Identifikácia osobnostných vlastností na základe údajov publikovaných v rámci sociálnych médií

Peter Koncz[1], Ján Paralič[1], Dominik Jakub[1]

[1]Katedra kybernetiky a umelej inteligencie, FEI TU v Košiciach, Slovenská republika
{peter.koncz, jan.paralic}@tuke.sk, dominik.jakub@student.tuke.sk

Abstrakt. Osobné profily sú cenným zdrojom informácií o používateľoch sociálnych médií. Hoci sa o týchto informáciách často predpokladá, že sú použiteľné pre identifikáciu osobnosti používateľov, systematický výskum v tejto oblasti je stále v počiatkoch. Práca je preto venovaná overeniu možností predikcie hodnôt dimenzií osobnosti podľa teórie Big Five na základe údajov zo sociálnej siete Facebook. Na základe údajov od 330 používateľov sociálnej siete, ako aj údajov z osobnostného dotazníka, bola overovaná možnosť vytvorenia predikčných modelov pre identifikáciu osobnosti. Získané predbežné výsledky potvrdzujú výsledky prác ktoré poukazujú na nedostatočnú presnosť takto vytvorených modelov.

Keywords: Big Five, dolovanie v dátach, sociálne siete

1 Úvod

Sociálne médiá predstavujú významný zdroj údajov o osobnostných vlastnostiach ich používateľov. Napriek tomuto široko akceptovanému tvrdeniu existuje relatívne malé množstvo prác, ktoré sa snažili toto tvrdenie výskumne overiť. Cieľom nášho výskumu je overenie možnosti použitia údajov publikovaných v rámci sociálnej siete Facebook pre tvorbu modelov na automatickú identifikáciu základných osobnostných vlastností používateľov. V nasledujúcich častiach najskôr stručne popíšeme existujúci výskum v tejto oblasti, následne popíšeme realizované experimenty a získané výsledky a napokon zhrnieme v podobe odporúčaní pre ďalší výskum.

2 Súvisiace práce

Ako sme spomínali, v súčasnosti je množstvo prác venovaných nami skúmanej problematike relatívne malé, čo zrejme súvisí s náročnosťou vytvárania vzoriek ktoré vyžadujú vyplnenie osobnostného dotazníka jednotlivými respondentmi. Z pohľadu analýzy existujúcich zdrojov sme sa zameriavali predovšetkým na použité vstupné atribúty, použité predikčné metódy ako aj výsledky realizovaných experimentov. Sumner a kol. [1] sa zamerali na korelácie medzi osobnostnými vlastnosťami a aktivitou na sociálnej sieti Facebook. Vytvorili zoznam 79 atribútov. Do ich expe-

rimentu sa zapojilo 537 respondentov. Výsledky experimentov poukazujú na viaceré závislosti hoci ich sila je vo väčšine prípadov veľmi nízka. Bai a kol. [2] realizovali podobný výskum v Číne kde v čase výskumu bola najobľúbenejšia sociálna sieť Renren. Všetci účastníci boli známi výskumníkov alebo známi ich známych, všetci Číňania a vekový priemer bol 24 rokov. Vo výslednej vzorke 209 účastníkov boli len tí, čo vyplnili osobnostný dotazník, mali minimálne 100 priateľov, počet statusov viac ako 50 a blogov viac ako 10. Použité boli viaceré skupiny premenných vrátane premenných súvisiacich s prítomnosťou emócií, klasifikovaných na základe naivného Bayesovského klasifikátora na príznakovom priestore tvorenom emočne zafarbenými slovami. Výsledné skóre každého faktoru rozdelili do troch skupín, s nízkym, stredným a vysokým skóre. Takto upravené zadanie potom riešili ako klasifikačnú úlohu do troch tried. Pri trénovaní modelu použili naivný Bayesov klasifikátor, metódu podporných vektorov a rozhodovacie stromy. Najväčšiu presnosť dosahovali s algoritmom C4.5. Hodnoty presnosti, návratnosti a hodnoty F1 miery dosahovali v priemere 0,7. Wehrli [3] sa vo svojej práci zameral na používateľov sociálnej siete StudiVZ pričom jeho vzorka bola tvorená 1560 respondentmi. Na týchto údajoch aplikoval analýzu hlavných komponentov a bol schopný replikovať 5 faktorov, teda rysov osobnosti z teórie Big Five. V rámci uvedených prác bol použitý dotazník Big Five Inventory [4]. Tento dotazník je pre výskumné účely voľne použiteľný bez nutnosti pýtať si povolenie od autorov. Dotazník bol preložený do slovenčiny dvakrát, následne boli obe verzie spojené do jednej. Dotazník obsahuje 44 otázok predstavovaných krátkymi vetami typu „Som niekto kto ...“, ktoré respondent ohodnotí mierou súhlasu na škále 1-5.

3 Experiment

Cieľom experimentu bolo vytvorenie predikčného modelu pre hodnoty osobnostných dimenzií podľa modelu Big Five na základe údajov publikovaných na sociálnej sieti Facebook. Vstupné premenné boli predstavované údajmi z profilov jednotlivých používateľov, kým výstupné premenné boli hodnoty piatich osobnostných dimenzií, ktorými boli extroverzia, prívetivosť, svedomitosť, emocionálna stabilita a otvorenosť voči skúsenosti.

3.1 Vzorka

Vzorka údajov pochádzala od 330 respondentov, pomer mužov a žien bol 53:48 s priemerným vekom 22 rokov. Hlavnou nevýhodou vzorky bolo, že všetci respondenti boli priateľmi jedného človeka. Respondenti boli požiadaní o vyplnenie dotazníka Big Five Inventory.

Respondentov sme požiadali aj o poskytnutie svojich údajov z Facebooku. Z týchto údajov bolo vytvorených 46 číselných atribútov ako napr. počet albumov, či statusov, nominálne atribúty ako napríklad obľúbené knihy, filmy, relácie či tzv. liky, a 4 textové atribúty, ktoré sme vytvorili na základe popisov kníh, filmov, relácií či tzv. likovaných objektov. Okolo zmienených troch skupín vstupných premenných sme organizovali aj jednotlivé experimenty.

3.2 Postup

Nominálne atribúty boli transformované na binominálne atribúty a v prípade textových atribútov sme vykonali odstránenie neplnovýznamových slov a vytvorili sme jedno a bi-gramy slov. Vyberali sme iba slová ktorých minimálny výskyt vo všetkých dokumentoch bol 10 a maximálny 200, pričom prítomnosť n-gramov bola indikovaná binárne. Následne sme vytvorili tri sady premenných. Prvú tvorili čisto numerické premenné, druhá bola tvorená kombináciou numerických a nominálnych a napokon sme použili čisto textové premenné. Cieľom bola predikcia hodnoty faktorov. Použitými metódami bola metóda podporných vektorov (SVM) v podobe dostupnej v knižnici LibSVM ako nu-SVR s radiálnou bázovou funkciou a základnými nastaveniami a metóda k najbližších susedov (K-NN) s počtom susedov 3. Pre trénovanie bolo použitých 90% záznamov a pre testovanie 10%.

Keďže dosahované hodnoty presností predikcie boli relatívne nízke, úlohu sme hodnotili aj ako klasifikačnú úlohu. Transformovali sme cieľové atribúty na nominálne s troma stupňami. Záznamy zoradené podľa hodnoty boli rozdelené do štyroch rovnako veľkých kvartálov, pričom stredné dva predstavovali strednú hodnotu premennej. V tomto prípade bola použitá desaťnásobná krížová validácia.

Pre porovnanie presnosti sme určili základné hladiny (baseline), ktoré boli predstavované hodnotou presnosti predikcie pri použití priemernej hodnoty cieľovej premennej v prípade predikčných úloh a pravdepodobnosťou najpravdepodobnejšej triedy v prípade klasifikačnej úlohy.

3.3 Výsledky

Na vyhodnotenie presnosti predikcie bola použitá stredná kvadratická chyba, pre vyhodnotenie presnosti klasifikácie bola použitá presnosť. Výsledky pre vzorky s použitím numerických aj nominálnych premenných sú zobrazené v tabuľke 1. Výsledky získané iba pre numerické premenné mali veľmi podobný charakter. Z výsledkov je možné vidieť vyššiu presnosť dosahovanú pri použití metódy podporných vektorov. Dosahované hodnoty chýb však nie sú významne nižšie ako hodnoty základnej hladiny, dokonca sú v niektorých prípadoch vyššie.

Tabuľka 1. Hodnoty strednej kvadratickej chyby pre jednotlivé dimenzie a metódy

	Extroverzia	Prívetivosť	Svedomitosť	Neuroticizmus	Otvorenosť
SVM	0,666	0,491	0,600	0,740	0,566
K-NN	0,899	0,639	0,834	0,963	0,756
Baseline	0,692	0.452	0.630	0,772	0.526

V prípade nominálnych premenných sa ako najväčší problém ukázal nedostatok premenných, ktoré by mali dostatočné množstvo pozitívnych hodnôt v rámci celej vzorky. Ani v prípade klasifikačnej úlohy neboli dosahované hodnoty ktoré by významne prekračovali hladinu náhodnosti (33% pre tri rovnako zastúpené úrovne hodnôt).

4 Záver

Na základe získaných výsledkov sme dospeli k záveru, že nami vytvorené modely nedosahujú hodnoty presnosti potrebné pre ich reálne nasadenie. Táto skutočnosť zrejme súvisí s nedostatočnou veľkosťou použitej vzorky ako aj potrebou zahrnutia ďalších skupín premenných. Ako potenciálne užitočné premenné môžeme uviesť premenné vyplývajúce z priateľstiev medzi používateľmi, či údaje zo správ používateľov, ktoré neboli v realizovaných experimentoch zahrnuté.

Poďakovanie

Táto práca bola podporovaná Agentúrou na podporu výskumu a vývoja na základe zmlúv č. APVV-0208-10 (50%) a č. SK-CZ-2013-0062 (50%).

English summary

Personal profiles are a valuable source of information about users of social media. Although this information is often assumed to be useful for identifying user personalities, systematic research in this area is still at its beginnings. The work is therefore devoted to verification of the possibilities of predicting personality dimension values according to the Big Five theory, based on data from Facebook. Based on data from 330 users as well data from personality questionnaire the possibility of creation of such prediction models was evaluated. The obtained results confirm some of the previous works with lack of precision of the created models. The paper also describes the problems associated with this prediction.

Referencie

1. Sumner, C., Byers, A., Shearing, M.: Determining personality traits & privacy concerns from facebook activity. Black Hat Brief. 11, (2011).
2. Bai, S., Zhu, T., Cheng, L.: Big-Five Personality Prediction Based on User Behaviors at Social Network Sites. ArXiv12044809 Cs. (2012).
3. Wehrli, S.: Personality on Social Network Sites: An Application of the Five Factor Model. ETH Zurich, Chair of Sociology (2008).
4. John, O., Naumann, L., Soto, C.: Paradigm Shift to the Integrative Big Five Trait Taxonomy: History, Measurement, and Conceptual Issues. In: John, O., Robbins, R., and Pervin, L. (eds.) Handbook of Personality: Theory and Research. pp. 114–156. Guilford (2008).