 | 26 |
| USA | 175 | 77 | 98 | 5 | 23 | 52 | 20 |
| Israel | 242 | 114 | 128 | 11 | 23 | 64 | 34 |
| England | 95 | 39 | 56 | 4 | 15 | 27 | 16 |
| Lithuania | 106 | 47 | 59 | 2 | 12 | 34 | 11 |
| Turkey | 91 | 46 | 45 | 3 | 5 | 30 | 11 |
| Kyrgyzstan | 104 | 51 | 53 | 8 | 11 | 29 | 10 |
| Italy | 70 | 41 | 29 | 1 | 11 | 20 | 8 |
| Spain | 67 | 27 | 40 | 0 | 8 | 20 | 5 |

## 5 Analysis

The purpose of this section is to analyze the data and the model described in Section 4. We apply a number of analytical steps to test our hypothesis that there are differences between genders and that these differences are also country-dependent.
The analytical steps are:
(1) Observation of the sampled data
(2) Observation of the quantity of rules that classify females and males

(3) Gender comparison
(4) Classification rules matching
(5) Gender characterization

### 5.1 Data observation

As was mentioned in Section 2, we applied four filtering steps to minimize the effect of false profiles. By inspecting the resulting number of females and males (Table 2), we can see the genders differences with respect to the profile creation. Many more females than males use different means of describing themselves through additional surveys, and many more females than males upload their photos. The largest difference between females and males can be observed in such countries as Italy (80%), Spain (76%), Latvia and Lithuania (72%), Moldova and USA (70%), while the smallest difference is in Armenia (18%), Russia (20%) and Azerbaijan (30%).

### 5.2 Model observation

The inspection of the quantity of generated rules that classify females and males (Table 3), shows that rules that classify females outnumber rules that classify males in 15 cases (countries), with the largest difference in Belarus. This finding may suggest that female users are more creative in profile construction and provide more heterogeneous information about themselves, while males use more homogeneous information to describe themselves. Moreover, the number of frequent rules is higher for females (19 cases) with the largest difference in Ukraine. This may also suggest the female users in different countries have more homogeneous behavior than men since they can be classified by relatively large amount of frequent rules. On the other hand, male users are heterogeneous with respect to the information they provide in their profiles, since most of them are classified by infrequent rules. This hypothesis is supported by inspecting how many rules cover the majority of the population. In all the cases, the number of rules that cover 90% of the population is larger for males with the greatest difference in Ukraine.

Any decision tree construction algorithm builds rules by determining the best attributes that build up the tree. The attribute at the root of the tree is the first attribute selected and, thus, is the best in the classification model. Inspection of the root attributes of the models reveals four groups of countries:
(1) Russia, Italy and Israel are characterized by the attribute *AimSex* (the aim on the site is to find a partner for having sex).
(2) Personality test (MonAmour test with more than 100 questions) is important for people from Azerbaijan and USA. This may suggest that people from Azerbaijan and USA consider the online dating as a very serious opportunity to find a romantic partner.
(3) Turkey is the only country where the classification tree is splitted according to the *Car* attribute. This may be explained by two reasons: (1) the number of

males is very high compared to the number of women and (2) most of the men like to “show off” by specifying what type of car they have.
(4) All other countries are characterized by the *photo* attribute that specifies how many photos were uploaded by a person.

### 5.3 Gender comparison

In Section 4.3 we applied a decision tree construction process to the user profiles from every country, and generated models that contain a number of rules that discriminate between females and males in a specific country. As mentioned already, classification trees are used for predicting the target class value. Usually, in order to estimate classifier's predictive performance, the model is evaluated on a separate test set. In the context of our analysis, we have evaluated the applicability of classification rules generated for each country to the data of other 19 countries. The high classification rate in this case should suggest that there is a high similarity between user profiles (and consequently between genders) across countries. We used 10-fold cross validation to estimate the testing accuracy of each country model on the data from the same country and used this result to report classification accuracy of the country's model. We selected 10% of profiles from the dataset of Russia and Ukraine using *StratifiedRemoveFolds* filter because Weka failed to run 10-fold cross validation on the entire dataset. Table 4 shows the classification accuracy for every model arranged in rows. The numbers on the diagonal represent the testing accuracy of each country model. The numbers arranged in columns represent classification accuracy of each of the country models on a test set of a given country. For simplicity of inspection, cells that have classification accuracy higher than 90% are colored in dark yellow, while cells that have classification accuracy less than 80% are colored in pink. It should be noted that the results are not symmetric. For example the classification accuracy of the Russian model on Moldova profiles is 79.93%, while it is 75.55% when Russian profiles are tested on the the Moldovan model. From the Table 4 we can see that the performance of most of the classifiers is around 80% to 90%. Classification accuracy of the Russian model on all other countries is below the average. Most classification models perform similar or even better on profiles from other countries than on profiles of their own countries, except for Azerbaijan and Armenia profiles, which have a lower performance with models of many other countries. Most accuracy differences in Table 4 were found statistically significant at the 99.9% confidence level.

The classification accuracy allows us to reason about cross-country prediction performance of each model, but it is not sufficient for comparing the countries behavioral patterns due to the non-symmetrical matrix (Table 4). The answer to the question “which countries are similar” is obtained by using the weighted Kappa statistic[3] [16–18], which is a measure of agreement between any two classifiers and defined as

[3] We calculated Kappa statistic using MedCalc statistical package `http://www.medcalc.be/`

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{1}$$

where P(A) is the proportion of times that the classifiers agree and P(E) is the proportional agreement expected by chance.

Table 5 shows the Kappa values between countries. The interpretation of Kappa values was adopted from [19]. Cells that denote a *very good agreement* (0.801-1.0) are colored in light red, *good agreement* (0.601-0.8) are colored in blue, *moderate agreement* (0.401-0.6) are colored in green and cells that represent *fair agreement* are colored in yellow. According to [19] values below 0.20 represent poor agreements. Accordingly, we assume that there is no agreement between classifiers when Kappa values are below 0.20.

We can see that Belarus and Ukraine are the only countries with a *very good* agreement (0.826), which indicates that behavioral patterns are very similar in these countries. Germany has the largest number of *good* agreements (four in total). Armenia and Lithuania are the countries that have the largest number of *moderate* agreements (six in total). Kazakhstan and Italy are the countries that have the largest number of *fair* agreements (seven in total). Kazakhstan, Belarus, Germany, Armenia, Latvia and Lithuania are the country with the most number of agreements (eleven in total). Russia, on the other hand, is the only country, which does not have any similarities to other 19 countries.

### 5.4 Rule matching

The gender classification rules that were generated for every model (Table 3) consist of the most important attributes and values that characterize females and males in a specific country. If we expect to have a similarity between same genders in different countries, then there should be a high number of similar rules found in the models. We performed the comparison of classification rules (rule matching) by taking every rule in a model, and searching for rules that have the same attributes and values in the precedent of the rule in any order in other models. For example the following two rules $A{=}x$ *AND* $B{=}y \rightarrow c$ and $B{=}y$ *AND* $A{=}x \rightarrow b$ match because they have common attributes ($A$ and $B$) in the precedent and those attributes take the same values.

Due to the space limitation we cannot provide the complete results of the comparison. However, it should be noted that the number of matching rules is very low. For example, the highest number of matching rules was observed between Latvia and England, and Germany and Israel (4 rules).

**Latvia-England:**

(1) *If there are no photos AND aim is sex → males*

and 3 rules that classify females:

(2) *If there is at least 1 photo AND aim is not sex AND have a car AND seek for a person older than 21 AND no kids AND no information about the body → females*

(3) *If there is at least 1 photo AND aim is not sex AND have a car AND seek for a person older than 21 AND have kids living together → females*
(4) *If there is at least 1 photo AND aim is not sex AND have a car AND seek for a person older than 36 AND have kids but live separately → females*

We can clearly see the differences between females and males on the example of the rules presented above. While males are characterized by intimate intention (to have sex) and lack of photos, females are characterized by availability of at least one photo in their profiles and the information regarding the desired age of the partner. It is possible that young female users who do not have children or those who have children search for a person older than 21, while older female users who have children not living in the same household, would like to meet a person older than 36.

Israel and Germany are another two countries that have four common rules. One of the precedents of the rule is the following:
*If there is at least 1 photo AND aim is sex AND personality test is filled AND sexual orientation is Bisexual AND no kids*
The German model classifies this rule as *females*, but the Israeli rule classifies this rule as *males*. It should be noted that except for this ambiguous classification due to specificity of sexual orientation, all other common rules are not ambiguous. This is a good indication that there is a consistency in common rules among the same genders across different countries.

### 5.5 Gender characterization

Since the space limitation does not allow us to present the whole list of rules generated for every gender and country, we provide a number of rule examples picked from the set of most frequent rules.
**Azerbaijan:** *If personality test is filled AND aim is sex → males*
**Azerbaijan:** *If personality test is not filled AND there is at least 1 photo AND have a car AND younger than 25 AND seek for a person older than 18 AND the body is slim → females*
**Belarus:** *If no photos AND aim is sex AND seek for a person younger than 22 → males*
**Belarus:** *If there is at least 1 photo AND aim is not sex AND no car AND personality test is not filled AND preferences in sex are provided AND seek for a person older than 21 AND have kids living together → females*
**Russia:** *If aim is sex AND seek for a person younger than 22 AND sexual orientation is Heterosexual AND older than 18 AND younger than 42 AND personality test is filled → males*
**Russia:** *If aim is not sex AND there is at least 1 photo AND seek for a person older than 25 AND older than 29 AND do not smoke AND have kids living together → females*

From these examples we can see that the sex component is present in the male rules (Azerbaijan, Belarus, Russia), photos are uploaded more by females (Belarus, Russia). In addition, the difference in age of the female and the person she seeks for is not significant (Azerbaijan, Russia), while males specify young females as their desired romantic partners (Belarus, Russia)

## 6 Conclusions

In this paper we investigated gender differences between countries in the context of dating sites using approaches from the field of Data Mining. We applied decision tree construction algorithm to the user profiles from 20 most active countries using more than 10 million profiles from one of the biggest dating sites in the Russian segment of the Internet. We analyzed the generated models and found countries where users behave similarly in terms of profile creation. However, the majority of countries are different from each other, which suggests that cultural aspects influences the way people behave in social networking sites. We also analyzed the induced classification rules and found almost no similarity between the same genders from different countries. This fact reinforces our hypothesis that cultural aspects influences behavior not only of different genders across countries but also of people of the same gender. We showed that social phenomena can be investigated using data mining methods if large quantities of data are available, and when statistical analysis alone is not enough for finding interesting patterns.

Our research overcomes the limitations of most previous studies, where the analysis was performed on small, non-representative and non-generalizable samples of the user population. However, some uncertainty is associated with the large-scale analysis of real profiles mined from a social networking site, since the analyst cannot verify the real purpose of profile creation (whether it has a serious intention or was created for fun). At this point, we assume that the majority of SNS users have real profiles that reflect their real self. Automated cleaning of profile data may be a subject of future research.

The preliminary results provided in the paper are encouraging. In our future work, we will apply more analytical methods to conduct all-embracing gender difference analysis and work closely with social scientists to test hypotheses that so far have been verified on very limited amounts of sampled data.

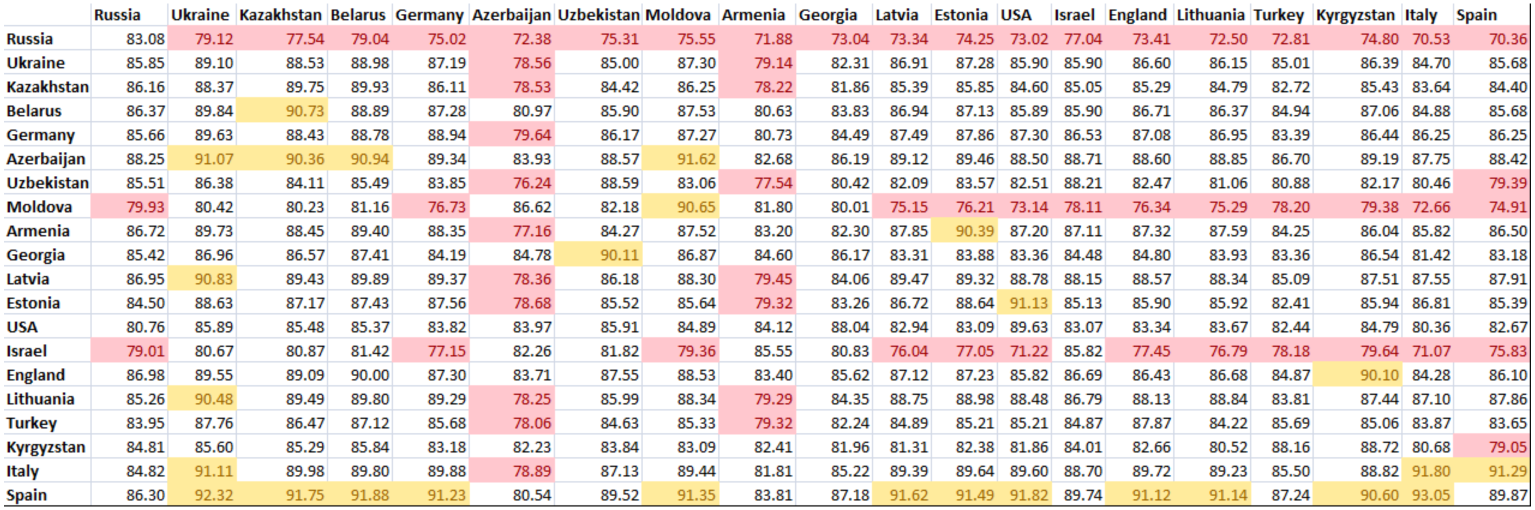
## Acknowledgements

This work was partially funded by the German Research Society (DFG) under grant GK-1042 (Research Training Group “Explorative Analysis and Visualization of Large Information Spaces”)

## References

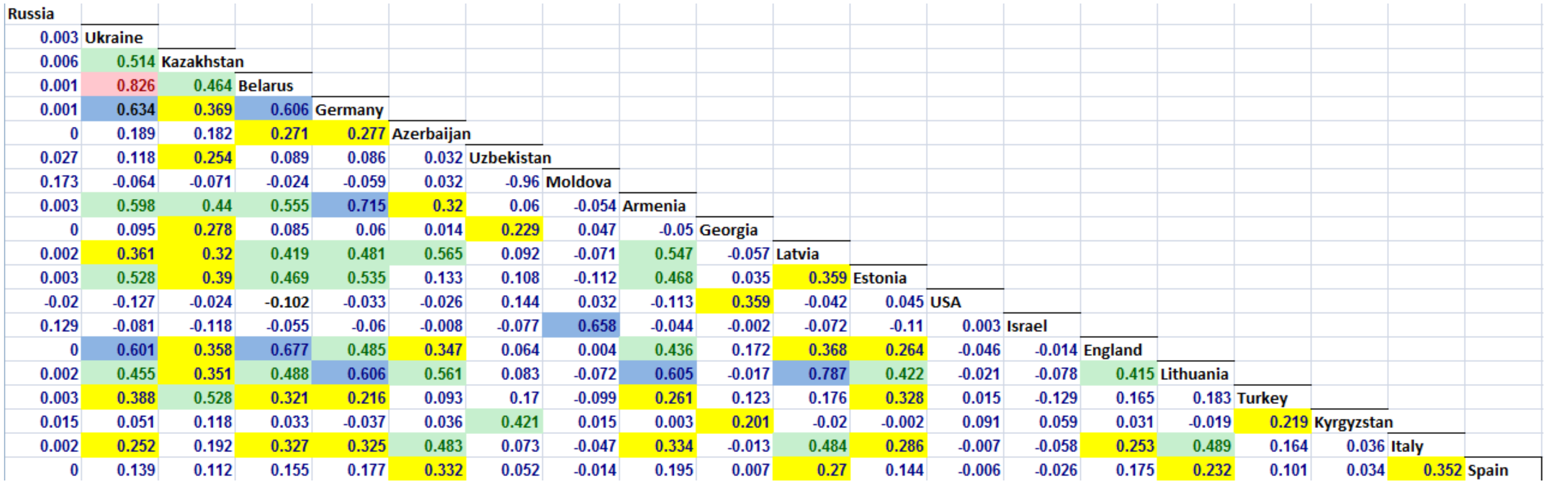
1. Nelson, S., Simek, J., Foltin, J.: The Legal Implications of Social Networking. Regent University Law Review **22**(1) (2009) 2
2. Acquisti, A., Gross, R.: Imagined communities: Awareness, information sharing, and privacy on the Facebook. In: Privacy Enhancing Technologies, Springer (2006) 36–58

3. Young, A., Quan-Haase, A.: Information revelation and internet privacy concerns on social network sites: a case study of facebook. In: Proceedings of the fourth international conference on Communities and technologies, ACM (2009) 265–274
4. Kisilevich, S., Mansmann, F.: Analysis of privacy in online social networks of Runet. In: Proceedings of the 3rd International Conference on Security of Information and Networks, ACM (2010)
5. Ellison, N., Heino, R., Gibbs, J.: Managing impressions online: Self-presentation processes in the online dating environment. Journal of Computer-Mediated Communication **11**(2) (2006) 415
6. Thelwall, M.: Social networks, gender, and friending: An analysis of MySpace member profiles. Journal of the American Society for Information Science and Technology **59**(8) (2008) 1321–1330
7. Pfeil, U., Arjan, R., Zaphiris, P.: Age differences in online social networking-A study of user profiles and the social capital divide among teenagers and older users in MySpace. Computers in Human Behavior **25**(3) (2009) 643–654
8. Grasmuck, S., Martin, J., Zhao, S.: Ethno-Racial Identity Displays on Facebook. Journal of Computer-Mediated Communication **15**(1) (2009) 158–188
9. Thelwall, M., Wilkinson, D., Uppal, S.: Data mining emotion in social network communication: Gender differences in MySpace. Journal of the American Society for Information Science and Technology (2009)
10. Pedersen, S., Macafee, C.: Gender differences in British blogging. Journal of Computer-Mediated Communication **12**(4) (2007) 1472
11. Goodchild, M., Anselin, L., Appelbaum, R., Harthorn, B.: Toward spatially integrated social science. International Regional Science Review **23**(2) (2000) 139
12. Kleinberg, J.: The convergence of social and technological networks. Commun. ACM **51**(11) (2008) 66–72
13. Golub, Y., Baillie, M., Brown, M.: Gender Differences in Internt Use and Online Relationships. American Journal of Psychological Research **3**(1) (2007)
14. Jones, S., Johnson-Yale, C., Millermaier, S., Pérez, F.: US College Students' Internet Use: Race, Gender and Digital Divides. Journal of Computer-Mediated Communication **14**(2) (2009) 244–264
15. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA data mining software: An update. ACM SIGKDD Explorations Newsletter **11**(1) (2009) 10–18
16. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1) (1960) 37
17. Cohen, J.: Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin **70** (1968) 213–220
18. Fleiss, J., Levin, B., Paik, M.: Statistical methods for rates and proportions. NY John Wiley & Sons (2003)
19. Altman, D.: Practical statistics for medical research. Chapman & Hall/CRC (1991)



| | Russia | Ukraine | Kazakhstan | Belarus | Germany | Azerbaijan | Uzbekistan | Moldova | Armenia | Georgia | Latvia | Estonia | USA | Israel | England | Lithuania | Turkey | Kyrgyzstan | Italy | Spain |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Russia** | 83.08 | 79.12 | 77.54 | 79.04 | 75.02 | 72.38 | 75.31 | 75.55 | 71.88 | 73.04 | 73.34 | 74.25 | 73.02 | 77.04 | 73.41 | 72.50 | 72.81 | 74.80 | 70.53 | 70.36 |
| **Ukraine** | 85.85 | 89.10 | 88.53 | 88.98 | 87.19 | 78.56 | 85.00 | 87.30 | 79.14 | 82.31 | 86.91 | 87.28 | 85.90 | 85.90 | 86.60 | 86.15 | 85.01 | 86.39 | 84.70 | 85.68 |
| **Kazakhstan** | 86.16 | 88.37 | 89.75 | 89.93 | 86.11 | 78.53 | 84.42 | 86.25 | 78.22 | 81.86 | 85.39 | 85.85 | 84.60 | 85.05 | 85.29 | 84.79 | 82.72 | 85.43 | 83.64 | 84.40 |
| **Belarus** | 86.37 | 89.84 | 90.73 | 88.89 | 87.28 | 80.97 | 85.90 | 87.53 | 80.63 | 83.83 | 86.94 | 87.13 | 85.89 | 85.90 | 86.71 | 86.37 | 84.94 | 87.06 | 84.88 | 85.68 |
| **Germany** | 85.66 | 89.63 | 88.43 | 88.78 | 88.94 | 79.64 | 86.17 | 87.27 | 80.73 | 84.49 | 87.49 | 87.86 | 87.30 | 86.53 | 87.08 | 86.95 | 83.39 | 86.44 | 86.25 | 86.25 |
| **Azerbaijan** | 88.25 | 91.07 | 90.36 | 90.94 | 89.34 | 83.93 | 88.57 | 91.62 | 82.68 | 86.19 | 89.12 | 89.46 | 88.50 | 88.71 | 88.60 | 88.85 | 86.70 | 89.19 | 87.75 | 88.42 |
| **Uzbekistan** | 85.51 | 86.38 | 84.11 | 85.49 | 83.85 | 76.24 | 88.59 | 83.06 | 77.54 | 80.42 | 82.09 | 83.57 | 82.51 | 88.21 | 82.47 | 81.06 | 80.88 | 82.17 | 80.46 | 79.39 |
| **Moldova** | 79.93 | 80.42 | 80.23 | 81.16 | 76.73 | 86.62 | 82.18 | 90.65 | 81.80 | 80.01 | 75.15 | 76.21 | 73.14 | 78.11 | 76.34 | 75.29 | 78.20 | 79.38 | 72.66 | 74.91 |
| **Armenia** | 86.72 | 89.73 | 88.45 | 89.40 | 88.35 | 77.16 | 84.27 | 87.52 | 83.20 | 82.30 | 87.85 | 90.39 | 87.20 | 87.11 | 87.32 | 87.59 | 84.25 | 86.04 | 85.82 | 86.50 |
| **Georgia** | 85.42 | 86.96 | 86.57 | 87.41 | 84.19 | 84.78 | 90.11 | 86.87 | 84.60 | 86.17 | 83.31 | 83.88 | 83.36 | 84.48 | 84.80 | 83.93 | 83.36 | 86.54 | 81.42 | 83.18 |
| **Latvia** | 86.95 | 90.83 | 89.43 | 89.89 | 89.37 | 78.36 | 86.18 | 88.30 | 79.45 | 84.06 | 89.47 | 89.32 | 88.78 | 88.15 | 88.57 | 88.34 | 85.09 | 87.51 | 87.55 | 87.91 |
| **Estonia** | 84.50 | 88.63 | 87.17 | 87.43 | 87.56 | 78.68 | 85.52 | 85.64 | 79.32 | 83.26 | 86.72 | 88.64 | 91.13 | 85.13 | 85.90 | 85.92 | 82.41 | 85.94 | 86.81 | 85.39 |
| **USA** | 80.76 | 85.89 | 85.48 | 85.37 | 83.82 | 83.97 | 85.91 | 84.89 | 84.12 | 88.04 | 82.94 | 83.09 | 89.63 | 83.07 | 83.34 | 83.67 | 82.44 | 84.79 | 80.36 | 82.67 |
| **Israel** | 79.01 | 80.67 | 80.87 | 81.42 | 77.15 | 82.26 | 81.82 | 79.36 | 85.55 | 80.83 | 76.04 | 77.05 | 71.22 | 85.82 | 77.45 | 76.79 | 78.18 | 79.64 | 71.07 | 75.83 |
| **England** | 86.98 | 89.55 | 89.09 | 90.00 | 87.30 | 83.71 | 87.55 | 88.53 | 83.40 | 85.62 | 87.12 | 87.23 | 85.82 | 86.69 | 86.43 | 86.68 | 84.87 | 90.10 | 84.28 | 86.10 |
| **Lithuania** | 85.26 | 90.48 | 89.49 | 89.80 | 89.29 | 78.25 | 85.99 | 88.34 | 79.29 | 84.35 | 88.75 | 88.98 | 88.48 | 86.79 | 88.13 | 88.84 | 83.81 | 87.44 | 87.10 | 87.86 |
| **Turkey** | 83.95 | 87.76 | 86.47 | 87.12 | 85.68 | 78.06 | 84.63 | 85.33 | 79.32 | 82.24 | 84.89 | 85.21 | 85.21 | 84.87 | 87.87 | 84.22 | 85.69 | 85.06 | 83.87 | 83.65 |
| **Kyrgyzstan** | 84.81 | 85.60 | 85.29 | 85.84 | 83.18 | 82.23 | 83.84 | 83.09 | 82.41 | 81.96 | 81.31 | 82.38 | 81.86 | 84.01 | 82.66 | 80.52 | 88.16 | 88.72 | 80.68 | 79.05 |
| **Italy** | 84.82 | 91.11 | 89.98 | 89.80 | 89.88 | 78.89 | 87.13 | 89.44 | 81.81 | 85.22 | 89.39 | 89.64 | 89.60 | 88.70 | 89.72 | 89.23 | 85.50 | 88.82 | 91.80 | 91.29 |
| **Spain** | 86.30 | 92.32 | 91.75 | 91.88 | 91.23 | 80.54 | 89.52 | 91.35 | 83.81 | 87.18 | 91.62 | 91.49 | 91.82 | 89.74 | 91.12 | 91.14 | 87.24 | 90.60 | 93.05 | 89.87 |

**Table 4.** Classification accuracy of 20 most active countries.



| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Russia | | | | | | | | | | | | | | | | | | | |
| 0.003 | Ukraine | | | | | | | | | | | | | | | | | | |
| 0.006 | 0.514 | Kazakhstan | | | | | | | | | | | | | | | | | |
| 0.001 | 0.826 | 0.464 | Belarus | | | | | | | | | | | | | | | | |
| 0.001 | 0.634 | 0.369 | 0.606 | Germany | | | | | | | | | | | | | | | |
| 0 | 0.189 | 0.182 | 0.271 | 0.277 | Azerbaijan | | | | | | | | | | | | | | |
| 0.027 | 0.118 | 0.254 | 0.089 | 0.086 | 0.032 | Uzbekistan | | | | | | | | | | | | | |
| 0.173 | -0.064 | -0.071 | -0.024 | -0.059 | 0.032 | -0.96 | Moldova | | | | | | | | | | | | |
| 0.003 | 0.598 | 0.44 | 0.555 | 0.715 | 0.32 | 0.06 | -0.054 | Armenia | | | | | | | | | | | |
| 0 | 0.095 | 0.278 | 0.085 | 0.06 | 0.014 | 0.229 | 0.047 | -0.05 | Georgia | | | | | | | | | | |
| 0.002 | 0.361 | 0.32 | 0.419 | 0.481 | 0.565 | 0.092 | -0.071 | 0.547 | -0.057 | Latvia | | | | | | | | | |
| 0.003 | 0.528 | 0.39 | 0.469 | 0.535 | 0.133 | 0.108 | -0.112 | 0.468 | 0.035 | 0.359 | Estonia | | | | | | | | |
| -0.02 | -0.127 | -0.024 | -0.102 | -0.033 | -0.026 | 0.144 | 0.032 | -0.113 | 0.359 | -0.042 | 0.045 | USA | | | | | | | |
| 0.129 | -0.081 | -0.118 | -0.055 | -0.06 | -0.008 | -0.077 | 0.658 | -0.044 | -0.002 | -0.072 | -0.11 | 0.003 | Israel | | | | | | |
| 0 | 0.601 | 0.358 | 0.677 | 0.485 | 0.347 | 0.064 | 0.004 | 0.436 | 0.172 | 0.368 | 0.264 | -0.046 | -0.014 | England | | | | | |
| 0.002 | 0.455 | 0.351 | 0.488 | 0.606 | 0.561 | 0.083 | -0.072 | 0.605 | -0.017 | 0.787 | 0.422 | -0.021 | -0.078 | 0.415 | Lithuania | | | | |
| 0.003 | 0.388 | 0.528 | 0.321 | 0.216 | 0.093 | 0.17 | -0.099 | 0.261 | 0.123 | 0.176 | 0.328 | 0.015 | -0.129 | 0.165 | 0.183 | Turkey | | | |
| 0.015 | 0.051 | 0.118 | 0.033 | -0.037 | 0.036 | 0.421 | 0.015 | 0.003 | 0.201 | -0.02 | -0.002 | 0.091 | 0.059 | 0.031 | -0.019 | 0.219 | Kyrgyzstan | | |
| 0.002 | 0.252 | 0.192 | 0.327 | 0.325 | 0.483 | 0.073 | -0.047 | 0.334 | -0.013 | 0.484 | 0.286 | -0.007 | -0.058 | 0.253 | 0.489 | 0.164 | 0.036 | Italy | |
| 0 | 0.139 | 0.112 | 0.155 | 0.177 | 0.332 | 0.052 | -0.014 | 0.195 | 0.007 | 0.27 | 0.144 | -0.006 | -0.026 | 0.175 | 0.232 | 0.101 | 0.034 | 0.352 | Spain |

**Table 5.** Agreements between classifiers using Kappa statistics. Light red: very good agreement (0.801 to 1.0), Blue: good agreement (0.601 to 0.8), Green: moderate agreement (0.401 to 0.6), Yellow: fair agreement (0.201-0.40)