Analiza uporabnosti podatkov iz družbenih medijev

Neli Blagus, Slavko Žitnik in Marko Bajec
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Večna pot 113, Ljubljana
{neli.blagus, slavko.zitnik, marko.bajec}@fri.uni-lj.si

Izvleček

Z razvojem svetovnega spleta in družbenih medijev smo vse bolj vpeti v uporabo interneta v vsakdanjem življenju. Družbena omrežja, različni forumi in splet na splošno spreminjajo način komunikacije in sodelovanja med nami. Podatki iz takšnih medijev so bogata zakladnica mnenj, ki nam ob pravilni obravnavi omogočajo celovitejše razumevanje družbe. V članku predstavljamo pet popularnih družbenih medijev: Twitter, Facebook, Tumblr, Google+[1] in YouTube. Raziskali smo njihove možnosti pridobivanja podatkov ter za izbrane ključne besede analizirali razlike med mediji in tematikami ključnih besed. Izkazalo se je, da sta najbolj uporabljena družbena medija Twitter in Facebook, a med njima obstajajo razlike, saj na Twitterju uporabniki raje objavljajo, na Facebooku pa komentirajo. Uporabniki, ki so pretežno moškega spola, v splošnem največ pišejo o znanih osebnostih ter politikih, objave pa so v večini v angleškem jeziku.

Ključne besede: Družbena omrežja, analiza podatkov, analiza omrežij, Twitter, Facebook, Tumblr, Google+, YouTube.

Abstract

Social Media Data Usability Analysis

In the past decade, social media has become an important part of our everyday life. Social media has changed the way we communicate, collaborate and gather information. Researchers from different fields exploit social media to provide deeper insight into human behavior. In this paper, we present five social media platforms with their potential for data analysis. Investigating several topics, we provide insight into the use of the social media data. The analyses show reach and engagement differences, usability of a social network based on user type and their goals or general behaviour. We observe statistics and compare data from Twitter, Facebook, Tumblr, Google+ and YouTube. The results reveal differences in posting or commenting. Twitter is the most influential media, yet the influence of Facebook is also considerable. Furthermore, posts from compared social media differ in language, sentiment and gender identification aspects.

Keywords: Social networks, data analysis, network analysis, Twitter, Facebook, Tumblr, Google+, YouTube.

1 UVOD

Junija 2018 je bilo na družbenem omrežju Facebook registriranih več kot dve milijardi uporabnikov. Od tega jih 66 odstotkov uporablja Facebook vsakodnevno, kar kaže na to, da so družbeni mediji pomemben del našega življenja (Statista, 2018). Družbena omrežja, različni forumi in splet na splošno spreminjajo način komunikacije in sodelovanja med nami. Podatki iz družbenih medijev so bogata zakladnica mnenj ter nam ob pravilni obravnavi omogočajo celovitejši pogled na družbene probleme in lažje razumevanje delovanja družbe.

Družbeni medij je definiran kot spletna ali mobilna aplikacija, s katero lahko uporabniki kreiramo različne vsebine in si jih izmenjujemo (Kaplan in Haenlein, 2010). Poznamo več tipov družbenih medijev, na primer blogi (angl. blogs; npr. Blogger) in mikroblogi (angl. microblogs; npr. Twitter), spletna družbena omrežja (angl. online social networks; npr. Facebook) in spletna mesta za izmenjavo vsebin (angl. media sharing sites; npr. YouTube) (Gundecha in Liu, 2012). Vsakodnevno uporabniki na družbenih medijih objavijo ogromne količine vsebin. Pridobivanje, shranjevanje in analiza tako velikih količin podatkov je velik izziv za raziskovalce. Večina družbenih medijev ima za to na voljo t. i. aplikacijski programski vmesnik (angl. application programming interface, API). API pomeni množico rutin, protokolov in orodij za izgradnjo programske opreme in aplikacij; med drugim prek API-jev družbeni mediji

[1] Storitev Google+ bo leta 2019 ukinjena za končne uporabnike.

ponujajo dostop do različnih podatkov tudi za raziskovalne namene.

V raziskavah je med vsemi družbenimi mediji najbolj priljubljen Twitter; podatke iz Twitterja največkrat uporabljajo v različnih študijah (Felt, 2016). Tufekci (2014) popularnost Twitterja pripisuje dostopnosti in preprostosti pridobivanja podatkov; Osborne in Dredze (2014) pa sta pokazala, da se na Twitterju najhitreje pojavijo objave, vezane na aktualne novice. Sicer raziskovalci podatke iz Twitterja uporabljajo na primer za identifikacijo pomembnih novic (Petrovič idr., 2013), zaznavanje potresov (Sakaki, 2010), prometnih dogodkov (Anantharam idr., 2015) ali epidemij (Aramaki idr., 2011). Prav tako je veliko raziskav, ki analizirajo značke (angl. hashtags) (Lotan idr., 2011) ter iščejo vzorce v komunikaciji med uporabniki Twitterja (Naaman idr., 2010).

Podatki iz družbenih medijev, kot so na primer Tumblr, YouTube in Facebook, so v raziskavah uporabljeni redkeje. Razlog za to najbrž leži v sami popularnosti posameznega medija ter v omejitvah pri pridobivanju podatkov iz njihovih API-jev. Podatki iz Tumblrja so uporabljeni v analizi sentimentov (angl. sentiment analysis) (Bourlai in Herring, 2014) ali za odkrivanje različnih vrst vsebin (Xu idr., 2014). Raziskave, ki uporabljajo podatke iz YouTuba, se ukvarjajo predvsem z napovedovanjem popularnosti spletnih vsebin (Figueiredo idr., 2011) ali z obnašanjem uporabnikov (Siersdorfer idr., 2010). Raziskave, ki uporabljajo podatke iz Facebooka, pa se ukvarjajo pretežno z analizo osebnosti uporabnikov glede na njihovo aktivnost (Moore in McElroy, 2012) in z zasebnostjo podatkov (Zimmer, 2010).

V prispevku analiziramo podatke iz petih družbenih medijev: Twitter, Facebook, Google+, Tumblr in YouTube. Predstavljamo njihove API-je ter možnosti in omejitve pri pridobivanju podatkov. Nato za izbrane ključne besede po izbranih medijih opazujemo statistiko objav, ki vsebujejo te ključne besede. Analizo smo izvedli v treh delih: A) v prvem merimo doseg (angl. reach) in aktivnosti (angl. engagement) objav s posamezno ključno besedo, B) v drugem delu analiziramo osnovno statistiko objav (število vseh objav, komentarjev in porabnikov), C) v tretjem delu pa analiziramo jezik in sentiment objav ter spol uporabnikov po posameznih družbenih medijih. Osredotočili smo se na opazovanje razlik med tematikami ključnih besed ter analizirali obnašanje uporabnikov. Izkazalo se je, da se to med družbenimi mediji razlikuje, na primer na Twitterju uporabniki več objavljajo, na Facebooku komentirajo. Skupen vsem analiziranim družbenim medijem je angleški jezik večine objav. Prav tako so na vseh medijih aktivni pretežno uporabniki moškega spola, ki objavljajo bolj negativne kot pozitivne objave.

Nadaljevanje prispevka je sestavljeno iz treh delov. Najprej v razdelku 2 opisujemo družbene medije in njihove API-je. V razdelku 3 predstavljamo ključne besede, uporabljene v analizi, ter rezultate in diskusijo, razdelek 4 je sklepni.

2 DRUŽBENI MEDIJI IN NJIHOVI API-JI

V raziskavi smo se osredotočili na družbene medije, ki so dobro poznani v Evropi: Facebook, YouTube, WhatsApp, Google+, Tumblr in Twitter (Reuter in Scholl, 2014). Aplikacija WhatsApp ne omogoča dostopa do podatkov, zato smo uporabili ostalih pet medijev.

Podatki so iz družbenih medijev dostopni prek API-jev. Med posameznimi mediji obstajajo razlike pri načinu dostopa do podatkov, omejitvah pri pridobivanju podatkov ter njihovi strukturi (Reuter in Schooll, 2014). Vprašanja, ki se porajajo pri pridobivanju in analizi teh podatkov, se tičejo pretežno zasebnosti ter načina nadzora upravljalcev družbenih medijev nad podatki (Schafer in van Es, 2017). Več avtorjev je poudarilo problem pristranskosti podatkov, saj lahko na primer nekateri uporabniki objavljajo občutno več kot drugi, hkrati pa ne poznamo vse populacije, iz katere je narejen vzorec podatkov (Lomborg in Bechman, 2014; Ruths in Pfeffer, 2014). Kljub naštetemu pomenijo API-ji učinkovit način za zbiranje podatkov v raziskovalne namene. V nadaljevanju predstavljamo posamezne družbene medije in možnosti ter omejitve njihovih API-jev (pri predstavitvi metod, ki so prek API-jev dostopne, smo se osredotočili na tiste, ki so zanimive za podobne raziskave in analize). Na koncu poglavja v tabeli 1 povzemamo možnosti posameznih API-jev.

2.1 Twitter

Twitter je spletno družbeno omrežje, ki ponuja storitev mikrobloganja (tj. objave imajo omejeno dolžino). Twitter API (Twitter API, 2018) je namenjen upravljanju oglasnih kampanj in dostopu do podatkov. Vmesnik API sestavlja več storitev: oglasni (angl. advertising), sporočilni (angl. direct message), iskalni (angl. search) in pretočni (angl. streaming) API. Pridobivanju podatkov sta namenjena zadnja dva.

Iskalni API ponuja podatke za objave zadnjih sedem dni. To velja za brezplačno javno verzijo, možne so tudi plačljive premium verzije, za 30 dni oziroma za vse objave od leta 2006. V naši analizi smo uporabili brezplačno verzijo. Prek API-ja ni mogoče priti do čisto vseh objav, ima pa možnost filtriranja podatkov na različne načine, na primer glede na lokacijo ali jezik. Po drugi strani lahko prek pretočnega API-ja pridobivamo objave, ki so objavljene v tistem trenutku in vsebujejo določeno ključno besedo ali jih objavlja določeni uporabnik. Ta API vrne več rezultatov kot iskalni, ima pa manj možnosti za filtriranje in prilagajanje rezultatov. Podatki so iz obeh API-jev dosegljivi z zahtevkom HTTP ali prek knjižnic programskih jezikov Python, PHP, JavaScript in drugih.

Do podatkov dostopamo z veljavnim žetonom za dostop (angl. access token). Ima pa API nekaj omejitev. Pretočni API lahko sledi hkrati 5000 uporabnikom in vrne približno odstotek vseh objav, ki so objavljene v danem trenutku. V iskalnem API-ju je mogočih 180 klicev za iskanje na uporabnika oziroma 450 klicev za iskanje na aplikacijo v časovnem oknu 15 minut.

Prek API-ja so dostopne štiri metode.

- Iskanje objav po ključni besedi. Za dano ključno besedo vrne objave, ki jo vsebujejo, skupaj z osnovnimi informacijami o objavi in uporabniku, ki je objavil objavo. Iščemo lahko glede na datum – pred določenim datumom in po njem (kot že omenjeno, pri brezplačni različici ni mogoče iskati po objavah, starejših od sedem dni, je pa mogoče starejše objave pridobiti prek izpisa objav določenega uporabnika). Med rezultati iskanja po objavah so tudi deljene objave (angl. retweet) in komentarji objav.
- Iskanje objav po lokaciji. Za dano ključno besedo in lokacijo (v okolici danih geografskih koordinat ali v podani državi) vrne vse objave, ki vsebujejo ključno besedo, njihov avtor pa ima podano informacijo o lokaciji oziroma državi.
- Iskanje uporabnikov po ključni besedi. Za dano ključno besedo vrne seznam uporabnikov, ki jo vsebujejo v imenu.
- Izpis podrobnejših informacij o objavi ali uporabniku. Za dano identifikacijsko številko objave vrne informacije, kot so na primer podatki o objavi, lokacija objave, statistike, kot so število ogledov, všečkov (angl. like), komentarjev ipd.

2.2 Google+

Google+ je spletno družbeno omrežje za povezovanje uporabnikov. API (Google+ API, 2018) je v osnovi namenjen izdelovanju aplikacij, ponuja pa tudi možnost dostopa do podatkov. API je dosegljiv prek zahtevkov HTTP in knjižnic programskih jezikov Python, PHP, Java, Ruby, Go in C#. Do podatkov dostopamo z veljavnim žetonom za dostop (je enak za vse Googlove storitve). V času pisanja tega prispevka (oktober 2018) je dostop do podatkov omejen s kvotami (angl. quota) na tri načine: za aplikacijo 10.000 zahtevkov na dan in 200 zahtevkov na 100 sekund in kvota za uporabnika 500 zahtevkov na 100 sekund.

Prek API-ja so dostopne te metode:

- iskanje objav po ključni besedi – za dano ključno besedo vrne objave, ki jo vsebujejo; iščemo lahko tudi objave glede na datum;
- iskanje uporabnikov po ključni besedi – za dano ključno besedo vrne seznam uporabnikov, ki jo vsebujejo v imenu;
- izpis komentarjev določene objave – za dano identifikacijsko številko objave vrne njegove komentarje (ime uporabnika, ki je zapisal komentar, besedilo komentarja, datum objave ipd.);
- izpis podrobnejših informacij o objavi – za dano identifikacijsko številko objave vrne njene osnovne informacije.

2.3 YouTube

YouTube je spletna storitev za objavljanje in deljenje video vsebin. YouTube API (YouTube API, 2018) je namenjen izdelavi aplikacij za interakcijo s storitvami YouTube in vsebuje analitični (angl. analytics), pretočni (angl. streaming) in podatkovni (angl. data) API. Prek njih je mogoče objavljanje in predvajanje posnetkov, iskanje in spreminjanje posnetkov ter dostop do različnih statistik. Za namene raziskav za dostop do podatkov uporabljamo podatkovni API. Dostopen je prek zahtevkov HTTP in knjižnic programskih jezikov Java, JavaScript, Python, PHP, Ruby, Go ter Node.js. Do podatkov dostopamo z veljavnim žetonom za dostop. V času pisanja tega prispevka (oktober 2018) je dostop do podatkov omejen s kvotami, pri čemer vsak klic stane nekaj kvot, na primer operacija branja stane 1 kvoto, nalaganje posnetka pa 1600 kvot. Vsak projekt ima na voljo milijon kvot na dan.

Prek API-ja so dostopne te metode:

- iskanje po ključni besedi – za dano ključno besedo vrne kanale, videe ali sezname predvajanja (angl. playlist), ki jo vsebujejo v naslovu ali opisu, skupaj z osnovnimi informacijami o videu, kanalu ali seznamu predvajanja; iščemo lahko glede na datum (pred določenim datumom in po njem) in glede na lokacijo (objave, ki so bile objavljene v okolici danih geografskih koordinat);
- izpis komentarjev določenega videa – za dano identifikacijsko številko videa vrne njegove komentarje skupaj z osnovnimi informacijami o avtorju komentarja, datumom objave ipd.;
- izpis podrobnejših informacij o videu – za dano identifikacijsko številko videa vrne informacije, kot so na primer podatki o kanalu, na katerem je objavljen video, lokacija objave, statistike, kot so število ogledov, všečkov, komentarjev ipd.

2.4 Tumblr

Tumblr je spletno družbeno omrežje, ki ponuja storitev mikrobloganja. Njegov API (Tumblr API, 2018) omogoča dostop do podatkov prek zahtevkov HTTP ali knjižnic v programskih jezikih Javascript, Ruby, Python, PHP, Java, C in Go. Do podatkov lahko dostopamo z veljavnim žetonom za dostop, omejitve pa niso točno specificirane in se spreminjajo glede na želene zahteve uporabnika API-ja.

Prek API-ja so dostopne te metode:

- iskanje objav po ključni besedi – za dano ključno besedo vrne objave, ki jo vsebujejo, skupaj z osnovnimi informacijami o blogu (blog je v tem primeru definiran kot uporabnik), v katerem je bila objava objavljena; iščemo lahko pred danim datumom ali po njem;
- izpis podrobnejših informacij o blogu – za dano identifikacijsko številko bloga vrne njegove osnovne informacije, kot so na primer ime, opis in število vseh objav bloga;
- izpis vseh objav bloga – za dano identifikacijsko številko bloga vrne njegove objave.

2.5 Facebook

Facebook je spletno družbeno omrežje, namenjeno komuniciranju, razvedrilu in vzpostavljanju družbenih odnosov. Glavni namen Facebookovega API-ja (Facebook API, 2018) je pomoč pri razvoju aplikacij za uporabnike omrežja. Vmesnik API sestavljajo oglasni (angl. ads), kreditni (angl. credits), klepetni (angl. chat) in omrežni (angl. graph) API-ji. Zadnji je namenjen pridobivanju podatkov tudi v raziskovalne namene. V času zbiranja podatkov za analizo (oktober 2018) je bil omrežni API dostopen prek zahtevka HTTP ali prek knjižnic različnih programskih jezikov. Javno dostopni so bili podatki o dogodkih, skupinah in straneh ter javnih objavah in komentarjih na straneh, ki so bile odprte za javnost (angl. open). Z aprilom 2018 so pri Facebooku zaostrili dostop do podatkov, tako dostop do objav v skupinah, dogodkih in straneh ni več mogoč, za dostop do nekaterih informacij pa je treba pridobiti posebna dovoljenja. Zato opozarjamo, da rezultate analize podatkov iz Facebooka v članku objavljamo, podobnih podatkov pa se ne da več pridobiti.

Do podatkov se je dostopalo z veljavnim žetonom za dostop; obstajajo štirje različni – za spreminjanje uporabnikovih osebnih podatkov, spreminjanje aplikacij, urejanje strani ter identifikacijo pri uporabi aplikacij.

Tabela 1: **Možnosti posameznih API-jev**

Družbeni medij	Iskanje po po objavah	Izpis podatkov in komentarjev objave	Iskanje po lokaciji (koordinate)	Iskanje po datumu	Iskanje uporabnikov
Twitter	✓	Izpis podatkov o objavi	✓	✓	✓
Google+	✓	✓	×	×	✓
YouTube	✓	✓	✓	✓	×
Tumblr	✓	×	×	✓	×

3 ANALIZA IN RAZPRAVA

Za namene raziskave smo iz družbenih medijev pridobili podatke o več ključnih besedah, ki se v medijih (npr. na spletu, v novicah) pojavljajo različno pogosto. Glede na tematiko smo razdelili izbrane ključne besede v šest skupin; prve tri skupine vsebujejo športnike, politike in znane osebnosti, druge tri skupine pa blagovne znamke, novice in dogodke. V vsaki izmed skupin smo izbrali tri ključne besede – eno poznano v svetu (globalna), drugo popularno v Evropi in Ameriki (Evropa & ZDA) ter tretjo poznano v Sloveniji (lokalna). Ključne besede smo v vsako skupino izbrali glede na to, koliko imajo zadetkov v iskalniku Google na posameznem družbenem mediju (za vsako ključno besedo smo sešteli število zadetkov v Googlu na straneh Facebook, Twitter, YouTube, Google+, Tumblr). Ključne besede iz globalne skupine imajo 7,7–12,5 milijona zadetkov, iz skupine Evropa&ZDA 2,5–4,3 milijona zadetkov ter iz lokalne skupine manj kot 50.000 zadetkov. S pomočjo API-jev smo iskali objave, ki vsebujejo samo izbrane ključne besede, ne pa njihovih sinonimov, izpeljank, sklanjatev ipd. Vse izbrane ključne besede z osnovnimi podatki so predstavljene v tabeli 2.

Števila v oklepajih označujejo število zadetkov v iskalniku Google za posamezne ključne besede. Krajšavi K in M pomenita 10^3 in 10^6.

V naslednjih razdelkih predstavljamo rezultate analize podatkov iz družbenih medijev, zbranih med 14. 11. in 31. 12. 2017. Izbrani družbeni mediji in opis ključnih besed so predstavljeni v prejšnjem razdelku. Poudarek analize je na razlikah med skupinami ključnih besed ter opazovanju, kako se razlikuje obnašanje uporabnikov med posameznimi družbenimi mediji.

Analizo smo opravili v treh delih.

A. Doseg in aktivnost. Doseg ključne besede označuje število ljudi, ki jih je dosegla določena objava (izračunamo ga kot vsoto sledilcev lastnika objave na Twitterju, število ogledov posnetka na YouTubu ter število všečkov strani, na kateri je objavljena objava, na Facebooku). Aktivnost ključne besede pomeni število aktivnosti objave, kar je seštevek všečkov, komentarjev ali delitev objave na vseh družbenih medijih.

B. Osnovna statistika. Za vsako ključno besedo smo izračunali statistike, kot so število uporabnikov, objav in komentarjev, dobljenih v izbranem časovnem obdobju.

C. Jezikovna analiza in analiza sentimenta objav ter identifikacija spola uporabnikov objav. Za vsako objavo smo določili jezik, sentiment in spol uporabnika.

A. Doseg in aktivnost

Dosegi in aktivnosti objav z izbranimi ključnimi besedami so predstavljeni v tabeli 3. Med globalnimi ključnimi besedami ima največji doseg Cristiano Ronaldo, največjo aktivnost pa Rihanna. Večjo aktivnost kot doseg ima samo Rihanna, kar nakazuje, da o pev-

Tabela 2: **Ključne besede, uporabljene v analizi**

	Globalne	**Evropa & ZDA**	**Lokalne (Slovenija)**
Šport	Cristiano Ronaldo Nogometaš (7,7 M)	Usain Bolt Atlet, sprinter (2,5 M)	Anže Kopitar Hokejist (49,4 K)
Politika	Hillary Clinton Političarka v ZDA (7,9 M)	Arnold Schwarzenegger Igralec in politik (2,9 M)	Miro Cerar Bivši predsednik vlade (36,2 K)
Znane osebnosti	Rihanna Pevka (11,7 M)	Melania Trump Prva dama ZDA (2,4 M)	Jan Plestenjak Pevec (37,3 K)
Znamke	Adidas Športna oblačila (12,5 M)	Motorola Telekomunikacijsko podjetje (3,0 M)	Alples Pohištveno podjetje (25,4 K)
Topic	Pokemon Go Resničnostna igra (9,5 M)	Brexit Izstop Velike Britanije iz Evropske unije (4,3 M)	Drugi tir Referendum za gradnjo novega tira (24,2 K)
Dogodki	Dirka po Franciji Kolesarska dirka (9,7 M)	Oktoberfest Nemški festival (3,1 M)	Odprta kuhna Kulinarična tržnica (37,3 K)

ki objavljajo aktivni uporabniki, ki o njej veliko pišejo in berejo. V skupini Evropa & ZDA ter med lokalnimi ključnimi besedami, imata Brexit in Anže Kopitar največji doseg in aktivnost. Z izjemo ključne besede Brexit, doseg in aktivnost ključnih besed ustrezata številu zadetkov v iskalniku Google – ključne besede z največ zadetki imajo tudi največji doseg in aktivnost. V splošnem pa iz rezultatov dosega in aktivnosti ne moremo narediti drugih sklepov glede razlik med ključnimi besedami oziroma skupinami ključnih besed.

Tabela 3: **Doseg in aktivnost ključnih besed**

Ključna beseda	Doseg	Aktivnost
Cristiano Ronaldo	39.580 M	4.561 M
Usain Bolt	870 M	736 M
Anže Kopitar	55 M	500 K
Hillary Clinton	23.952 M	17.050 M
Arnold Schwarzenegger	771 M	248 M
Miro Cerar	16 M	50 K
Rihanna	22.026 M	50.043 M
Melania Trump	8.431 M	1.364 M
Jan Plestenjak	101 K	38
Adidas	20.003 M	7.300 M
Motorola	13.972	183 M
Alples	20 K	7
Pokemon go	2.902 M	465 M
Brexit	61.858 M	5.404 M
Drugi tir	923 K	2 K
Dirka po Franciji	824 M	55 M
Oktoberfest	87 M	1 M
Odprta kuhna	112	16

Krajšavi K in M pomenita 10^3 in 10^6.

B. Osnovne statistike

V tabeli 4 so predstavljeni rezultati osnovnih statistik za ključne besede število objav, uporabnikov in komentarjev, ki so bili pridobljeni iz družbenih medijev za posamezne ključne besede. Pri interpretaciji rezultatov je treba upoštevati, da Tumblr API ne omogoča pridobivanja komentarjev objav, iz Facebookoveg API-ja pa niso dosegljivi podatki o uporabniku, ki je objavil neko objavo.

Rezultati pokažejo, da uporabniki na Twitterju več objavljajo, saj ima Twitter največje skupno število objav (zadnji dve vrstici tabele 4). Po drugi strani na Facebooku uporabniki raje komentirajo. To potrjujejo rezultati razmerja med objavami in komentarji. Razmerje je večje kot 1 v primeru Facebooka in YouTuba, kar pomeni, da na obeh družbenih medijih uporabniki raje komentirajo. Drugače pa je na Twitterju in Google+, na katerih raje objavljajo kot komentirajo (razmerje je manjše kot 1). Razmerje med objavami in uporabniki je večje kot 1 le v primeru YouTuba, kar nakazuje aktivnost manjšega števila uporabnikov pri objavljanju in komentiranju. Na drugih omrežjih pa je bolj aktivnih več različnih uporabnikov.

Iz rezultatov glede na skupine ključnih besed po tematiki (zadnji stolpec tabele 4), opazimo, da uporabniki na vseh medijih največ objavljajo o aktualnih novicah. Prav tako je veliko objavljenega o znanih osebnostih in politikih, najmanj pa o dogodkih. Z izjemo ključne besede Brexit vse druge ključne besede po številu objav sledijo enakemu vrstnemu redu, kot je vrstni red po številu zadetkov na iskalniku Google (glej tabelo 2).

Na Facebooku uporabniki najraje objavljajo o znanih osebnostih in športnikih. Tudi ključne besede iz skupine blagovnih znamk imajo veliko število objav, kar lahko pripišemo popularnosti oglaševanja na Facebooku. Tej domnevi ustreza tudi manjše število komentarjev teh objav. Na Twitterju in YouTubu uporabniki najraje objavljajo o aktualnih novicah in znanih osebnostih, na Google+ o politikih, na Tumblrju pa je največ objav v zvezi z blagovnimi znamkami in znanimi osebnostmi.

C. Jezikovna analiza, sentiment in identifikacija spola

V zadnjem delu analize smo raziskali še jezik in sentiment objav ter spol uporabnikov objav. Za identifikacijo jezika smo uporabili Googlovo knjižnico za programski jezik Python (Langdetect, 2018). Rezultate prikazujemo za šest najpogostejših jezikov in slovenščino. Za sentiment in identifikacijo spola smo uporabili naivni Bayesov klasifikator in klasificirali vsako objavo v pozitivni ali negativni razred v primeru sentimenta ter v ženski in moški spol v primeru identifikacije spola. Obe klasifikaciji delujeta po principu najpogostejših besed, uporabljenih v objavah (npr. moški uporabljajo v objavah določene besede pogosteje kot ženske, podobno negativne objave vsebujejo druge besede kot pozitivne) (Schwartz, 2013), klasifikator se uči na obstoječi bazi objav z znanim sentimentom in spolom uporabnika.

Rezultati analize so predstavljeni v tabeli 5. Izkazalo se je, da je algoritem za določanje jezika objav določil jezik več kot 75 odstotkom objav na Google+, Tumblrju in YouTubu. Odstotek je veliko manjši za

Tabela 4: **Osnovne statistike ključnih besed: število objav, uporabnikov in komentarjev, dobljenih za posamezne ključne besede v analiziranih družbenih medijih**

Ključna beseda	**Facebook**			**Twitter**			**Google+**			**Tumblr**		**YouTube**			**Skupna vsota**
	Objave	**Uporabniki**	**Komentarji**	**Objave**	**Uporabniki**	**Komentarji**	**Objave**	**Uporabniki**	**Komentarji**	**Objave**	**Uporabniki**	**Objave**	**Uporabniki**	**Komentarji**	
Cristiano Ronaldo															
Usain Bolt	120.821	163.941	480.663	1.810.987	638.202	35.628	5.074	2.913	2.277	2.284	454	15.570	119.069	161.780	3.559 K
Anže Kopitar															
Hillary Clinton															
A. Schwarzenegger	21.260	27.813	96.559	3.767.288	702.988	159.548	10.433	7.172	14.333	1.923	785	13.497	54.389	113.674	4.992 K
Miro Cerar															
Rihanna															
Melania Trump	331.091	163.379	1.520.378	3.996.206	1.526.960	155.653	11.617	7.185	7.868	6.950	2.393	19.379	48.591	62.390	7.859 K
Jan Plestenjak															
Adidas															
Motorola	193.412	85.021	394.521	1.829.267	790.94	143.101	7.872	6.959	8.154	7.772	3.447	18.075	87.907	111.391	3.686 K
Alples															
Pokemon go															
Brexit	99.217	97.011	416.955	7.090.994	1.011.360	636.486	11.400	7.703	8.952	4.801	1.668	25.852	131.844	332.921	9.876 K
Drugi tir															
Dirka po Franciji															
Oktoberfest	15.029	24.278	34.833	77.579	50.876	7.002	930	823	654	286	100	3.400	4.184	2.731	222 K
Odprta kuhna															
Skupna vsota	780 K	561 K	2.943 K	18.572 K	4.721 K	1.137 K	47 K	32 K	42 K	24 K	9 K	96 K	446 K	785 K	30.199 K
Razmerje		0,72	3,77		0,25	0,06		0,69	0,89		0,37		4,65	8,18	3.559 K

Krajšavi K in M pomenita 10^3 in 10^6.

Twitter (26,8 %) in Facebook (41,4 %). Za vse družbene medije velja, da je prek 73 odstotkov objav napisanih v angleščini. Drugi najpogostejši jezik je španščina na Facebooku in Tumblrju, portugalščina na Twitterju in nemščina na Google+ in YouTubu.

Rezultati analize sentimenta pokažejo, da je tudi sentiment v najmanjšem odstotku določljiv objavam na Facebooku in Twitterju (manj kot 32 %). V drugih družbenih medijih je odstotek večji, prek 50 odstotkov. Večina objav v vseh medijih je negativnih, največ od tega na Twitterju, najmanj pa na Tumblrju.

Največjemu odstotku objav je spol uporabnika določljiv na Google+ (več kot 80 %), najmanj pa na Twitterju (manj kot 27 %). Sicer večino objav objavljajo uporabniki moškega spola, najnižji odstotek moških uporabnikov pa je na Twitterju (tam je 25 % ženskih uporabnic).

V splošnem so objave na Twitterju in tudi Facebooku najmanj primerne za analizo jezika, sentimenta in spola. Vse troje smo lahko določili manj kot 42 odstotkom objav na Facebooku in manj kot 27 odstotkom objav na Twitterju. Za druge medije je delež večji kot 56 odstotkov. Skupna vsem družbenim medijem je angleščina kot večinski jezik objav (več kot 73 %). Prav tako je večina objav v vseh medijih negativnih (med 50 in 70 %); večino objav so napisali moški uporabniki (74–90 %). Glede na vse rezultate ugotavljamo, da v primeru uporabe podatkov samo iz enega medija lahko pridemo do zavajajočih ugotovitev pri analizi jezika, sentimenta in spola objav.

4 SKLEP

V prispevku se ukvarjamo z analizo podatkov iz različnih družbenih medijev ter njihovo uporabnostjo v raziskovalne namene. Za izbrane ključne besede z različnih področij, kot so šport, politika in dogodki, smo raziskali, kako pogosto se pojavljajo v petih medijih (Twitter, Facebook, Google+, Tumblr in YouTube). Družbene medije smo primerjali med seboj glede na doseg in aktivnost ključnih besed ter jezik in sentiment objav.

V raziskavi uporabljene ključne besede smo izbrali glede na število zadetkov v iskalniku Google. Rezultati analize so pokazali, da so ključne besede z največ zadetki tudi v družbenih medijih uporabljene v največ objavah. Izkazalo se je, da je Twitter najpriljubljenejši medij glede na analizirano statistiko, naslednji za njim je Facebook. Razlike med mediji opazimo pri načinu objavljanja – na Twitterju uporabniki več objavljajo, na Facebooku pa raje komentirajo objave. Pri opazovanju statistike objav glede na ključne besede opazimo, da uporabniki vseh medijev najraje govorijo o znanih osebnostih in politikih, prav tako so popularne trenutno aktualne teme, kot je na primer Brexit. V zadnjem delu analize smo analizirali še jezik in sentiment objav ter spol uporabnikov objav. Izkazalo se je, da ima Twitter največji delež negativnih objav ter skupaj s Facebookom najnižji delež uporabnic. Skupni vsem medijem so večinski jezik angleščina, večji delež negativnih kot pozitivnih objav ter večinski delež moških uporabnikov.

Tabela 5: **Rezultati analize jezika in sentimenta objav ter identifikacije spola uporabnikov objav**

	Facebook		**Twitter**		**Google+**		**Tumblr**		**YouTube**	
Vseh objav	780.830		1.870.565		48.668		23.916		95.953	
Objave z jezikom	322.046	41,2 %	5.004.991	26,8 %	42.003	86,3 %	18.262	76,4 %	72.199	75,2 %
Nemški	7.841	2,4 %	110.924	2,2 %	372	0,9 %	431	2,4 %	4.654	6,4 %
Angleški	146.555	76,6 %	4.482.096	89,6 %	39.337	93,7 %	13.475	73,8 %	53.879	74,6 %
Španski	12.966	4,0 %	123.941	2,5 %	91	0,2 %	563	3,1 %	1.905	2,6 %
Francoski	4.074	1,3 %	46.899	0,9 %	18	0,0 %	29	0,2 %	84	0,1 %
Italijanski	3.690	1,1 %	67.186	1,3 %	164	0,4 %	279	1,5 %	2.149	3,0 %
Portugalski	6.642	2,1 %	176.578	3,5 %	185	0,4 %	455	2,5 %	2.230	3,1 %
Slovenski	674	0,2 %	11.866	0,2 %	99	0,2 %	42	0,2 %	1.175	1,6 %
Objave s sentimentom	147.229	31,7 %	4.093.965	24,0 %	39.436	81,0 %	13.517	56,5 %	55.054	57,4 %
Pozitivne	161.528	65,3 %	3.128.212	69,6 %	24.443	62,0 %	6.870	50,8 %	33.410	60,7 %
Negativne	85.699	34,7 %	1.365.753	30,4 %	14.993	38,0 %	6.647	49,2 %	21.644	39,3 %
Objave s spolom	247.220	31,7 %	4.980.829	26,6 %	40.184	82,6 %	14.345	60,6 %	61.902	64,5 %
Ženski	53.080	21,5 %	1.264.623	25,4 %	4.977	12,4 %	2.699	18,8 %	6.417	10,4 %
Moški	194.149	78,5 %	376.206	74,6 %	35.207	87,6 %	11.646	81,2 %	55.485	89,6 %

Iz rezultatov analize se je pokazalo, da se posamezni mediji ne razlikujejo samo po priljubljenosti, temveč tudi po načinu objavljanja ter obnašanju uporabnikov. Osredotočanje na en sam medij za namene različnih raziskav lahko vodi do nepravih sklepov, zato je pri analizah smotrno uporabiti podatke iz več različnih medijev. Prispevek torej predstavlja ogrodje za uporabo vsebin iz družbenih medijev. Ključnega pomena pri tem je vedenje, kaj lahko iz katerega medija pridobimo ter kakšni so interesi uporabnikov posameznega družbenega medija. V prispevku prikazujemo primer pridobivanja podatkov in njihovo analizo za različne ravni in domene, kar lahko služi kot vodilo pri analizah na novih področjih. Opozarjamo pa, da je pri uporabi predlaganega pristopa pomembno upoštevati tudi dejstvo, da se tip uporabnikov, vsebina in funkcionalnosti družbenih medijev ves čas spreminjajo. V nadaljnjih raziskavah se bomo osredotočili na obnašanje uporabnikov pri objavljanju v različnih družbenih medijih. Raziskali bomo, ali med mediji obstajajo skupni vzorci, kdaj uporabniki objavljajo največ ter ali nenadni dogodki spremenijo te vzorce obnašanja.

5 VIRI IN LITERATURA

[1] Anantharam, P., Barnaghi, P., Thirunarayan, K. in Sheth, A. (2015). Extracting city traffic events from social streams. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *6*(4), 43.

[2] Aramaki, E., Maskawa, S. in Morita, M. (2011). Twitter catches the flu: detecting influenza epidemics using Twitter. V *Proceedings of the conference on empirical methods in natural language processing* (str. 1568–1576). Association for Computational Linguistics.

[3] Bourlai, E. in Herring, S. C. (2014, junij). Multimodal communication on Tumblr: I have so many feels! V: *Proceedings of the 2014 ACM conference on Web science* (str. 171–175). ACM.

[4] Facebook API. Objavljeno na https://developers.facebook.com/ (zadnji ogled februarja 2018).

[5] Felt, M. (2016). Social media and the social sciences: How researchers employ Big Data analytics. *Big Data & Society*, *3*(1), 2053951716645828.

[6] Figueiredo, F., Benevenuto, F., in Almeida, J. M. (2011, februar). The tube over time: characterizing popularity growth of YouTube videos. In *Proceedings of the fourth ACM international conference on Web search and data mining* (str. 745–754). ACM.

[7] Google+ API. Objavljeno na https://developers.google.com/+/ (zadnji ogled februarja 2018).

[8] Gundecha, P. in Liu, H. (2012). Mining social media: a brief introduction. *Tutorials in Operations Research*, *1*(4), 1–17.

[9] Kaplan, A. M., in Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons*, *53*(1), 59–68.

[10] Langdetect. Objavljeno na https://pypi.python.org/pypi/langdetect? (zadnji ogled februarja 2018).

[11] Lomborg, S. in Bechmann, A. (2014). Using APIs for data collection on social media. *The Information Society*, *30*(4), 256–265.

[12] Lotan, G., Graeff, E., Ananny, M., Gaffney, D. in Pearce, I. (2011). The Arab Spring| the revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *International journal of communication*, *5*, 31.

[13] Moore, K. in McElroy, J. C. (2012). The influence of personality on Facebook usage, wall postings, and regret. *Computers in Human Behavior*, *28*(1), 267–274.

[14] Naaman, M., Boase, J. in Lai, C. H. (2010, februar). Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (str. 189–192). ACM.

[15] Osborne, M. in Dredze, M. (2013). Facebook, Twitter and Google Plus for breaking news: Is there a winner? V: *ICWSM*.

[16] Petrovic, S., Osborne, M., McCreadie, R., Macdonald, C., Ounis, I. in Shrimpton, L. (2013). Can twitter replace newswire for breaking news? V: *ICWSM*.

[17] Reuter, C. in Scholl, S. (2014). Technical Limitations for Designing Applications for Social Media. V: *Mensch & Computer Workshopband* (str. 131–139).

[18] Ruths, D. in Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063–1064.

[19] Sakaki, T., Okazaki, M. in Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. V: *Proceedings of the 19th international conference on World wide web* (str. 851–860). ACM.

[20] Schäfer, M. T. in van Es, K. (2017). *The datafied society: Studying culture through data*. Amsterdam University Press.

[21] Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M. in Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, *8*(9), e73791.

[22] Siersdorfer, S., Chelaru, S., Nejdl, W. in San Pedro, J. (2010). How useful are your comments?: analyzing and predicting YouTube comments and comment ratings. V: *Proceedings of the 19th international conference on World wide web* (str. 891–900). ACM.

[23] Statista: Number of daily active Facebook users worldwide as of 3rd quarter 2018 (in millions). Objavljeno na https://www.statista.com/statistics/346167/facebook-global-dau/ (zadnji ogled decembra 2018).

[24] Tufekci, Z. (2014). Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. *ICWSM*, *14*, 505–514.

[24] Tumblr API. Objavljeno na https://www.tumblr.com/docs/en/api/v2 (zadnji ogled februarja 2018).

[25] Tumblr API. Objavljeno na https://developer.twitter.com/en/docs (zadnji ogled februarja 2018).

[26] Xu, J., Lu, T. C., Compton, R. in Allen, D. (2014, april). Civil unrest prediction: A tumblr-based exploration. V: *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* (str. 403–411). Springer, Cham.

[27] YouTube API. Objavljeno na https://developers.google.com/youtube/ (zadnji ogled februarja 2018).

[28] Zimmer, M. (2010). »But the data is already public«: on the ethics of research in Facebook. *Ethics and information technology, 12*(4), 313–325.

Neli Blagus je raziskovalka v Laboratoriju za podatkovne tehnologije na Fakulteti za računalništvo in informatiko Univerze v Ljubljani. Po opravljenem doktoratu iz področja analize omrežij se še naprej ukvarja z raziskavami na področju družbenih omrežij in rezultate redno objavlja v mednarodnih znanstvenih revijah.

▪

Slavko Žitnik je docent na Fakulteti za računalništvo in informatiko Univerze v Ljubljani, kjer poučuje predmete s področja podatkovnih baz in obdelave podatkov. Raziskovalno se ukvarja z obdelavo naravnega jezika, predvsem na semantični ravni. Je predsednik Sveta za elektronske komunikacije RS, sodeluje pri organizaciji konferenc s področja informatike in pri projektih, povezanih z obdelavo podatkov na področju interneta stvari.

▪

Marko Bajec je redni profesor na Fakulteti za računalništvo in informatiko Univerze v Ljubljani, kjer poučuje dodiplomske in podiplomske predmete s področja razvoja informacijskih sistemov in podatkovnih baz. Raziskovalno se ukvarja z metodami in pristopi k snovanju in razvoju informacijskih sistemov in obvladovanjem informatike ter v zadnjih letih predvsem s podatkovnimi tehnologijami za predstavitev, analizo in vizualizacijo podatkov. Leta 2009 je ustanovil Laboratorij za podatkovne tehnologije ter prevzel njegovo vodenje. Je član številnih domačih in tujih združenj, komisij in odborov. V okviru fakultete je vodil več aplikativnih in raziskovalnih projektov. Svoje raziskovalne rezultate in dosežke iz prakse redno objavlja v domačih in mednarodnih znanstvenih in strokovnih krogih.