IMPAKTI I BIG DATA NË METODAT TRADICIONALE TË KËRKIMIT

ELA DUKAJ (VRENOZI).[1], ANA KTONA.[2]

[1]Instituti i Kërkimeve të Tregut (GFS Research), Liqeni i Thatë, Tiranë, Shqipëri
[2]Universiteti i Tiranës, Fakulteti i Shkencave të Natyrës, Departamenti i Informatikës
email: elavrenozi@yahoo.com

Përmbledhje

Me rritjen e Big Data vitet e fundit, disa mund të mendojnë se, duke pasur teknikat e duhura dhe aftësinë për të manovruar me këto Big Data, do gjejmë përgjigjet për të gjitha pyetjet tona, ndaj metodat tradicionale të kërkimit mund tëzëvendësohen plotësisht. Qëllimi i këtij studimi është të shpjegojë përse Big Data nuk duhet konsideruar si zëvendësuese e metodave tradicionale të kërkimit dhe sesi këto dy burime të dhënash mund të kombinohen për të gjeneruar avantazhe të mëdha nga njëra-tjetra. Për këtë arsye janë analizuar grupet e studenteve në facebook "Informatika FSHN" dhe "TIK FSHN" nëpërmjet kombinimit të Facebook API dhe programimit R ku janë evidentuar tematikat më të diskutuara në postimet dhe komentet nga anëtaret e grupeve, ndër to drejtimi "Master Profesional". Për të shpjeguar këtë trend janë zhvilluar 277 intervista me studentë të TIK dhe Informatikë Viti 3 bachelor. Nëpërmjet analizimit të përgjigjeve shpjegohet arsyeja e orientimit të studentëve në drejtimin master profesional (hyrja sa më herët në tregun e punës). Një studim tjetër është zhvilluar në fushën e politikës, ku në bazë të analizës së faqes publike në facebook të kryetarit të bashkisë së Barit, Itali, është vënë re një rritje e vlerësimit për këtë politikan nëpërmjet numrit në rritje të pëlqimeve dhe komenteve pozitive. Për të shpjeguar këtë trend, janë zhvilluar 600 intervista telefonike, ku qytetarët shprehin arsyet se PSE e vlerësojnë këtë personazh, (besueshmëria, idetë inovative etj, janë disa nga karakteristikat e evidentuara nga të intervsituarit). Pra Big Data mund të na tregojë çfarë ka ndodhur në të shkuarën, ndoshta me modele përkatëse edhe të parashikojë disa evente të ardhshëm, por nuk mund të shpjegojë saktësisht se PSE diçka ka ndodhur. Për të kutpuar PSE-në na nevojiten metodat tradicionale të kërkimit, si një burim të dhënash komplementar për Big Data.

Fjalëkyçe: Big Data, metoda tradicionale të kërkimit, burim të dhënash, programim R.

Abstract

With the Big Data growth in recent years, some may think that by having the right techniques and the ability to handle these Big Data, we may find answers to all of our questions, so traditional survey research can be completely replaced. The purpose of this study is to explain why Big Data should not be considered as a substitute for traditional survey research and how these two data sources can be combined to generate great advantages from each another. For this reason, the "Informatics FNS" and "ICT FNS" student groups in facebook, were analyzed through the combination of Facebook API and R programming, where the most

discussed topics in posts and comments were highlighted, among them as a more discussed topic is the "Professional Master". To explain this trend, 277 interviews were conducted with ICT and Informatics students of third year in bachelor program. By analyzing the answers, the model generated explains the reason for the orientation of the students in the Professional Master program (entry into the labor market as soon as possible). Another study has been conducted in the politics field. The analysis of the facebook public page of the Bari (Italy) mayor, shows that there has been a growing level of appreciation for this politician through the growing number of positive comments and likes. This trend si followed by a survey research, for which 600 telephone interviews have been conducted, where citizens express the reasons WHY they appreciate this politician (reliable, innovative ideas, etc,.). So Big Data can tell us what has happened in the past, maybe predict some future events by using relevant models, but it can not explain exactly WHY the specific event has happened. To understand the WHY we need the traditional survey research, as a complementary data source for Big Data.

Key words: Big Data, traditional survey research, data source, R programming.

Hyrje

Vitet e fundit kemi dëshmuar një rritje të konsiderueshme të sasisë së statistikave që përshkruajnë fenomene të ndryshëm në shoqëri bazuar në Big Data. Ndërsa në të shkuarën vuanim nga mungesa e informacionit ndaj nevojiteshin studimet e tregut për të mbushur boshllëqet e njohurive, sot jemi me fat që kemi bollëk informacioni. Çdo ditë ne gjenerojmë 2.5 kuintilion bajt të dhëna që vijnë nga burime të ndryshme: postime në rrjete sociale, foto dhe video dixhitale, rekorde transaksionesh, sinjale GPS etj,. Si rezultat disa mund të mendojnë se, duke pasur teknikat e duhura dhe aftësinë për të manovruar me Big Data, do gjejmë përgjigjet për të gjitha pyetjet tona, ndaj metodat tradicionale të kërkimit mund të zëvendësohen plotësisht (Couper & Mick P. *et al.*, 2013). Big Data nuk është zgjidhja universale. Përdorimi i metodave statistikore të vlefshme mbi Big Data është një sfidë e vërtetë, kjo për shkak të varietetit të të dhënave (më shumë se 80% sot janë të dhëna të pastrukturuara). Gjithashtu ekzitson edhe një ide e gabuar që sasia e të dhënave mund të zëvendësojë çfarëdolloj defiçiti në të dhëna. (Japec, 2015; Fan *et al.*, 2014).

Shumica e kërkimeve të tregut, politike apo sociale bëhen për të kuptuar arsyet e zhvillimit të një ngjarjeje (Biemer & Paul P. *et al.*, 2015). Big Data mund të na tregojë çfarë ka ndodhur në të shkuarën, dhe ndoshta me modele përkatëse edhe të parashikojë disa evente të ardhshëm, por nuk mund të shpjegojë saktësisht se PSE diçka ka ndodhur. Pa kuptuar PSE-në, Big Data-t nuk janë edhe aq të zbatueshme (Fan *et al.*, 2014), ndaj edhe dy gjigandët e Big Data sot, Google dhe Facebook, shpesh përdorin metoda tradicionale survejimi për te kuptuar më shumë mbi përdoruesit e tyre. Qëllimi i këtij studimi është të shpjegojë mbi kërkime reale (në fushën e arsimit dhe në politikë), se përse Big Data nuk duhet konsideruar si zëvendësuese e metodave tradicionale të kërkimit dhe sesi këto dy burime të dhënash mund të kombinohen për të gjeneruar avantazhe të mëdha nga njëra-tjetra. Për të

realizuar këtë qëllim janë kryer dy studime paralele në fushën e arsimit dhe konkretisht pranë Fakultetit të Shkencave të Natyrës me studentët e viteve të treta të Informatikës dhe TIK, si edhe në fushën e politikës ku sondazhet janë gjithmonë një burim kryesor informacioni për çdo parti apo personazh politik.

Metodologjia

Punimi bazohet në dy studime paralele ku krahasohen të dhënat e analizuara nga rrjetit social facebook dhe sondazheve të drejtpërdrejta.

Studimi është kryer për të kuptuar orientimin e studentëve në programet master.

Nëpërmjet gjuhës së programimit R janë mbledhur dhe analizuar të dhënat nga grupet e studentëve të vitit të tretë në degën e Informatikës dhe Teknologjisë së Informacionit në Fakultetin e Shkencave të Natyrës. Këto të dhëna i analizuam duke përdorur metoda statistikore të grupimit dhe text mining që ofron katalogu i paketave në programim R. Gjithashtu hartuam një pyetësor me disa pyetje demografike si gjinia, mosha, vendbanimi, punësimi, pyetje mbi njohuritë profesionale në fushën e teknologjisë së informacionit dhe preferencat e tyre për drejtimin master. Linku i pyetësorit iu dërgua me email të gjithë studentëve të viteve të treta Informatikë dhe Teknologji Informacioni dhe Komunikimi, Fakulteti i Shkencave të Natyrës, si kandidatët për studimet master. U mblodhën 273 pyetësorë.

Paralelisht me këtë studim është zhvilluar një studim në fushën e politikës në qytetin e Barit në Itali. Është studiuar faqja zyrtare në facebook e kryetarit të bashkisë së këtij qyteti për të kuptuar nivelin e pëlqyeshmërisë nga ndjekësit e tij, gjithashtu janë analizuar komentet nëpërmjet text mining për të kuptuar reagimet pozitive dhe ato negative, ku është vënë re një orientim pozitiv ndaj këtj personazhi. Për këtë arsye janë intervistuar nëpërmjet telefonatave 600 banorë të këtij qyteti ku përvec pyetjeve profilizuese si gjinia, mosha, titulli i studimeve, niveli i jetesës, profesioni etj, banorët janë pyetur në lidhje me figurën e kryetarit të tyre të bashkisë, për të kuptuar nivelin real të besueshmërisë dhe të pëlqyeshmërisë së këtij politikani si dhe për të bërë krahasimet nga të dhënat e analizuara nga facebook. Kampioni prej 600 pjesëmarrësish në sondazh është zgjedhur përfaqsues për gjininë, moshën dhe nivelin social, bazuar në të dhënat demografike të marra nga www.demo.istat.it, faqja zyrtare e institutit të statistikave në Itali.

Janë përdorur SPSS për mbledhjen dhe normalizimin e të dhënave të sondazheve dhe Weka për analizimin e tyre. Ndërkohë që mbledhja dhe analizimi i të dhënave nga facebook është bërë duke përdorur facebook API dhe facebook developer tools në kombinim me programim R, ku gjenerimi i grafikëve është realizuar me R ggplot.

Rezultatet

Studimi i parë: Orientimi në studimet master i studentëve të Informatikës dhe TIK në Fakultetin e Shkencave të Natyrës.

Nëpërmjet Big Data sugjerohet se: Shumica e studentëve ndjekin (studiojnë apo diskutojnë) mbi Master profesional.

Vrojtim i përgjithshëm i anëtarëve të grupeve studentore në rrjetin social facebook:

- Master profesional në Informatikë - 327 anëtarë
- Master i Shkencave në Informatikë - 64 anëtarë
- Master profesional në Teknologji informacioni - 98 anëtarë
- Master i Shkencave në Teknologji informacioni - 28 anëtarë

Nga ky vrojtim i përgjithshëm vihet re që grupet e titulluar me Master profesional kanë më shumë anëtarë e si rrjedhim më shumë studentë të interesuar për këtë drejtim sesa grupet e tjera prezente në facebook.

Por le të shohim më hollësisht studimin e grupit më të madh në facebook në lidhje me teknologjinë e informacionit pranë Fakultetit të Shkencave të Natyrës, grupi "Teknologji Informacioni dhe Komunikimi at FSHN" me 2172 anëtarë. Nga analiza e postimeve dhe komenteve të bëra nga studentët anëtarë të këtij grupi u gjenerua grafiku i mëposhtëm në të cilin paraqiten fjalët kyçe të përdorura më së shumti.

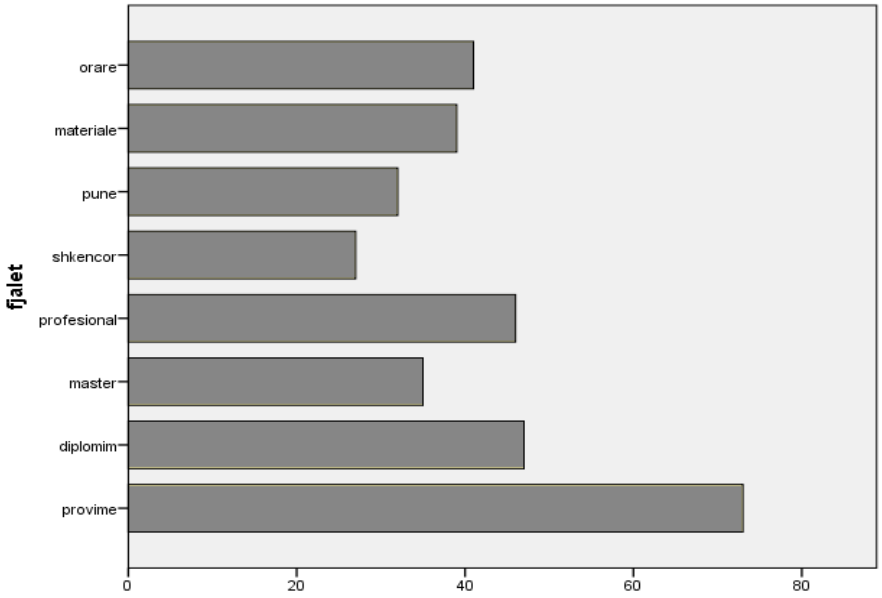

Figura 1: Fjalët më të përdorura në komente nga studentët në grupin e facebook

Sikurse vihet re edhe nga grafiku përveç tematikave më shqetësuese për studentët si *provime* dhe *diplomim*, fjala e tretë më e përdorur në postime dhe komente i referohet masterit profesional çka indikon që ky drejtim është më i preferuari i studentëve të kësaj fushe. Por cilat janë arsyet e kësaj preference? Sigusisht përveç mesatares së tre viteve të studimeve, e cila sikurse e dimë është faktori klasifikues që ndikon në ndarjen në master profesional dhe shkencor të studentëve. Për të kuptuar më mirë këtë orientim na vjen në ndihmë metoda tradicionale e kërkimit nëpërmjet pyetësorëve. Le t'iu referohemi rezultateve të mëposhtëm të gjeneruar nga pyetësorët e mbledhur pas analizës me anë të algoritmit k-means.

18:08:19 - SimpleKMeans

Cluster centroids:

Attribute	Full Data (277)	Cluster# 0 (92)	1 (68)	2 (42)	3 (75)
i»¿DrejtimGjimnaz	0.0217	0.0109	0.0294	0.0238	0.0267
FormA2	0.7617	0.8152	0.7794	0.7381	0.6933
KlasifikPik	2.1336	2.0978	2.1471	2.119	2.1733
Niveli_real	1.4513	1.2283	1.6618	1.119	1.72
VetVlersimi	1.9856	1.7826	2.2059	1.5476	2.28
SistemeSH	0.0722	0.0543	0.1029	0	0.1067
FormimProfesional	0.3141	0.2065	0.3088	0.3333	0.44
masterIA	0.0794	0.0435	0	0	0.24
masterSHW	0.1805	0.2609	0	0	0.3467
masterSistInfSoft	0.2455	0	1	0	0
masterInfBiznes	0.1516	0	0	1	0
masterMesuesi	0.0181	0.0326	0	0	0.0267
masterITmirmbajtje	0.1444	0.413	0	0	0.0267
masterING	0.1336	0.1522	0	0	0.3067
masterLigj	0.0433	0.087	0	0	0.0533
Pune	1.8448	1.8478	1.8676	1.9286	1.7733
PuneDeshire	0.5343	0.0326	0.7353	0.4762	1
Vendbanim	1.213	1.1957	1.2059	1.1905	1.2533
FormimIPergj	0.4332	0.4457	0.3971	0.5238	0.4

Figura 2: Rezultatet e algoritmit k-means mbi të dhënat e mbledhura nga pyetësorët

Algoritmi k-means me katër cluster-a (k=4) nxjerr një rezultat me nivel stë lartë saktësie (100%) përsa i përket orientimit të studentëve në master professional, ku sikurse shohim kemi një vlerë 1 (përputhshmëri të plotë) për drejtimin master profesional në Informatikë Biznesi. Në të njëjtën kolonë ku shfaqet vlera 1 shohim që edhe variabli "Punë' i cili i referohet pyetjeve për punësimin, ka vlerën më të lartë (1.9286) në gjithë rreshtin e vet çka tregon se ky grup është më tepër i interesuar për t'u futur sa më herët në tregun e punës.

Gjithashtu vlera përkatëse e variablit "FormimIPërgjithshëm", që i referohet grupit të pyetjeve për formimin e përgjithshëm në fushën e teknologjisë së informacionit, ka vlerën më të lartë në krahasim me grupet e tjera, çka tregon që jo gjithmonë ata që preferojnë apo mbeten (për shkak të mesatares) në master profesional janë domosdoshmërisht studentët më të dobët (duke qenë mesatarja pikë ndarjeje në pranimin në programet e studimit master profesional dhe shkencor). Pra nëpërmjet pyetësorëve të drejtpërdrejtë (metodë tradicionale kërkimi) arritëm të kuptojmë disa nga arsyet (PSE-të) e

ngritura nga orientimi që na dha analiza e rrjetit social facebook (implementim i Big Data).

Studimi i dytë: Opinioni mbi kryebashkiakun e qytetit të Barit, Itali

Big Data: Nga studimi i faqes zyrtare të facebook, vihet re një rritje e reagimeve pozitive ndaj kryebashkiakut të qytetit të Barit Itali, kjo nga numri në rritje i komenteve, pëlqimeve dhe shpërndarjeve të posteve të tij.

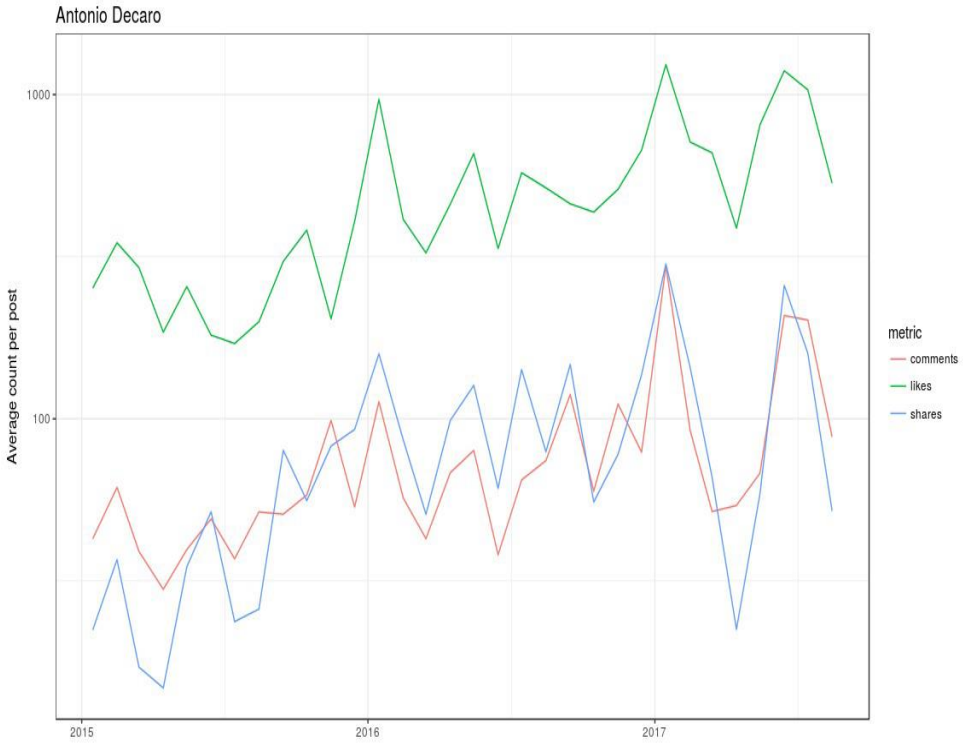

Figura 3: Grafiku i aktivitetit të ndjekësve 3 vitet e fundit në faqen publike të kryebashkiakut

Nga analiza e fjalëve më të përdorura në komente është nxjerrë grafiku i mëposhtëm, i cili tregon se fjalët më të përdorura kanë semantikë pozitive si *bravo - i zoti, bene - mirë, grande - i madh* etj., çka vërteton edhe njëherë një rritje të opinionit pozitiv në lidhje me këtë figure politike.

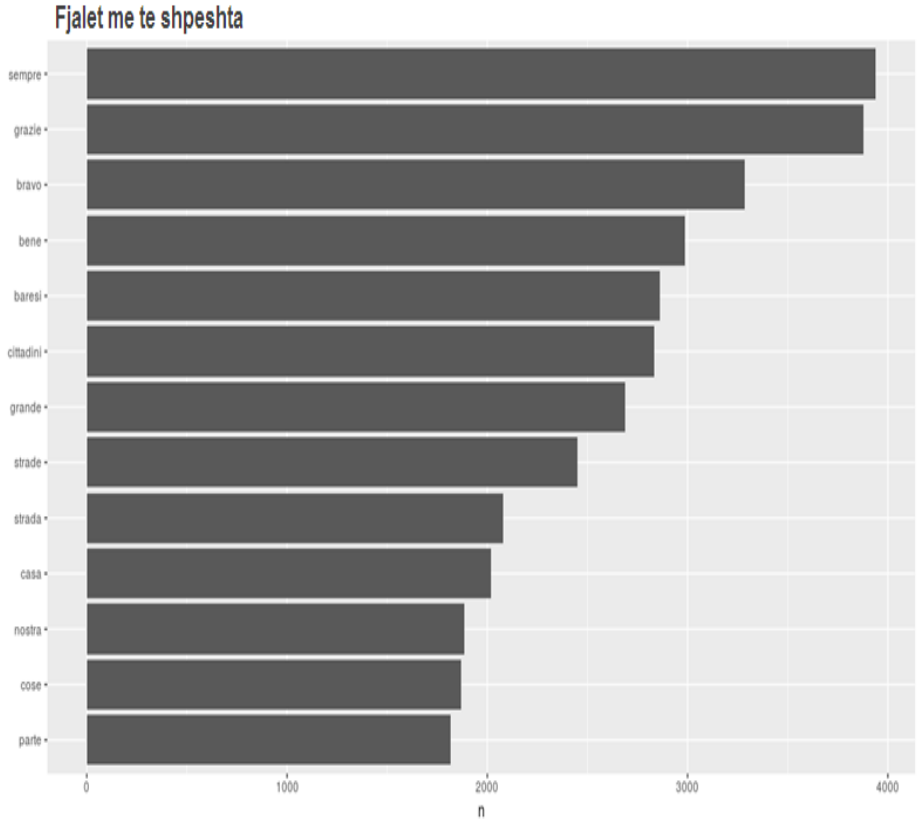

Figura 4: Fjalët më të përdorura në komentet e posteve

Metoda tradicionale të kërkimit: Sërish për të kuptuar më mirë arsyet e këtij orientimi pozitiv na duhet ti drejtohemi pyetjeve të drejtpërdrejta, pra metodave tradicionale të kërkimit. Sigurisht sikurse dihet, personat që pëlqejnë faqen publike të një figure politike janë ata që përgjithësisht e pëlqejnë dhe vlerësojnë këtë figurë ndaj mund të jetë e pritshme që të ketë më tepër reagime pozitive nga ndjekësit specifik të një figure politike (Antenucci *et al.*, 2014). Për këtë arsye për të qënë më asnjanës janë intervistuar nëpërmjet telefonatave 600 banorë në mënyrë të rastësishme për të cilët nuk dihej prirja politike. Nga të dhënat e mbledhura nga intervistimi i banorëvë të këtij qyteti rezultoi se një pjesë e madhe e tyre mendonin se cilësia e jetesës në qytet është përmirsuar shumë gjatë përiudhës së drejtimit të bashkisë nga kryetari aktual.

Sikurse duket edhe në grafikun e mëposhtëm ku pjesa më e madhe e të intervistuarëve pohojnë që niveli i jetesës është përmirsuar shumë dhe mjaftueshëm gjatë tre viteve të fundit, ndaj edhe dëshmojnë një nivel besueshmërie më të lartë karshi kryetarit të tyre të bashkisë.

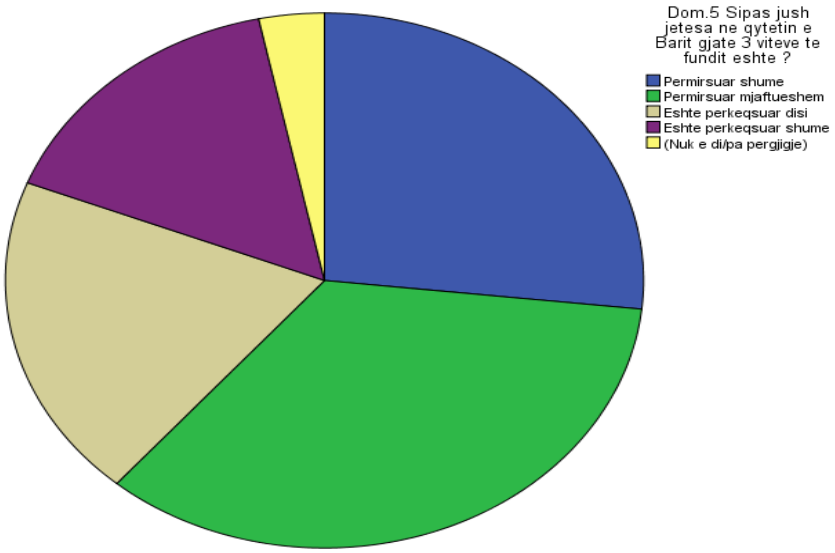

Figura 5: Përgjigjet mbi nivelin e jetesës në qytet gjatë tre viteve të fundit

Për sa më sipër, pra për nivelin e besueshmërisë, kur të intervistuarit u pyetën se sa besim kanë në kryetarin e tyre të bashkisë, shumica prej tyre u përgjigjën shumë dhe mjaftueshëm, çka është në një linjë me orientimin që na dha studimi i faqes facebook. Por nga ana tjetër studimi i faqes facebook ndoshta nuk është edhe aq përfaqësues për qytetin, duke pasur parasysh që përdoruesit e facebook janë kryesisht të rinj si edhe që jo të gjithë ndjekësit e një figure politike si kryetari i bashkisë mund të jenë banorë të qytetit që ai drejton. Pra grupi i ndjekësve të faqes publike të këtij personazhi mund të mos jetë shumë përfaqsues për qytetin e Barit. Metoda tradicionale e kërkimit në këtë rast e mbush këtë boshllëk duke hartuar një kampion me vetëm banorë të qytetit, të zgjedhur në mënyrë përfaqësuese të përbërjes demografike të popullsisë së zonës.

Sikurse shihet edhe në grafikun e mëposhtëm shpërndarja e grup-moshave është gjithëpërfshirëse dhe në proporcion me të dhënat demografike të qytetit, pra me një numër të madh të moshuarish. E ndërsa në facebook kryesisht ndjekësit janë të rinj, shohim që edhe të moshuarit e këtij qyteti kanë një vlerësim mjaftueshëm pozitiv për kryetarin e bashkisë. Pra Big Data na dha një indikacion të saktë dhe një pikë nisjeje për të zhvilluar më pas një sondazh, i cili na konfermoi edhe njëherë orientimin pozitiv por gjithashtu plotësoi hapësirat e krijuara duke përdorur kampion gjithëpërfshirës dhe pyetje specifike për të gjetur edhe arsyet konkrete.

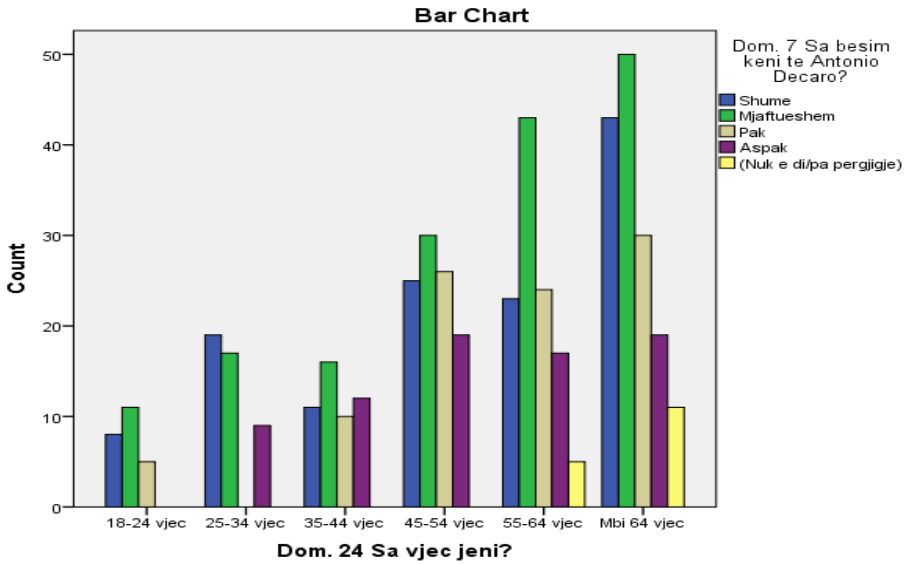

Figura 6: Niveli i besueshmërisë sipas grup-moshave të qytetarëve rezidentë në Bari

Përfundime

Bazuar në studimin e kryer rezulton se studentët që janë të orientuar në masterin profesional jo gjithmonë janë më të dobët nga ana e njohurive profesionale se studentët e tjerë, por ata kërkojnë që të hyjnë më herët në tregun e punës.

Në fushën politike Big Data na dha një indikacion të saktë në lidhje me një figurë politike e cila na ndihmoi më pas të hartonim një sondazh për të plotësuar më mirë tablonë për këtë përsonazh. Në pikëpamje më të gjerë mund të themi se:

Big Data na ofron mundësi të reja për matjen e sjelljeve njerëzore, duke marrë të dhëna nga web dhe duke analizuar rrjetet sociale (Antenucci , 2014; Couper & Mick P. *et al.*, 2013). Rasti i grupeve të studentëve dhe faqes publike të personazhit politik në këtë studim.

Big Data mund të përdoret për të gjetur modele dhe vendosur marrëdhënie ndërmjet disa faktorëve. Në rastin e studimit të figurës së një personazhi politik vendosëm marrëdhënie ndërmjet nivelit të besueshmërisë dhe të dhënave demografike, në rastin e studentëve u vendos marrëdhënie ndërmjet preferencës për master profesional, nivelit të njohurive dhe synimit për në tregun e punës. Big Data-t edhe pse nuk iu përgjigjën të gjitha pyetjeve tona, padyshim janë një shtesë tmerrësisht e rëndësishme në kërkim dhe shërbejnë si plotësuese e metodave tradicionale të hulumtimit.

Kërkimi tradicional kryhet për t'iu pergjigjur pyetjeve dhe mënyra më e mirë për të filluar është duke studiuar gjithë informacionin disponibël (Japec, 2015; Biemer & Paul P. *et al.*, 2010). Falë studimit paraprak të facebook (fushë e Big Data) arritëm të krijonim pyetësorë përkatësisht për studentët ë

Informatikës dhe TIK dhe për banorët e qytetit të Barit, bazuar në atë çfarë Big Data na ofroi dhe informacionet për të cilat kishim nevojë për më tepër pyetje dhe përgjigje. Pra Big Data na shërbeu si një burim që ofroi mënyra të reja për të drejtuar pyetje tradicionale si edhe për të krijuar pyetje të reja. Nga ana tjetër metodat tradicionale të kërkimit na siguruan të dhëna më specifike për të shpjeguar trendet që Big Data zbuloi. Ndaj mund të themi se të dy këto burime të rëndësishme të dhënash duhen konsieruar si komplementare dhe jo si konkurente të njëra-tjetrës.

Literatura

Lilli Japec, Frauke Kreuter, Marcus Berg, Paul Biemer, Paul Decker, Julia Lane, Cathy O'Neil,Abe Usher (2015): American Association for Public Opinion Research (AAPOR) Big Data Task Force

Antenucci, Dolan, Michael Cafarella, Margaret Levenstein, Christopher Ré, and Matthew D. Shapiro (2014): Using Social Media to Measure Labor Market Flows

Biemer, Paul P. (2010): Total survey error: Design, implementation, and evaluation. Public Opinion Quarterly, 74(5): 817-848

Couper, Mick P. (2013): Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys. Survey Research Methods, 7(3): 145-156

Fan, Jianqing, Fang Han, and Han Liu. (2014): Challenges of Big Data analysis. National Science Review, 1: 293-314