Računalna statistička analiza jezika religijskih rasprava na internetskim forumima

Autor: Josip Torić
Mentor: izv. prof. dr. sc. Jan Šnajder

Fakultet elektrotehnike i računarstva, Sveučilište u Zagrebu

11. srpnja 2018.

Uvod

Religija je sastavni dio života svakog čovjeka, a pri tome je uopće ne bitno je li čovjek religiozan ili nije.

Cilj rada je analizirati tekstove religijskih rasprava uz pomoć statističke analize.

Statistička analiza jezika

Deskriptivna statistička analiza

Deskriptivna statistička analiza bavi se mjerama centralne tendencije i mjerama rasipanja.

Najvažnije mjere centralne tendencije su:

- aritmetička sredina
- medijan
- mod

Odnos mjera centralne tendencije

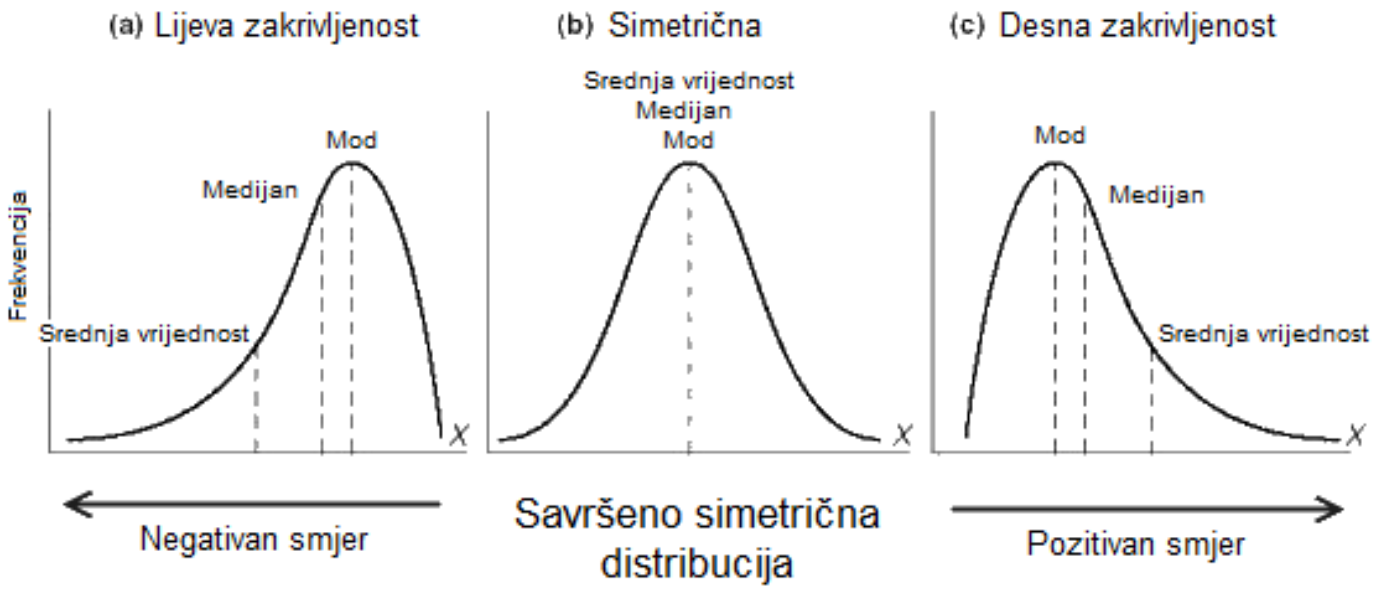

Figure: Odnos aritmetičke sredine, medijana i mode ovisno o zakrivljenosti distribucije.

Mjere rasipanja

Najvažnije mjere rasipanja su:

- rang
- interkvartilni rang
- varijanca
- standardna devijacija
- koeficijent varijacije

Vizualizacija standardne devijacije

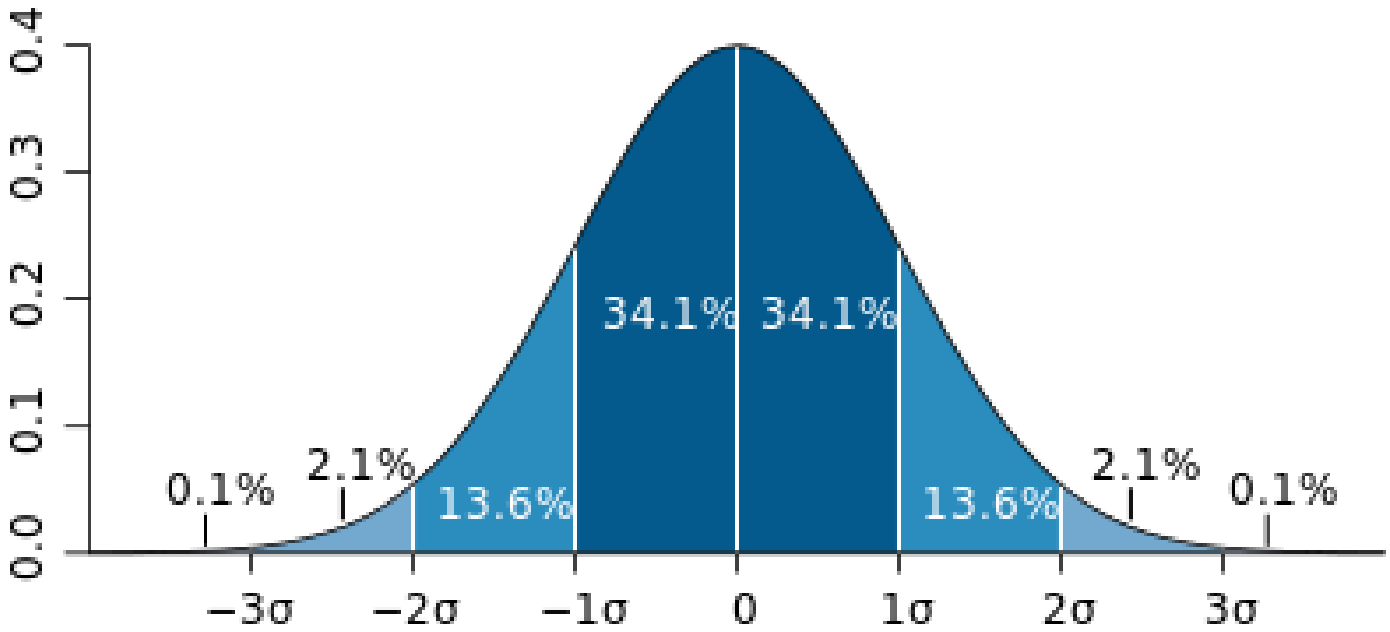

Figure: Standardna devijacija u normalnoj distribuciji.

Inferencijalna statistička analiza

Inferencijalna analiza odnosi se na provjeravanje postavljenih hipoteza uz pomoć statističkih testova.

Inferencijalna statistička analiza se provodi na sljedeći način:

- Postavljanje hipoteza
- Odabiranje testne statistike
- Odabiranje nivoa značajnosti α
- Izračunavanje vrijednosti statistike i usporedba s α

Eksplorativna statistička analiza

Eksplorativna statistička analiza za razliku od deskriptivne i inferencijalne analize ne fokusira se na osnovne podatke o podacima i testiranje hipoteza.

Cilj eksplorativne statistička analiza je istražiti podatke bez definiranih ciljeva u početku te pokušati formulirati sasvim nove hipoteze o podatcima.

Proces obrade podataka

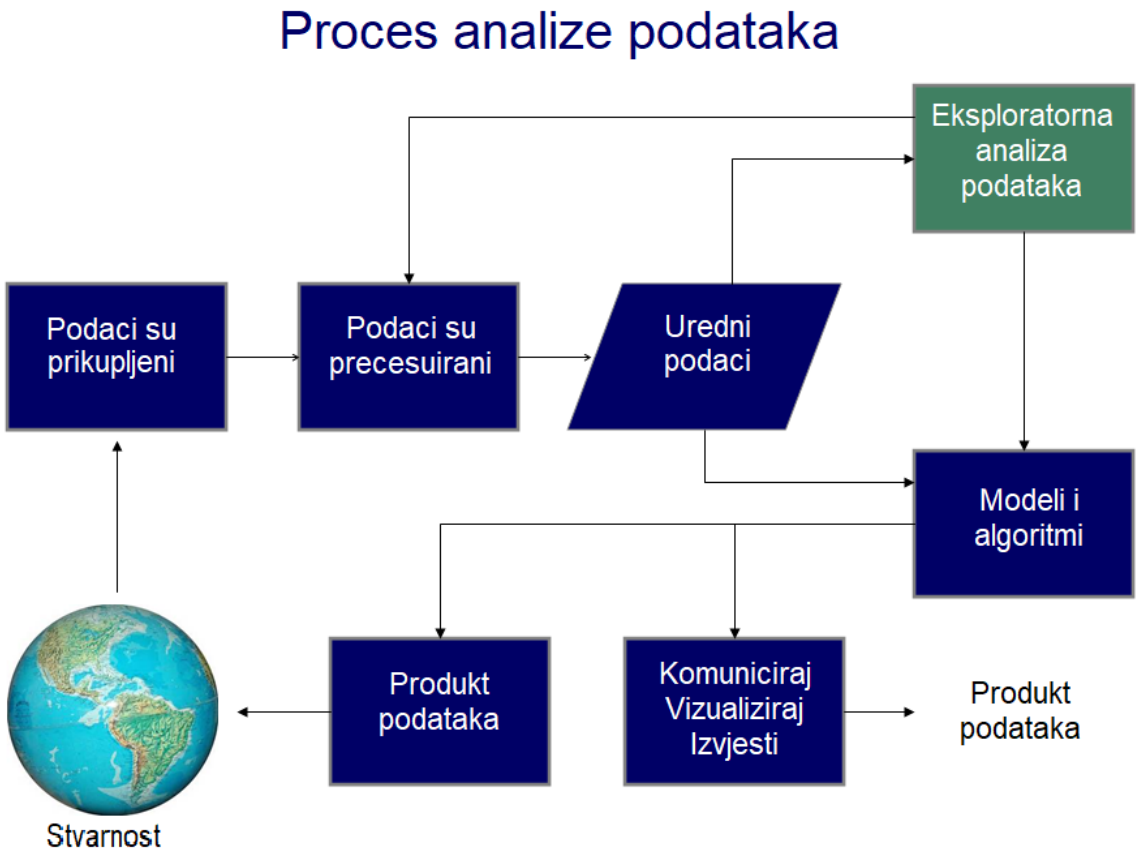

Figure: Proces obrade podataka u eksplorativnoj statističkoj analizi.

Značajke LIWC

LIWC je kratica od Lingvističko ispitivanje i brojanje riječi (engl. *Linguistic Inquiry and Word Count*).

LIWC uzima dani tekst i za svaku značajku LIWC vraća frekvenciju njenog pojavljivanja u tekstu.

Značajke LIWC podijeljene su u četiri kategorije:

- lingvističke varijable
- interpunkcijske varijable
- ostale gramatičke varijable
- psihološke varijable

Strojno učenje

Strojno učenje je podgrana umjetne inteligencije koja se razvila iz raspoznavanja uzoraka i statistike.

Strojno učenje nije ništa drugo nego predviđanje svojstva novih podataka na temelju svojstva starih podataka.

Algoritam strojnog učenja pomoću starih podataka stvara model koji je definiran parametrima čije se vrijednosti određuju iz starih podataka, tj. podataka koje smo ustupili algoritmu.

Logistička regresija

Logistička regresija je metoda za klasifikaciju podataka u diskretne klase. Logistička funkcija definirana je prema sljedećoj formuli:

$$f(z) = \frac{1}{1 + e^{-z}} \tag{1}$$

gdje se varijabla z računa na sljedeći način:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n = \beta + \sum_{n=1}^{N} (B_i x_i) \tag{2}$$

Graf logističke regresije

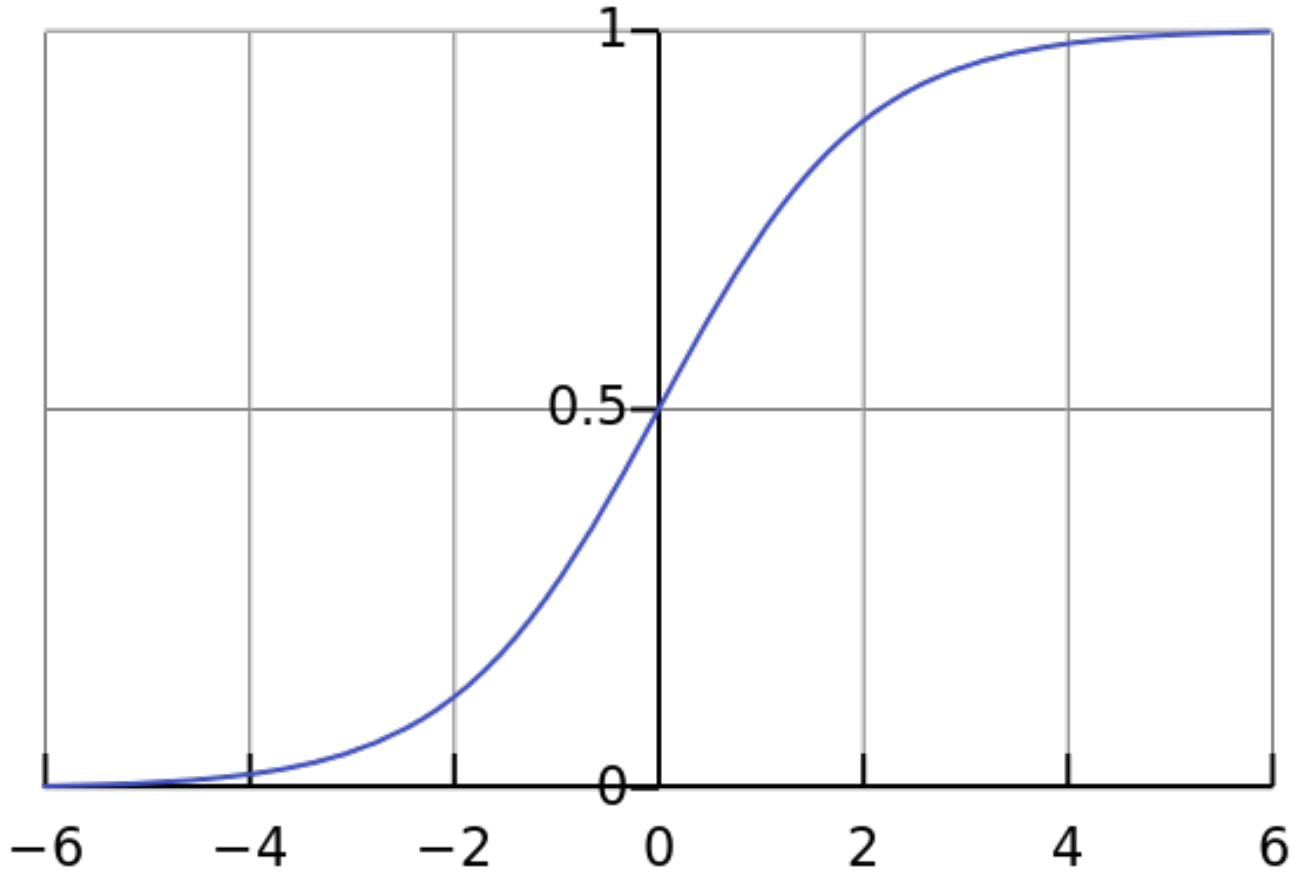

Figure: Graf logističke funkcije.

Stroj potpornih vektora

Stroj potpornih vektora jedan je od najpopularnijih modela strojnih učenja koji se danas koristi pri bilo kojoj klasifikaciji.

Stroj potpornih vektora na ulaz dobiva podatke povezane s klasom kojoj pripadaju te ih prikazuje kao točke u prostoru raspoređene na način da su točke koje predstavljaju podatke koji pripadaju različitim klasama međusobno što razmaknutije.

Zadatak stroja potpornog vektora je odabrati optimalnu hiperravninu razdvajanja.

Odabir hiperravnine u stroju potpornih vektora.

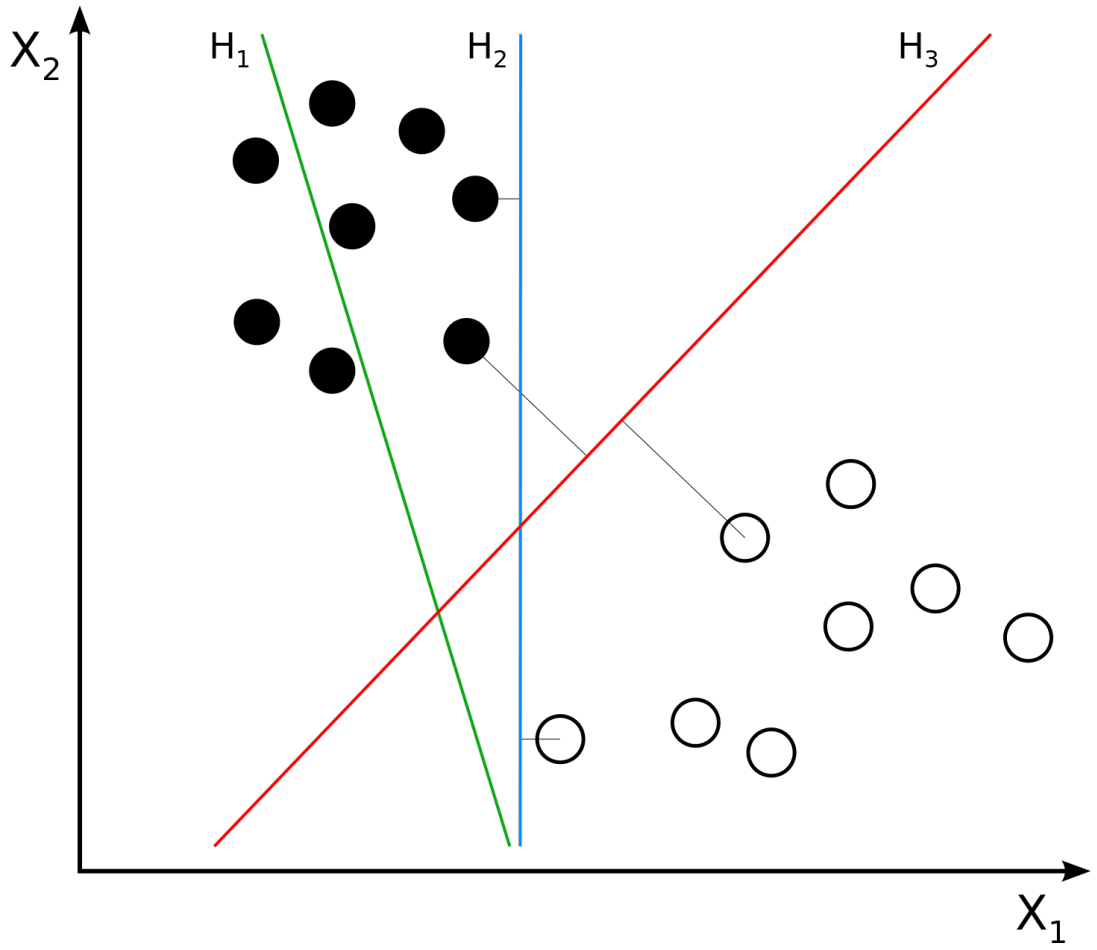

Figure: Odabir hiperravnine u stroju potpornih vektora.

Podaci

Prikupljanje podataka

Podatke za ovaj rad prikupili smo s društvene mreže Reddit, s dretve naziva “debateReligion".

Odabrali smo 4000 korisnika koji su raspravljali na spomenutoj dretvi te za svakog od njih su preuzeti svi komentari koje su ti korisnici ostavljali na društvenoj mreži Reddit, bez obzira o kojoj se dretvi radilo.

Obrada podataka

Obradu smo napravili u sljedećim koracima:

- Odbacivanje svih podatka koji nam nisu bitni za analizu
- Spajanje svih komentara istog korisnika u jedan komentar
- Standardizacija oznaka

Model podataka

Nakon obrade podataka u našim podatcima imamo 3733 redaka, svaki redak sadrži oznaku religije i tekst komentara.

U model podataka dodat ćemo frekvencije 700 najčešćih riječi koje se pojavljuju u svim komentarima nakon izbacivanja zaustavnih (engl. *stop*) riječi iz teksta.

U model podatka dodat ćemo i značajke LIWC koje smo dobili uz pomoć programa LIWC2015.

Broj korisnika po religijama

Table: Broj korisnika po religijama

Religija	Broj korisnika	Postotak
Sekularni	2228	59,68%
Kršćani	567	15,18%
Ostali	506	13,55%
Bivši	208	5,57%
Muslimani	106	2,83%
Židovi	63	1,69%
Budisti	40	1,07%
Hindusi	15	0,40%

Rezultati

Razlike u deskriptivnoj analizi

Table: Broj komentara po korisniku, broj riječi po komentaru i dužina riječi po religijama

Religija	Broj komentara	Broj riječi po komentaru	Dužina riječi
Bivši	685,80	12,06	5,39
Budisti	388,80	11,94	5,35
Hindusi	617,40	11,82	5,38
Kršćani	498,69	12,84	5,38
Muslimani	536,65	12,99	5,37
Ostali	694,77	12,83	5,50
Sekularni	823,48	12,75	5,38
Židovi	629,33	13,00	5,41

Razlike u vokabularu

Table: Deset najčešće korištenih riječi po religijama

Religija	Deset najčešće korištenih riječi
Bivši	*people*, *good*, *will*, *time*, *well*, *shit*, *pretty*, *thing*, *going*, *fuck*
Budisti	*people*, *will*, *good*, *time*, *well*, *sure*, *pretty*, *going*, *better*, *thing*
Hindusi	*people*, *will*, *good*, *india*, *time*, *well*, *pretty*, *better*, *guy*, *indian*
Kršćani	*people*, *good*, *will*, *god*, *time*, *well*, *going*, *pretty*, *sure*, *thing*
Muslimani	*people*, *will*, *good*, *islam*, *time*, *muslim*, *muslims*, *god*, *well*, *allah*
Ostali	*people*, *good*, *will*, *time*, *well*, *going*, *thing*, *better*, *pretty*, *watch*
Sekularni	*people*, *good*, *will*, *time*, *well*, *going*, *thing*, *pretty*, *sure*, *better*
Židovi	*people*, *jews*, *will*, *good*, *jewish*, *fine*, *god*, *israel*, *thing*, *going*

Razlike u značajkama LIWC

Table: Tablica značajki LIWC

Religija	Religioznost	Uskličnici	Točka zarez	Obitelj	Hrana	Novac	Psovanje
Bivši	1,60	0,37	0,12	0,29	0,35	0,59	0,40
Budisti	1,08	0,56	0,12	0,18	0,39	0,65	0,29
Hindusi	2,13	0,25	0,09	0,21	0,23	0,42	0,32
Kršćani	2,25	0,36	0,13	0,29	0,28	0,60	0,24
Muslimani	2,97	0,29	0,13	0,33	0,22	0,39	0,27
Ostali	1,23	0,34	0,20	0,22	0,35	0,60	0,35
Sekularni	1,06	0,32	0,12	0,21	0,31	0,67	0,36
Židovi	1,80	0,49	0,11	0,31	0,33	0,60	0,28

Zaključci inferencijalne analize

Table: Rezultati testne statistike

Značajka	Vrijednost testne statistike
Frekvencija “muslim"	2*10^-16
Frekvencija “jews"	2*10^-16
Frekvencija “shit"	1,5*10^-8
LIWC “Religioznost"	2*10^-16
LIWC “Uskličnici"	0,000587
LIWC “Točka zarez"	0,00195
LIWC “Obitelj"	2*10^-16
LIWC “Hrana"	1,02*10^-6
LIWC “Novac"	5,21*10^-12
LIWC “Psovanje"	3,22*10^-16

Predikcije logističke regresije

Table: Rezultati na skupu za treniranje

Pogodak	Promašaj	Točnost	95%-tni interval pouzdanosti
2012	988	67,07%	67,04%-67,10%

Table: Rezultati na skupu za testiranje

Pogodak	Promašaj	Točnost	95%-tni interval pouzdanosti
457	276	62,36%	62,30%-62,43%

Predikcije stroja potpornih vektora

Table: Rezultati na skupu za treniranje

Pogodak	Promašaj	Točnost	95%-tni interval pouzdanosti
2077	923	69,23%	69,20%-69,27%

Table: Rezultati na skupu za testiranje

Pogodak	Promašaj	Točnost	95%-tni interval pouzdanosti
473	260	64,53%	64,46%-64,59%

Zaključak

Glavna motivacija ovog rada je bila vidjeti postoje li razlike u internetskim raspravama između religija.

Uspjeli smo statistički uvidjeti da postoje razlike u načinu izražavanja na internetu između religija.

U ovom radu uspješno smo izgradili model koji s točnošću od otprilike 65% predviđa koje je religijske orijentacije korisnik koji je napisao komentare.

Rezultati dobiveni u ovom radu, otvaraju daljnje mogućnosti istraživanja međureligijskih odnosa i razlika.

Osim statistički, ovaj problem bi se sigurno trebao sagledati s teološke i psihološke perspektive, pogotovo nakon što nam je statistika pokazala da razlike postoje.

Hvala na pažnji! Pitanja?