Development and Evaluation of a Facebook-based Product Advisor for Online Dating Sites

Martin Winter[1], Sebastian Goebl[1], Nina Hubig[2], Christopher Pleines[3]
and Christian Böhm[1]
[1]University of Munich, Germany,
winterm@cip.ifi.lmu.de, {goebl,boehm}@dbs.ifi.lmu.de
[2]Helmholtz Zentrum Munich, Germany,
nina.hubig@helmholtz-muenchen.de
[3]Zu-Zweit.de, pleines@zu-zweit.de

Abstract: Social networks play an important role in Web 2.0. For many users establishing contacts and staying in touch in the virtual world is more than just a spare time filler. In social networks like Facebook they provide much information about themselves in user profiles. Also for online dating the focus is on establishing new contacts. In general, three types of dating sites can be distinguished: more serious dating agencies, less focused singles' platforms, and casual dating sites. In the proposed paper we develop a product advisor that uses the Facebook profile information provided by a user to classify her or him to one of the three dating site categories which is most suitable for the user's purpose. For classification we use Naive Bayes. To train the classifier we investigate the correlation between profile information and choice of dating sites. We also evaluate this correlation on collected data of representative German dating sites. Although a sharp distinction is hard to find, tendencies and enlightening insights are revealed by the collected data.

1 Introduction

Since the evolution from Web 1.0 to 2.0 the focus is no longer exclusively on providing information but on involving people. Social networks represent a central building block for Web 2.0, providing virtual communities for social relations online. In Germany, social networks are an essential part of life. With 46.1 million of 81.8 million inhabitants more than half of Germany uses social networks, where Facebook is leading with 13 million users per day (24 million users in total) far ahead of other social networks like Ask.fm (0.6 million users per day) or Xing.com (0.4 million users per day) [MB13]. Sharing news with friends is a quick and easy thing to do in social networks. This simplicity helps to keep information up to date. An immense data pool has been formed by all this profile information. It is used e.g. for target group oriented advertising. Also, developers of applications for Facebook can access the complete profile information if a user gives her or his consent. This is what we make use of in the development of our product advisor for online dating sites.

We distinguish three categories of online dating sites. The first category are *dating agen-*

Dating Agencies	Singles' Platforms	Casual Dating Sites
Parship.de	Neu.de	Secret.de
Elitepartner.de	DatingCafe.de	C-Date.de
eDarling.de	FriendScout24.de	Lovepoint.de

Table 1: Dating sites selected for our product advisor.

cies. These are aimed at users who are interested in serious relationships. The sites make use of psychological questionnaires to find a matching partner. Then, there are less focused *singles' platforms* for users who want to flirt and date as well — but with not as serious intentions as dating agencies' users. Finally, *casual dating* sites aim at users who are looking for erotic adventures without being interested in a relationship. For the proposed product advisor for online dating sites we select three German dating sites for each category, shown in Table 1.

User questionnaires on dating agencies' websites aim to collect information about the user to offer him the optimal matching partner. We go a step further and analyze user interests already before she or he even chooses a specific dating website and can be confronted with a questionnaire. Also, by using Facebook user profile information we access an even larger data pool as is possible by questionnaires. The dating agency Parship.de ask every user 70 questions for matching purposes. On the other hand, in average our Facebook-based product advisor collects 124 likes per user, which is used for classification.

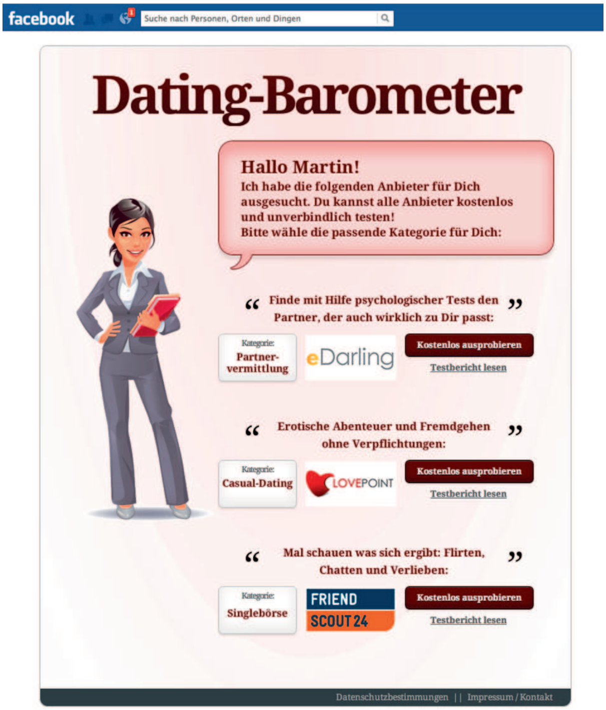
Our product advisor is trained by user decisions: in the data collecting phase it has suggested three random dating sites — one of each of the dating site categories. The user's decision to access a dating site (by clicking on one of the three suggested links) as well as information about him signing up has been stored together with her or his Facebook profile information. Using Naive Bayes we construct a classifier for each dating site category. The classifier decides which of the three dating site categories is most suitable for the user. We build the classifier based on the preferences of dating site categories that have been gained in the data collecting phase. Using cross validation the collected data is also used to test the quality of our classifier. Additionally, we present insights into interesting correlations that have been discovered when analyzing the data.

2 Related Work

Research has addressed social networks in general and Facebook especially, mostly concerning privacy issues as well as users' generosity of giving away private information. Stutzman et al. [SGA12] investigated from 2005 to 2011 how Facebook users handle their own privacy. It was found out that exposure of private information to the public has decreased during this period, while on the other hand sharing information to Facebook friends has increased. Privacy issues and the amount of shared information were investigated by Gross et al. [GA05] on 4000 students of the Carnegie Mellon University. Especially, attacks on user like stalking were considered. It was found out that users provide information to non-direct ’friends’ which in fact are total strangers. Facebook’s Gross National Happiness Index investigated the connection between positive and negative status updates, and the mood of the population of different countries. It was analyzed by Wang et al. [WKSR14] who found out that it does not sufficiently express the nations’ mood. Kosinski et al. [KSG13] drew attention to their study on predicting private attributes from Facebook information. By only few information about Facebook likes it was possible to construct an exact personal profile about Facebook users. In the study 88% of the male users were classified into their correct sexual orientation and 82% of them properly distinguished between Christians and Moslems. The authors developed the Facebook application ’You are what you like’ allowing users to receive a personality analysis based on their Facebook likes. In average Facebook users provided more than 200 likes on their Facebook profile. Bachrach et al. [BKG$^+$12] also addressed the possibility to draw conclusions from Facebook likes to the user’s personality. For an exact personality analysis the ’Big Five’ personality traits (openness, conscientiousness, extroversion, agreeableness, and neuroticism) were reconstructed for each user. As result it was proved feasible to extract the Big Five from the provided personal information. By asking the user to fill out a questionnaire online dating agencies follow the same intention to analyze the user’s personality in order to provide optimal matches. However, these algorithms are not made public. The related work concentrates on extracting personality traits from user information provided more or less intentionally through Facebook profile information. The mentioned related work proves that drawing conclusions from Facebook profile information is possible. We tie on this research and use the user-provided Facebook information to implicitly also analyze his personality. We aim to develop a product advisor that suggests the online dating site category that is most suited to his character as far as it is reconstructed from the provided information. As an additional benefit the training of the product advisor allows an evaluation of which types of personality structure are interested in which type of only dating site category. This represents an insight which can be used to tailor online dating sites to their target group.

3 The Dating Barometer

Facebook offers the possibility to create web applications providing tools for Facebook developers. These applications (called *apps*) are available for Facebook users only. When accessing an app for the first time, the user is asked if he agrees in sharing his profile information with the Facebook application. In case of consent the app has access to private profile information. Our application is called *Dating Barometer* (as it measures dating intentions) and is designed for German users, since it is comparing German online dating sites. First, training data for the product advisor needs to be collected to enable classification. Therefore, the Dating Barometer randomly creates a suggestion of dating sites for the user. The suggestion consists of a ranking of the three dating site categories represented by one of the sites as assigned to in Table 1 (cf. Figure 1). Although ordered randomly for collecting training data, the final product recommendation based on Facebook profile information is presented in the same way. The data collected by the Dating



facebook Suche nach Personen, Orten und Dingen

Dating-Barometer

Hallo Martin!
Ich habe die folgenden Anbieter für Dich ausgesucht. Du kannst alle Anbieter kostenlos und unverbindlich testen!
Bitte wähle die passende Kategorie für Dich:

“Finde mit Hilfe psychologischer Tests den Partner, der auch wirklich zu Dir passt:”

Kategorie: Partnervermittlung | eDarling | Kostenlos ausprobieren | Testbericht lesen

“Erotische Abenteuer und Fremdgehen ohne Verpflichtungen:”

Kategorie: Casual-Dating | LOVEPOINT | Kostenlos ausprobieren | Testbericht lesen

“Mal schauen was sich ergibt: Flirten, Chatten und Verlieben:”

Kategorie: Singlebörse | FRIEND SCOUT 24 | Kostenlos ausprobieren | Testbericht lesen

Datenschutzbestimmungen || Impressum / Kontakt

Figure 1: Dating Barometer creating a ranked suggestion of dating sites for the user.

Barometer consists basically of two parts: the *general Facebook profile information* and the *extended Facebook profile information*. The first comprises *name*, *e-mail address*, *sex*, *birthday*, *place of residence*, and *profile picture*. The second comprises additional profile information about *activities*, *interests*, *favorite music*, *favorite TV shows* and *favorite movies*, as well as contents that are connected by pressing the like button, called Facebook *likes*. Additional profile information about *age*, *interest in men and/or women*, and *relationship status* is expected to be very relevant and, thus, is collected, too. We name the extended Facebook profile information *meta (profile) information* in contrast to *basic* rest of the profile information.

The Dating Barometer is implemented in PHP using the PHP-SDK provided by Facebook. All data is stored in a MySQL database. On his first visit a user receives a HTTP cookie containing a distinct identifier. This is necessary, since the Dating Barometer has no profile information at all, before the user consents to sharing. Thus, if a user visits the Barometer but leaves without sharing and revisits again, only a cookie can identify him. All user interaction with the Dating Barometer is stored in our database. New users or revisiting users who have not shared profile information are welcomed by our virtual character Lisa and informed that the application offers a recommendation for online dating sites based on his profile information. If consenting the user is given a random recommendation as defined above. Then, the user can ask for further information about a specific dating site and receives a test review (button “Testbericht lesen”(read review) on Figure 1), or he can directly access the registration form for the dating site (button “Kostenlos ausprobieren”(try for free) on Figure 1). Clicking one of these buttons is recorded for this user in our database as a *click-out*. In case the user signs up for the online dating site, this is stored as a *sign-up* for this user.

4 Evaluation

For about three months the Dating Barometer has been advertised to attract users and been active for collecting data on Facebook. Additional to the user information on Facebook, the accessed online dating sites (cf. Table 1) reported if a click-out resulted in a sign-up. In this section we present general observations derived from the collected data and construct a Naive Bayes classifier that predicts a dating site category for a user providing her or his Facebook information. We used LIBSVM for the Naive Bayes classifier.

4.1 General Observations

4.1.1 Conversion Rate

Until the end of the e n guided to the Dating Barometer. As Figur onsent when asked for profil information. With 896 click-outs by 606 different users the rate is 43.60% with a mean of 1.48 click-outs per user. This means that nearly the half of the users sharing profile information also made used of the Dating Barometer. 221 (36.47%) of these users also signed up at an online dating site.

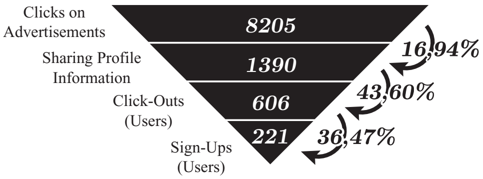


Figure 2: Conversion rate of users visiting the Dating Barometer.

4.1.2 Gender

The Dating Barometer attracted men more then women. Only 15.97% of all users sharing profile information were female users (cf. Figure 2). Of these female users 31.08% resulted in a click-out and 10.36% in a sign-up. For men the relative click-out and sign-up rates are higher: 45.97% clicked on a dating site suggesting and 16.98% signed up for a dating site.

	Total	Female	Male	Unknown
Clicks on Advertisements	8205	—	—	—
Sharing Profile Information	1390	222	1166	2
Click-Outs	606	69	536	1
Sign-Ups	221	23	198	0

Table 2: Gender of Dating Barometer users.

4.1.3 Age

The mean age of all Dating Barometer users was 31.12 ± 11.35 years. Most of the users (52.37%) were between 18 and 29 years of age (cf. Figure 3). 23.60% of the users were between 30 and 39 years, whereas 16.12% of the users were between 40 and 49 years. The remaining users were over 50 years, constantly decreasing. Some high ages might not be too realistic. However, they do not appear to impact the overall picture. The male users' age distribution corresponds to the overall distribution, whilst the distribution of female users between 18 and 55 is relatively uniform and only slightly decreasing. Users under 18 years of age are not regarded in the following observations as they are not allowed to use most of the online dating sites because of the sites' general terms and conditions.

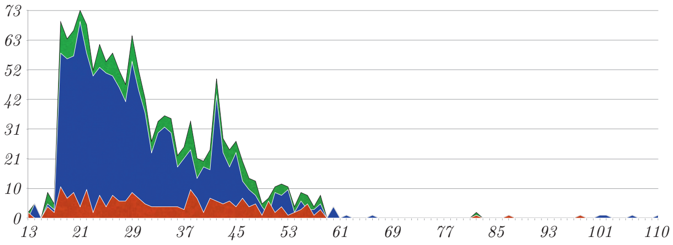


Figure 3: Total number of users (vertical axis) and their age (horizontal axis) for female (red), male (blue) and all (green) users.

Meta Profile Information	Mean	Minimum	Maximum
Likes	124.03 ± 306.69	0	4552
TV Shows	0.03 ± 0.17	0	1
Activities	2.97 ± 8.78	0	213
Interests	1.48 ± 2.87	0	36
Music	0.50 ± 0.95	0	8
Movies	0.03 ± 0.17	0	1
Total	129.04 ± 289.49	0	4621

Table 3: Mean, minimal and maximal number of provided meta profile information entries for all Dating Barometer users (who consented to sharing profile information).

4.1.4 Meta Profile Information

For every meta profile information there are users with no details. 32.59% users have provided no information to Facebook for at least one meta information. A total of 61 users have given no meta information at all. Table 3 shows mean, minimal and maximal number of entries of meta information for all users who share information.

4.1.5 Relationship Status

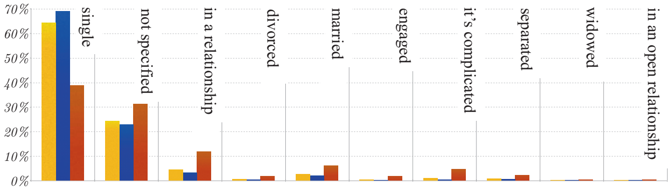
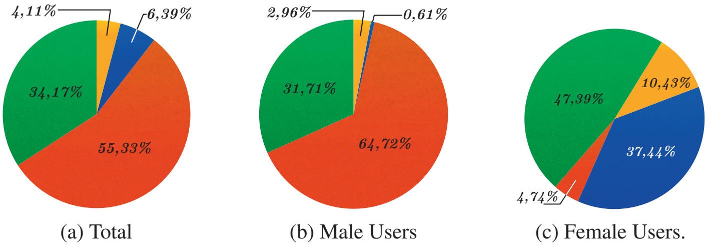


Figure 4: Occurrences of types of relationship status under Dating Barometer users showing male (blue), female (red) and all (yellow) users.

63.09% of all Dating Barometer users declared themselves as being single, followed by 23.81% not specifing their relationship status. All other types of relationship status were represented by maximal 4.53% of the Dating Barometer users (cf. Figure 4).

4.1.6 Sexual Orientation



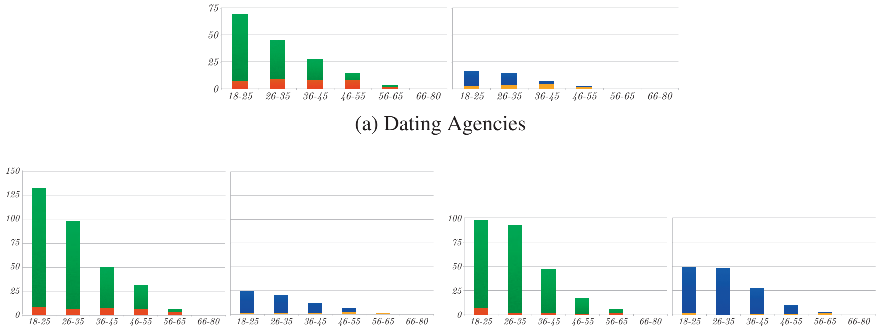
(a) Total (b) Male Users (c) Female Users.

Figure 5: Sexual Orientation: interest in men (blue), women (red), both (yellow) or not specified (green).

Illustrated by Figure 5, most of the Dating Barometer users state being heterosexual. Women tend to declare this less (37.44%) than men (64.72%). A total of 4.11% states to be bisexual, women a bit more often than men. Interest in the same sex is told by 4.74% of the female and 0.61% of the male users. 34.17% of all users have not specified their sexual orientation.

4.1.7 Distribution of Age Groups over Online Dating Site Categories

Figure 6 distinguishes the age groups related to the three online dating site categories. The charts also differentiate between click-outs and sign-ups for male and female users. Some results can already be read: singles' platforms were most successful in click-outs (23.53%), while casual dating sites showed the highest sign-up rate (9.86%). Dating agencies were in click-outs (11.73%) and sign-ups (2.88%) at the end. 4.60% of the users signed up at a singles' platform, while 18.92% clicked in casual dating sites. Singles' platforms and dating agencies both have a strong peak in the 18-25 age group. Instead, casual dating sites are nearly as popular in the 26-35 age group.



(a) Dating Agencies

(b) Singles' Platforms

(c) Casual Dating Sites

Figure 6: Distribution of click-outs/sign-ups (total number: vertical axis) over age groups (horizontal axis) for male click-outs (green), male sign-ups (blue), female click-outs (red) and female sign-ups (yellow) for the three different online dating site categories.

4.2 Naive Bayes Classification

4.2.1 Data Selection

To identify relevant meta profile information and to exclude unspecific data, e.g. likes as 'McDonalds' or 'CocaCola', we calculate the relevance of meta information. We derive it from the relation of Dating Barometer users *liking* a meta profile information (MPI) to all Facebook users *liking* it as

$$relevance_{MPI} = \frac{\textit{\# of Dating Barometer users with MPI}}{\textit{\# of all Facebook users with MPI}}.$$

For data classification we only use the top 20 most relevant meta profile information entries.

We address another not trivial problem: meta data is generated by and, thus, named by the Facebook users. Therefore, for the same subject there exist several different meta information entries. This can be due to spelling mistakes or to different levels of concept hierarchies, e.g. 'The Simpsons' compared to 'Homer Simpson' or 'Bart Simpson', all relating to the same TV series. However, other meta information entries are not connected in any way even if written in a similar way. E.g., 'Jessica Simpson' is equal to the TV series if comparing text strings. To deal with this problem we use the rich possibilities of Google's search. Google's syntax recognition allows to differentiate between search keywords by using the top 100 URLs. On this basis, we construct a Google crawler that fills a database with the top 100 URLs for each meta profile information entry.

After crawling the search keywords, a main URL is set for each of the top 20 meta information entries. By comparing strings of meta information entries in combination with comparing the top 100 URLs of each entry we associate the top 20 meta information entries with its matching entries and use the combined new meta information entries as input data for our evaluation. To tie on the example from above, a Google search for the TV series meta information entry 'The Simpsons' retrieves the URL `www.thesimpsons.com` as first search result. This URL is also suitable to be set as main URL for this entry. If searching for 'Bart Simpson', the same URL appears as tenth search result. On the other hand, a search for 'Jessica Simpson' does not show this URL at all in the top 100 results. Therefore, defining similarity over Google search results is a promising way to group meta information.

Details to religion are free text fields where users can describe their attitude using own words. In order to make use of this information, the main religious orientations in Germany, *Catholicism*, *Protestantism*, *Judaism* and *Islamism*, were extracted using string comparison. Altogether, for each user we extract the attributes *gender*, *age*, *sexual orientation*, *relationship status*, *religion*, *spoken language(s)*, *operating system*, *education*, *top 20 meta likes* and the connected *click-outs* and *sign-ups*.

4.2.2 Naive Bayes Classifier

We create a Naive Bayes classifier for a simplified data set using all but meta profile information. The numbers of click-outs (606) and sign-ups (221) suggest using k-fold cross validation with $k = 10$ to validate our classifier. Table 4 presents the confusion matrix for the Naive Bayes classifier for click-outs trained on the simplified data set. For dating agencies and casual dating sites the negative predictive value is 74.62% and 73.76%. The predictive value for singles' platforms is 55.77%. The weighted mean of the negative predictive value of this classifier is 65.73%.

DA	**SP**	**CD**	
28	129	6	**DA**
35	234	11	**SP**
11	139	13	**CD**

Table 4: Confusion matrix for click-out classifier for all dating site categories: dating agencies (DA), singles' platforms (SP) and casual dating sites (CD)

We also construct a Naive Bayes classifer constructed for sign-ups using all profile information including meta information. It is validated to a negative predictive value showing a weighted mean of 68,54%.

5 Conclusion

The click-outs showed that the Dating Barometer users in the age groups 18-25 and 26-35 are more interested in singles' platforms and casual dating sites than dating agencies. However, singles' platforms and dating agencies show about the same number of sign-ups in these age groups, but the number of sign-ups in casual dating sites is more than double the number. Thus, this means that these users are more interested in erotic adventures.

With altogether 1390 users of which could be used only 44% for click-out classification and 16% for sign-up classification, the amount of used data is rather small. But still, results clearly show facts as such that dating agencies are also most relevant for the age group 18-25. We can also define with an accuracy of 81% if users sign up in a dating agency or not. The accuracy for singles' platforms here is 71% and for casual dating sites 67%.

However, classification by Naive Bayes classifiers to predict the type of dating site proves to be a difficult task. However, it is also due to the small numbers of click-outs and sign-ups that it is hard to establish useful classifiers.

References

[BKG$^+$12] Yoram Bachrach, Michal Kosinski, Thore Graepel, Pushmeet Kohli, and David Stillwell. Personality and Patterns of Facebook Usage. In *Proceedings of the 3rd Annual ACM Web Science Conference*, WebSci '12, pages 24–32, New York, NY, USA, 2012. ACM.

[GA05] Ralph Gross and Alessandro Acquisti. Information Revelation and Privacy in Online Social Networks. In *Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society*, WPES '05, pages 71–80, New York, NY, USA, 2005. ACM.

[KSG13] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 2013.

[MB13] Ayaan Mohamud and Berit Block. 2013 Future in Focus - Digitales Deutschland. `http://www.comscore.com/ger/Insights/Presentations_and_Whitepapers/2013/2013_Future_in_Focus_Digitales_Deutschland`, 2013. Accessed: 2014-03-26.

[SGA12] Fred Stutzman, Ralph Gross, and Alessandro Acquisti. Silent listeners: The evolution of privacy and disclosure on Facebook. *Journal of privacy and confidentiality*, 4(2):7–41, 2012.

[WKSR14] N. Wang, M. Kosinski, D.J. Stillwell, and J. Rust. Can Well-Being be Measured Using Facebook Status Updates? Validation of Facebookâs Gross National Happiness Index. *Social Indicators Research*, 115(1):483–491, 2014.