Prepoznavanje neželjenih SMS poruka

Tubić, Eugen

SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA

Sveučilišni studij

PREPOZNAVANJE NEŽELJENIH SMS PORUKA

Završni rad

Eugen Tubić

Osijek, 2018. g.

FERIT
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA OSIJEK

Obrazac Z1P - Obrazac za ocjenu završnog rada na preddiplomskom sveučilišnom studiju

Osijek, 19.09.2018.

Odboru za završne i diplomske ispite

Prijedlog ocjene završnog rada

Ime i prezime studenta:	Eugen Tubić
Studij, smjer:	Preddiplomski sveučilišni studij Računarstvo
Mat. br. studenta, godina upisa:	R3511, 24.10.2017.
OIB studenta:	88599023212
Mentor:	Prof.dr.sc. Goran Martinović
Sumentor:	Dr.sc. Bruno Zorić
Sumentor iz tvrtke:	
Naslov završnog rada:	Prepoznavanje neželjenih SMS poruka
Znanstvena grana rada:	**Programsko inženjerstvo (zn. polje računarstvo)**
Predložena ocjena završnog rada:	Izvrstan (5)
Kratko obrazloženje ocjene prema Kriterijima za ocjenjivanje završnih i diplomskih radova:	Primjena znanja stečenih na fakultetu: 3 bod/boda Postignuti rezultati u odnosu na složenost zadatka: 2 bod/boda Jasnoća pismenog izražavanja: 3 bod/boda Razina samostalnosti: 3 razina
Datum prijedloga ocjene mentora:	19.09.2018.
Datum potvrde ocjene Odbora:	26.09.2018.

Potpis mentora za predaju konačne verzije rada u Studentsku službu pri završetku studija:	Potpis:
	Datum:

FERIT
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA
I INFORMACIJSKIH TEHNOLOGIJA **OSIJEK**

IZJAVA O ORIGINALNOSTI RADA

Osijek, 26.09.2018.

Ime i prezime studenta:	Eugen Tubić
Studij:	Preddiplomski sveučilišni studij Računarstvo
Mat. br. studenta, godina upisa:	R3511, 24.10.2017.
Ephorus podudaranje [%]:	8

Ovom izjavom izjavljujem da je rad pod nazivom**: Prepoznavanje neželjenih SMS poruka**

izrađen pod vodstvom mentora Prof.dr.sc. Goran Martinović

i sumentora Dr.sc. Bruno Zorić

moj vlastiti rad i prema mom najboljem znanju ne sadrži prethodno objavljene ili neobjavljene pisane materijale drugih osoba, osim onih koji su izričito priznati navođenjem literature i drugih izvora informacija.
Izjavljujem da je intelektualni sadržaj navedenog rada proizvod mog vlastitog rada, osim u onom dijelu za koji mi je bila potrebna pomoć mentora, sumentora i drugih osoba, a što je izričito navedeno u radu.

Potpis studenta:

Sadržaj:

1. UVOD

U ovom radu bit će opisan problem neželjenih elektroničkih poruka i njihov utjecaj na svakodnevno korištenje digitalnih komunikacijskih sustava. Elektroničke neželjene poruke mogu se opisati kao bilo kakva bezvrijedna elektronička poruka koju primatelj smatra neželjenom. Elektroničke poruke koje se smatraju neželjenima najčešće sadrže reklame, poveznice do proizvoda za koje primatelj nema interesa, ponude za razne usluge i slično. Neželjene poruke često se smatraju opasnima jer mogu biti „zaražene“ računalnim virusom ili mogu sadržavati razne oblike prijevare poput mrežne krađe identiteta (engl. *phishing*). Neželjene poruke se koriste u mnogim tvrtkama kao sredstvo reklamiranja. Isplative su s ekonomske strane jer oglašavačke tvrtke nemaju nikakvih radnih troškova osim ažuriranja liste primatelja te održavanja servera i domena. Pobliže će biti objašnjen utjecaj neželjenih SMS (engl. *short message service*) poruka te će biti prikazano jedno od mnogih programskih rješenja za filtriranje neželjenih poruka. Neželjene SMS poruke razlikuju se od ostalih oblika neželjenih poruka jer je broj znakova koji se može koristiti u poruci ograničen. Također, neželjene SMS poruke specifične su po tome što korisnicima može biti naplaćena naknada za slanje poruke iako su oni samo primatelji te poruke.

Nakon uvoda, u drugom će poglavlju rada biti prikazani osnovni problemi neželjenih poruka u određenim komunikacijskim kanalima te utjecaj neželjenih poruka na slanje poruka u tim komunikacijskim kanalima. Objasnit će se različiti načini klasifikacije i filtriranja neželjenih poruka koji već postoje u literaturi. U trećem poglavlju biti će prikazan program koji će pokušati klasificirati poruke pomoću pet različitih klasifikatora te će biti provedena analiza uspješnosti klasifikacije. U zaključku će biti komentirana uspješnost klasifikatora u prepoznavanju neželjenih SMS poruka.

1.1. Zadatak završnog rada

U teorijskom dijelu rada potrebno je opisati problem neželjenih elektroničkih poruka i njegov utjecaj na svakodnevno korištenje digitalnih komunikacijskih sustava poput e-pošte, SMS-a, IM-a i slično. Prikazati najčešće metode rješavanja ovog problema korištenjem nadzirane klasifikacije i značajki dobivenih na temelju sadržaja poruke. U praktičnom dijelu rada ostvariti sustav za prepoznavanje neželjenih poruka i testirati ga na podatkovnom skupu neželjenih SMS poruka iz literature.

2. OPISNE ZNAČAJKE I METODE ZA RASPOZNAVANJE NEŽELJENIH PORUKA

Neželjene poruke su postale veliki problem u digitalnim komunikacijskim sustavima. Digitalni komunikacijski sustavi, poput elektroničke pošte, SMS poruka ili instant poruka, postali su dio svakodnevne privatne, ali i poslovne komunikacije. Neželjene poruke najviše štete poslovnoj komunikaciji jer ju usporavaju. Poseban problem se stvara zbog korištenja elektroničke pošte koja je iznimno podložna primanju neželjenih poruka od običnih reklama do pokušaja mrežne krađe identiteta. Prema [1], zbog neželjenih poruka najviše ispaštaju sami korisnici, ljudi koji primaju neželjene poruke. Elektronička pošta korisnika postaje zatrpana stotinama neželjenih poruka, njihova računala moraju spremati te poruke na tvrdi disk ako koriste neki program za elektroničku poštu, a i internet promet (engl. *bandwidth*) korisnika se koristi za primanje tih poruka. Ljudi koji šalju takve poruke nemaju gotovo nikakvih troškova osim svoga vremena koje im je potrebno za održavanje liste primatelja te slanje neželjenih poruka. Kod poslovnih korisnika, najveći problem stvaraju neželjene poruke oglašavačkog karaktera pri čemu njihova elektronička pošta postaje zatrpana raznim reklamnim porukama koje umanjuju preglednost njihovog računa i brzinu komunikacije.

Prema [2], svrha filtriranja neželjenih poruka je umanjivanje utjecaja neželjenih poruka na digitalni komunikacijski sustav, odnosno poboljšavanje dostavljanja bitnih podataka korisniku. Za filtriranje neželjenih poruka, često se koriste razne metode koje uključuju korištenje klasifikatora. Klasifikatori se koriste za određivanje neželjenih poruka, odnosno koriste se da korisniku prikažu koja je poruka neželjena, a koja je poruka prava. Postoji više različitih klasifikatora, a za različite slučajeve se koriste različiti klasifikatori. Neki klasifikatori su bolji za filtriranje elektroničke pošte, dok su neki su bolji za filtriranje SMS poruka. U ovom radu će se prikazati utjecaj neželjene pošte na razne digitalne komunikacijske kanale, kao i neke metode za filtriranje neželjene pošte.

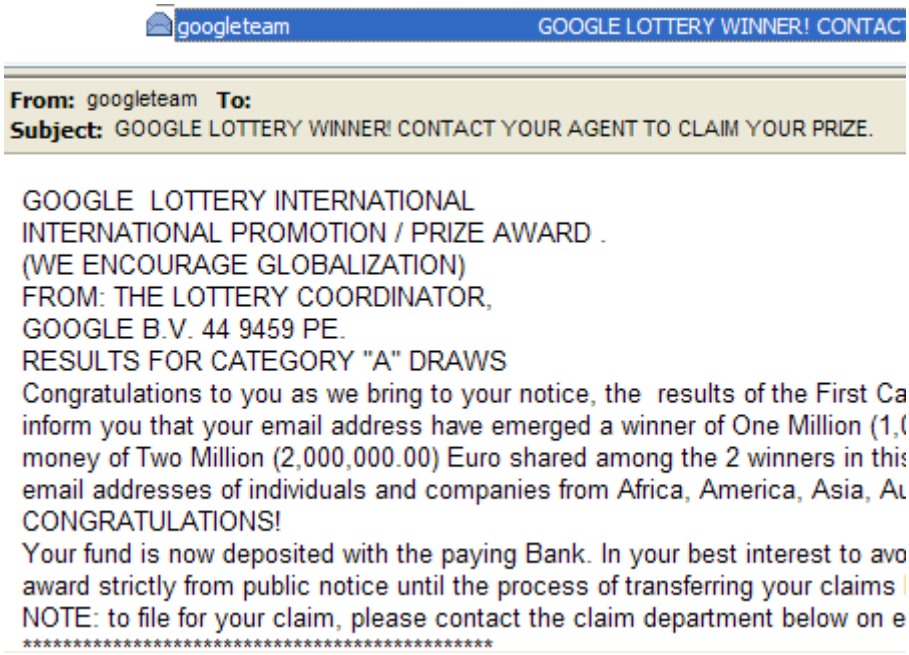
2.1. Neželjene poruke

Prema [3], elektronička poruka je neželjena ako je osobni identitet primatelja te kontekst poruke nebitan jer se poruka može primjeniti na razne druge potencijalne primatelje i ako primatelj nije dao provjereno dopuštenje da mu se takva poruka šalje. Postoje razni oblici neželjenih poruka. Neželjene poruke mogu biti elektroničke poruke koje obično budu napisane od većeg broja znakova i to su najčešće reklame, no u zadnjih nekoliko godina počele su se pojavljivati kratke poruke koje služe za slanje preko servisa za razgovor (engl. *chat*) ili SMS poruka koje najčešće

budu pozivnice za neke nagradne igre ili poruke koje govore korisniku da je osvojio neku nagradu. Takve poruke su u većini slučajeva obična prijevara i treba se odgovorno nositi s takvim porukama. Poruke tog oblika su opasne jer mogu od korisnika saznati važne podatke poput broja i PIN-a kreditne kartice. Neželjene poruke se šalju iz raznih izvora, od običnih ljudi koji jednostavno iz dosade žele poslati tisuće neželjenih poruka nekome na elektroničku poštu ili organiziranog slanja neželjene pošte, što se najčešće odnosi na razne tvrtke koje se bave marketingom. Naravno, uvijek postoji i druga strana, koja jednim dijelom graniči i s kriminalnim aktivnostima, a to je slanje malicioznog sadržaja kroz neželjenu poštu poput trojanskih konja, crvi ili računalnih virusa.

2.1.1. Neželjena elektronička pošta

Neželjenom elektroničkom poštom smatra se bilo kakva poruka koja je s gotovo identičnim sadržajem poslana velikom broju primatelja na njihovu e-poštu. Elektroničke poruke tog tipa često sadrže korisniku poznate sadržaje koji služe da korisnika navuku na otvaranje internetske poveznice ili nečeg sličnog. Takve poveznice najčešće vode do daljnjih reklama ili služe za *phishing* informacija od korisnika, poput broja i PIN-a njegove kreditne kartice i slično. Preko elektroničke pošte se isto šalju velike količine reklamnog sadržaja kojeg najčešće šalju razne oglašivačke tvrtke. Takva vrsta elektroničkih poruka je uglavnom bezopasna, a jedina mana joj je nepotrebno zatrpavanje elektroničkog poštanskog sandučića korisnika. Većina elektroničkih neželjenih poruka se šalje preko elektroničke pošte jer ne postoji ograničenje na broj znakova poruke. Prema [4], može se primijetiti da čak 45% svih elektroničkih poruka poslanih preko elektroničke pošte u jednom danu predstavljaju neželjene poruke. Od toga broja, 36% neželjenih poruka su reklame. Iz ovih brojeva se može zaključiti da su neželjene poruke veliki problem u digitalnim komunikacijskim kanalima. Također, u provedenom istraživanju prema [4], se može vidjeti da neželjene poruke stvaraju velike troškove tvrtkama te se smatra da neželjene poruke uzrokuju oko 20 milijardi dolara gubitaka godišnje. Primjer neželjene poruke u elektroničkoj pošti se može vidjeti na slici 2.1.



Slika 2.1. *Neželjena elektronička pošta [5]*

2.1.2. Neželjene SMS poruke

Neželjene SMS poruke su oblik neželjenih tekstualnih poruka koje se šalju preko mobilnih uređaja. Početkom 21. stoljeća počinje i ubrzani razvoj mobilnih telefona pri čemu se pojavljuje mogućnost slanja neželjenih poruka preko mobilnih uređaja. U usporedbi s neželjenim porukama u elektroničkoj pošti, zastupljenost neželjenih poruka na mobilnim uređajima je relativno mala, osim u određenim azijskim zemljama poput Kine gdje prema [6], zastupljenost neželjenih SMS poruka je 20 do 30%. Neželjene SMS poruke se manje koriste od drugih oblika neželjenih poruka jer je broj znakova SMS poruka ograničen pa je zato moguće slati samo kratke poruke, većinom pozivnice za nagradne igre koje su često oblik *phishinga*. Prema [7], slanje neželjenih SMS poruka je u Europi u porastu jer prema statistici gotovo svaka osoba preko 15 godina starosti posjeduje mobilni uređaj te prosječni korisnik mobilnog uređaja dnevno pošalje 10 SMS poruka, što SMS poruke stavlja u odličnu poziciju za zlouporabu. Neželjene SMS poruke su lakše za prepoznati nego poruke elektroničke pošte zbog ograničenosti znakova. Često sadrže svih 160 mogućih znakova jer je pošiljatelju neželjenih poruka svejedno šalje li se 16 ili 160 znakova kada uvijek plaća istu cijenu poruke. Također, prema [6] postoji još jedna važna značajka prepoznavanja neželjenih poruka, a to su pravopis i gramatika. U neželjenim porukama je općenito manje gramatičkih pogrešaka nego u običnima, a to se može primjetiti u određenim dijelovima poruka poput stavljanja razmaka nakon zareza. Razlozi slanja neželjenih SMS poruka su drugačiji nego kod slanja neželjene elektroničke pošte jer je SMS puno manje podložan oglašavanju zbog ograničenosti znakova jedne poruke. Neželjene SMS poruke se koriste uglavnom za prijevaru

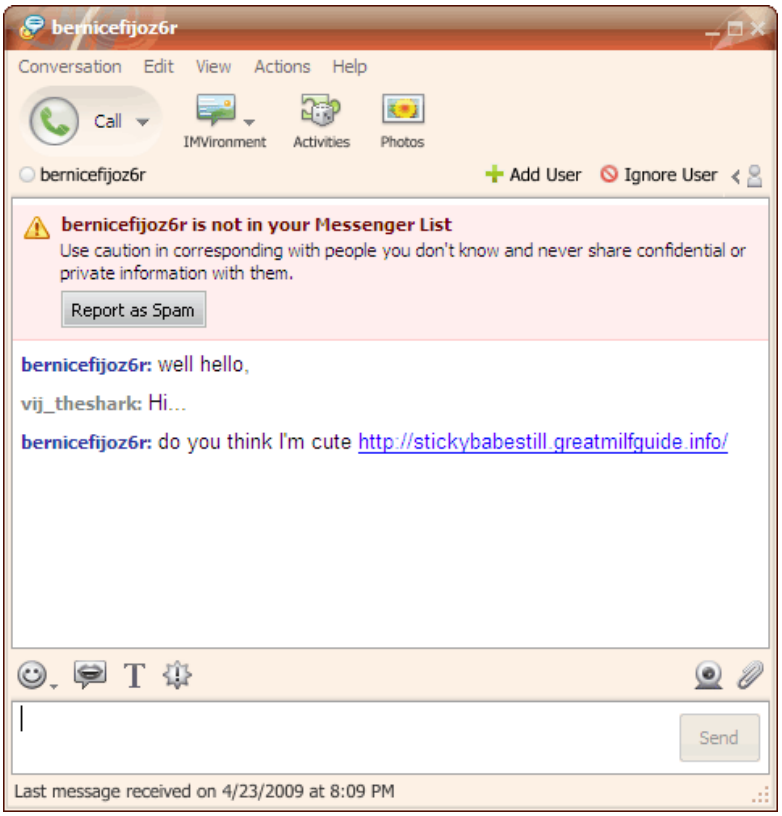
poput poruka da je korisnik nešto osvojio ili poziva za nagradne igre. Primjer neželjene SMS poruke se može vidjeti na slici 2.2.

Slika 2.2. Neželjena SMS poruka *[8]*

2.1.3. Ostali komunikacijski kanali

Neželjene poruke nisu samo problem elektroničke pošte i SMS poruka. Postoje razni oblici neželjenih poruka i na drugim servisima. U daljnem tekstu, prikazat će se još nekoliko komunikacijskih kanala koji imaju problem s neželjenim porukama.

Instant poruke su jedan od digitalnih komunikacijskih kanala koje bilježe porast u slanju neželjenih poruka. Prema [9], teže je ignorirati neželjene instant poruke jer one „iskoče“ čim se korisnik prijavi na svoj račun. Također, kod instant poruka se pojavljuje veća vjerojatnost da će korisnik stvarno i otvoriti sadržaj poruke jer će poruka u realnom vremenu biti dostavljena korisniku te se može dogoditi da korisnik slučajno stisne na poruku. Primjer neželjene poruke u instant porukama je prikazan na slici 2.3.



Slika 2.3. *Neželjena poruka u instant porukama [10]*

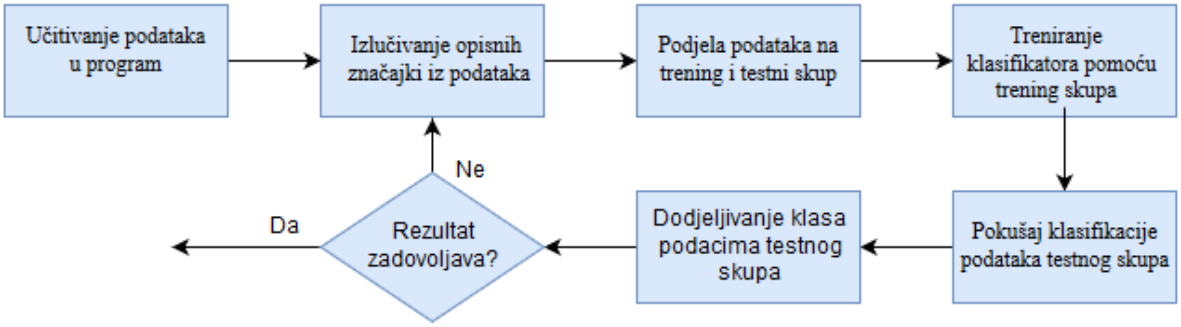
Neželjene poruke na društvenim mrežama također postaju problem u zadnjih nekoliko godina. Prema [11], neželjene poruke na društvenim mrežama imaju puno veću vjerojatnost da će ih netko otvoriti ili pročitati jer prema istraživanjima, korisnici će prije otvoriti poruku kada znaju da ih je poslao njihov prijatelj ili kolega s posla. Zbog toga, filtriranje neželjenih poruka na društvenim mrežama je puno teže, a s druge strane, pošiljateljima takvih poruka je puno lakše dobiti klikove na njihove neželjene poruke ili internetske poveznice. Primjer neželjene poruke na društvenoj mreži je prikazan na slici 2.4.

Slika 2.4. Neželjena poruka na društvenoj mreži Twitter [12]

2.2. Metode raspoznavanja zasnovane na strojnom učenju

Prije nego se u radu prikažu konkretni klasifikatori, potrebno je prvo objasniti razliku između nadzirane i nenadzirane klasifikacije. U ovome radu koristiti će se klasifikatori koji pripadaju nadziranoj klasifikaciji.

Prema [13], nadzirana klasifikacija je klasifikacija u kojoj postoji poznati skup podataka, odnosno trening skup na kojem se klasifikator istrenira kako razlikovati različite klase koje postoje u tom trening skupu kako bi se kasnije mogle odrediti klase za nove podatke koji će se unositi. Klasifikator će izlučiti zadane opisne značajke podataka iz trening skupa i znati će njihovu pravu klasu. Nakon toga, prema poznatim klasama trening skupa klasifikator će pokušati dodjeliti klase novim podacima. Kod nenadzirane klasifikacije ne postoji trening skup nego će klasifikator bez ikakvog prijašnjeg saznanja o skupu podataka, pokušati nove podatke razdvojiti u grupe sličnih karakteristika. U [13], ova problematika je objašnjenja na primjeru košare s voćem gdje će kod nadzirane klasifikacije klasifikator imati prijašnja saznanja o boji ili veličini voća, dok će kod nenadzirane klasifikacije sustav morati pregledati svako voće, sortirati ga prvo po boji pa onda po veličini i nakon toga im pokušati odrediti prave grupe. Slika 2.6. prikazuje blok dijagram nadzirane klasifikacije.



Slika 2.6. *Blok dijagram nadzirane klasifikacije*

Kod nadzirane klasifikacije se pojavljuje i problematika neujednačenih klasa. Prema [14], problem neujednačenih klasa se pojavljuje kada u skupu podataka ne postoji jednak broj podataka za sve klase. Pri djeljenju podataka na trening i test skup, potrebno je voditi računa o omjerima podjele podataka, odnosno u svakom skupu treba postojati ujednačen omjer podataka s različitim klasama koje čine skup podataka.

Postoje razne metode za filtriranje neželjenih poruka, ali ne koriste se sve metode za filtriranje neželjenih SMS poruka. Određene metode raspoznavanja su bolje za određivanje neželjenih poruka u instant porukama ili elektroničkoj pošti, dok su neke metode bolje za određivanje neželjenih poruka u SMS porukama. SMS poruke su specifične zbog ograničenosti znakova. Teže je filtrirati poruke kada je broj znakova ograničen jer poruke postaju previše slične. Također, u SMS porukama se često koriste posebni znakovi kako bi se određene riječi skratile ili kako bi se ispisali smajlova. Za pronalaženje neželjenih poruka koriste se klasifikatori. Postoji veliki broj različitih klasifikatora, no u tekstu će biti objašnjeni klasifikatori koji se najviše koriste u praksi.

Naivni Bayesov klasifikator, prema [15], se temelji na računanju posteriorne vjerojatnosti. Kada se koristi veliki broj značajki, klasifikator postaje spor i težak za korištenje. Da bi se ovaj klasifikator mogao koristiti, potrebno je poznavati vjerojatnost da neki podatak pripada određenoj klasi. Tu vjerojatnost se može izračunati iz trening skupa podataka, pri ćemu se pretpostavlja da su sve značajke međusobno neovisne. Naivni Bayesov klasifikator se često koristi u praksi, ali prema [16], obično s nekoliko optimizacija, ovisno o kakvom se filtriranju radi kako bi se najlakše mogao prilagoditi stvarnim situacijama koje se mogu dogoditi. Naivni Bayesov klasifikator se mora prvo istrenirati s kontroliranim porukama koje su unaprijed određene kao neželjene poruke ili kao prave poruke. Kroz klasifikator se propusti skup poruka za treniranje pri čemu se riječima koji se često pojavljuju pridodaje neutralna vjerojatnost 0,5 pri čemu klasifikator najčešće takve riječi jednostavno ignorira. Riječi kojima se pridodaje neutralna vrijednost su najčešće riječi poput

članova u engleskom jeziku, kao *the* ili nešto što se piše često poput riječi *it*. Nakon što se klasfikator istrenira na trening skupu, klasifikator se može koristiti za raspoznavanje neželjenih poruka. Rezultat naivnog Bayesovog klasifikatora su vrijednosti u rasponu od 0 do 1, odnosno vjerojatnost da je ta poruka neželjena. Kako je u tekstu već navedeno, naivni Bayesov klasifikator se najčešće implementira s određenim dodacima te se tako dobiju multinominalni i multivarijabilni Bayesovi klasifikatori koji se često koriste u praksi. Uspješnost Bayesovih klasifikatora, prema [16], je visoka. Prema [16], multinominalni Bayesov klasifikator ima uspješnost između 93 i 97% u raspoznavanju neželjenih poruka, dok multivarijabilni Bayesov klasifikator je uspješno prepoznao neželjene poruke u 89% slučajeva.

Metoda vektora podrške ili SVM (engl. *support vector machine*), prema [15], je binarni klasifikator koji se koristi pri nadziranom učenju, primjeniv je u klasifikaciji i problemu regresije. Promatrano u prostoru, SVM klasifikator će pokušati odvojiti značajke suprotnih klasa hiper-ravninom pri čemu će se pokušati stvoriti najveći razmak između najbližih elemenata i hiper-ravnine. Margine su hiper-ravnine koje odvajaju podatke i između sebe imaju najveću udaljenost. Pri korištenju SVM klasifikatora kod linearno neodvojivih podataka, treba se dozvoliti određena razina pogreške prilikom stvaranja hiper-ravnine. U ovom slučaju margine će se pokušati stvoriti sa što većim razmakom, ali pritom će broj pogrešno klasificiranih elemenata pokušati održati što manjim.

Šume odluke (engl. *random forest*), prema [15], su klasifikatori koji se sastoje od više stabala odluke. Kod ovog klasifikatora često dolazi do preprilagođavanja (engl. *overfitting*) pri ćemu se klasifikator previše prilagodi trening skupu, dok je na stvarnim podacima preciznost manja. Svako stablo kod ovog klasifikatora se stvara s podskupom trening podataka, a ta se metoda zove odvajanje (engl. *bagging*). Sami postupak klasifikacije se izvodi tako da svako stablo odredi klasu, a onda se za klasu samog podatka uzima klasa koja ima najviše glasova.

Umjetna neuronska mreža (engl. *artificial neural network*) je skup umjetnih neurona. Ova metoda se temelji na principima rada ljudskog mozga. Neuronska mreža se sastoji od umjetnih neurona pri čemu svaki neuron je funkcija koja na izlazu daje logičku jedinicu ako suma ulaza prelazi zadani prag, dok u slučaju ako suma ulaza ne prelazi zadani prag daje logičku nulu. Višeslojni perceptron (engl. *multiLayer perceptron*) je vrsta neuronske mreže koja je sposobna raditi sa nelinearnim podacima. Višeslojni perceptron sadrži jedan ili dva skrivena sloja koji mu omogućavaju rad s nelinearnim podacima [15].

Jedan od najčešće korištenih klasifikatora je kNN (engl. *k nearest neighbour*) klasifikator. Implementacija ovoga klasifikatora je jednostavna jer način na koji kNN klasifikator raspoznaje neželjene poruke je lako razumljiv. Prema [17], kNN klasifikator je najpopularniji klasifikator za raspoznavanje neželjenih poruka jer su mnogi istraživači otkrili da kNN klasifikator ostvaruje odlične rezultate za razne skupine podataka. Ideja kNN klasifikatora je lako razumljiva. Klasifikator se prvo trenira na skupini podataka za treniranje, tj. prikupi opisne značajke svih poruka iz trening skupa pri čemu će klasifikator znati koja poruka je prava, a koja neželjena iz trening skupa. Kada se klasifikatoru da novi podatak, klasifikator će računati udaljenost između novog podatka i svakog podatka iz skupine podataka za trening. Udaljenost će računati na temelju određenih značajki poput broja riječi, informatičkoj pismenosti, udjelu brojeva u poruci i slično. Nakon što se izračuna udaljenost između novog podatka i svih podataka iz trening skupa, klasifikator razmatra *k* podataka iz trening skupa koji su najbliži novom podatku. Ukoliko je k=3, razmtrati će se 3 najbliža podatka iz trening skupa. Ovisno o 3 najbliža podatka, klasifikator će odlučiti je li novi podatak neželjena ili prava poruka. Klasifikacija se određuje na temelju većine, odnosno ako su barem 2 podatka prava poruka, i novi podatak će biti klasificiran kao prava poruka i obrnuto. Glavni problem kNN klasifikatora je njegova sporost jer klasifikator će računati udaljenost između novog podatka i svakog drugog podatka iz trening skupa. Udaljenost između podataka se računa pomoću euklidske udaljenosti, no ne i nužno samo pomoću nje. To će raditi tako što će sustav računati euklidsku udaljenost između opisnih značajki nove učitane poruke i poruka iz trening skupa, pri ćemu je euklidska udaljenost ustvari udaljenost između dvije točke u euklidovom prostoru:

$$d(p,q) = d(q,p) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \tag{2-1}$$

gdje je:

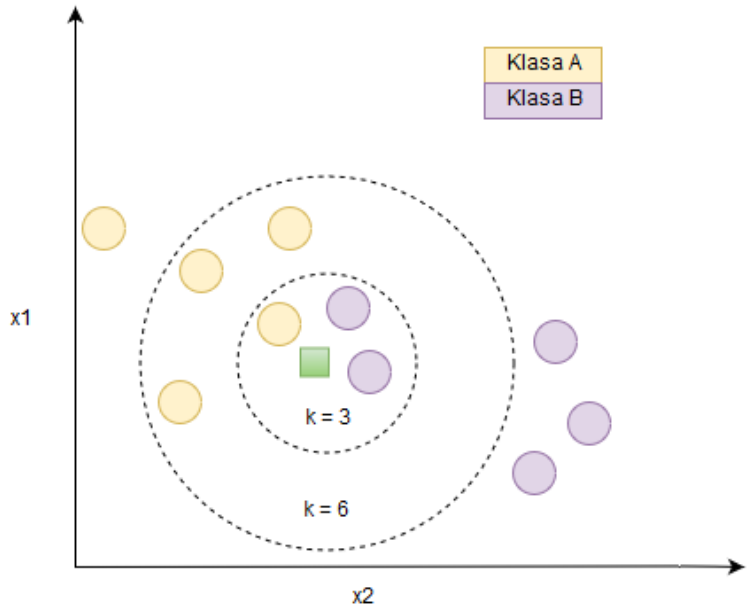
q_i - *i*-ta koordinata prve točke

p_i - *i*-ta koordinata druge točke.

Dvije točke su dvije poruke. Prva točka je poruka iz trening skupa, a druga točka je nova poruka učitana u sustav. Koordinate točaka, odnosno poruka, su opisne značajke koje je sustav izlučio. U ovom slučaju, sustav je iz poruka izlučio pet opisnih značajki, što znači da točke imaju pet koordinata. Nakon što sustav izračuna euklidsku udaljenost između nove učitane poruke i svake poruke iz trening skupa, sustav treba klasificirati novu poruku kao pravu ili kao neželjenu.

Postavlja se vrijednost *k*, odnosno koliko će se najbližih susjeda koristiti pri klasificiranju. Ako je *k*=5, koristiti će se pet najbližih susjeda. Pregledavaju se klase pet najbližih susjeda i ako većina klasa je prava poruka, nova poruka će biti klasificirana kao prava i obrnuto.

Udaljenost se može računati drugin metodama, poput Manhattan udaljenosti ili udaljenosti po Minkowskom. Na slici 2.5. izrađenoj prema [18], može se vidjeti graf koji prikazuje koliko na klasifikaciju utječe odabir broja *k*, odnosno broja najbližih podataka podatku kojeg se treba klasificirati. Ukoliko je *k*=3, podatak će biti klasificiran kao klasa B, a ukoliko je *k*=6, podatak će biti klasificiran kao klasa A.



Slika 2.5. *Klasifikacija novog podatka*

2.3. Matrica zabune

Prema [19], matrica zabune je matrica u kojoj svaki redak predstavlja predviđenu klasu dok stupac predstavlja stvarnu klasu. Matrica zabune je uvijek simetrična, odnosno ima jednak broj stupaca koliko ima redaka, a taj broj je jednak broju klasa. Matrica zabune prikazana je u tablici 2.1.

Tablica 2.1 Matrica Zabune

		Stvarne klase	
		0	1
Predviđene klase	0	TP	FP
	1	FN	TN

U tablici 2.1. oznaka *TP* predstavlja stvarno pozitivne (engl. *true positive*) poruke, odnosno broj pravih poruka koje je klasifikator označio kao prave. Oznaka *FP* predstavlja lažno pozitivne (engl. *false positive*) poruke, odnosno broj neželjenih poruka koje je klasifikator označio kao prave. Oznaka *TN* predstavlja stvarno negativne (engl. true negative) poruke, odnosno broj neželjenih poruka koje je klasifikator označio kao neželjene. Oznaka *FN* predstavlja lažno negativne (engl. *false negative*) poruke, odnosno broj pravih poruka koje je klasifikator označio kao neželjene. Iz matrice zabune može se izračunati pokazatelje tog klasifikacijskog modela a oni su sljedeći:

$$Točnost = \frac{TP + TN}{TP + TN + FP + FN} \quad (2-2)$$

$$UPK = 1 - Točnost \quad (2-3)$$

$$PK0 = \frac{TP}{TP + FP} \quad (2-4)$$

$$PK1 = \frac{TN}{TN + FN} \quad (2-5)$$

$$Odziv = \frac{TP}{TP + FN} \quad (2-6)$$

$$Specifičnost = \frac{TN}{TN + FP} \quad (2-7)$$

$$F1 = 2 * \frac{PK0 * Odziv}{PK0 + Odziv} \quad (2-8)$$

Iz jednadžbe (2-2) se može izračunati postotak točno klasificiranih poruka u odnosu na sve klasificirane poruke. Jednadžba (2-3) se koristi za izračun postotka krivo klasificiranih poruka u odnosu na sve klasificirane poruke. Jednadžba (2-4) se koristi za izračun postotka koji prikazuje koliki postotak od ukupno klasificiranih poruka kao klasa nula je prava klasa nula. Iz jednadžbe (2-5) se može izračunati koliki postotak od ukupno klasificiranih poruka kao klasa jedan je prava klasa jedan. Jednadžba (2-6) se koristi da bi se odredilo koliko je poruka program uspio klasificirati kao klasa nula od svih poruka koje su u stvarnosti klasa nula. Jednadžba (2-7) se koristi da bi se odredilo koliko je poruka program uspio klasificirati kao klasa jedan od svih poruka koje su u stvarnosti klasa jedan. Jednadžba (2-8) prikazuje vrijednost za F-mjeru koja se može objasniti kao težinski prosjek preciznosti i odziva, pri čemu najveću vrijednost ostvaruje kod rezultata 1 a najmanju vrijednost pri rezultatu 0.

2.4. Opisne značajke korištene u klasifikaciji

Kako bi klasifikacija započela, prvo je potrebno naučiti klasifikator što je neželjena poruka, a što nije. Zbog toga je potrebno odrediti neke opisne značajke koje će se koristiti u klasifikaciji poruka. Postoji veliki broj informacija u svakoj poruci koji se mogu koristiti kao opisnu značajku. Opisne značajke su značajke koje opisuju određeni podatak. U slučaju SMS poruka, opisne značajke opisuju tu poruku kroz broj riječi poruke, broj alfanumeričkih znakova, broj znakova u samoj poruci i slično. Zbog toga, u daljnjem tekstu će biti prikazane opisne značajke koje se najčešće koriste u literaturi. Odabir opisnih značajki je iznimno važan za uspješnost klasifikatora jer neće svaka opisna značajka davati isti rezultat. Potrebno je tražiti posebnost u porukama i na njih se usmjeriti. Prema [20], u istraživanju se koristio brojač koji je brojio u porukama sva mjesta na kojima nije bilo alfanumeričkih znakova ili određenih posebnih znakova poput točke, zareza, povlake i slično. Takva opisna značajka je dobra jer učestalo pojavljivanje znakova koji se ne koriste u normalnom razgovoru, poput čestog pojavljivanja smajlova mogu značiti da je riječ o neželjenoj poruci. Često se, kao jedna od opisnih značajki, koristi broj znakova u poruci. Takva opisna značajka je iznimno korisna kod filtiranja SMS poruka. SMS poruke su ograničene na 160 znakova što dovodi do toga da pošiljatelj neželjene poruke će gotovo uvijek iskoristiti svih 160 znakova. Opisne značajke koje su se koristile u [17], su se pokazale dobrima jer su s određenim klasifikatorima pokazali dobre rezultate. Linearni SVM klasifikator je uspio uspješno prepoznati 97% neželjenih poruka koristeći opisnu značajku koja broji sve znakove koji nisu alfanumerički. Prema [6], dobre opisne značajke su broj riječi u poruci, dužina poruke, nepismenost koja se označava s 1 ili 0 pri čemu se provjerava da li nakon točke ili zareza ide razmak ili odmah sljedeća riječ. Nepismenost je dobar indikator jer ako je poruka prava, najčešće će biti informatički pravilno napisana. Informatički pravilno napisana poruka podrazumjeva određene norme koje se koriste pri pisanju elektroničkih poruka, poput stavljanja razmaka nakon zareza ili točke. Dobra opisna značajka je i količina brojeva u poruci jer neželjene poruke često sadrže brojeve telefona. U [6], kao opisne značajke, su se koristile i ključne riječi koje se mogu izolirati iz poruka, a često se nalaze u neželjenim porukama. Također, korisno je poistovjećivati riječi istog korijena, poput *teacher* i *teaching* jer ako se jedna može poistovjetiti s neželjenom porukom ili pravom porukom, najvjerojatnije se onda i druga riječ, koja je istog korijena, može svrstati u tu skupinu.

2.5. Dostupni podatkovni skupovi

Podatkovni skupovi su kolekcije raznih neželjenih i pravih poruka koji su prikupljeni zbog lakšeg daljnjeg istraživanja i u svrhu testiranja. Postoje razni podatkovni skupovi, koje su najčešće

napravili istraživači ili tvrtke koje se bave telekomunikacijama. U ovome poglavlju prikazat će se neki podatkovni skupovi i objasnit će se kako su nastali, tko ih je skupio i gdje su bili korišteni. Također, objasniti će se i podatkovni skup koji će biti korišten u programskom dijelu ovog rada.

SMS Spam Collection v. 1 je javni podatkovni skup SMS poruka koji se koristi za istraživanja neželjenih poruka u SMS porukama. Sadrži 5574 engleskih poruka, koje su označene sa *ham* ili *spam* što označava da li je poruka prava ili neželjena. Ova kolekcija je sastavljena iz raznih izvora, potpuno besplatnih ili besplatnih ako se koriste u svrhu istraživanja. Kolekcija se sastoji od 425 poruka koje su ručno prepisane sa stranice *Grumbletext*. To je forum u kojem korisnici objavljuju neželjene SMS poruke. Drugi dio čini 3375 pravih poruka nasumično odabranih iz *NUS SMS Corpus* podatkovnog skupa koji se sastoji od 10000 pravih poruka sakupljenih na zavodu za računarstvo sveučilišta u Singapuru. Sljedeći dio skupa čini 450 pravih poruka prikupljenih iz doktorskog rada Caroline Tag. Završni dio čini 1002 prave poruke i 322 neželjenih poruka koje su skupili istraživači T.A. Almeida i J.M.Gomez, koji su ujedno i stvorili ovu kolekciju [21].

SMS Spam Corpus v.0.1 je kolekcija SMS poruka, koja je sakupljena u svrhu istraživanja prepoznavanja neželjenih SMS poruka. Sadrži dvije kolekcije SMS poruka i obje kolekcije sadrže poruke na engleskom jeziku. Jedna kolekcija sadrži 1084 poruke, dok druga sadrži 1319 poruka. U kolekcijama, poruke su označene sa *ham* ili *spam*, odnosno da li je neka poruka prava ili neželjena. Kolekcija je sakupljena iz besplatnih izvora s interneta. Manja verzija ove kolekcije sadrži 1002 prave poruke i 82 neželjene poruke, a veća verzija sadrži 1002 prave poruke i 322 neželjene poruke. Kolekciju su sakupili istraživači J.M.G.Hidalgo i E.P.Sanz [22].

DIT SMS Spam Dataset je kolekcija koja sadrži 1353 različite SMS neželjene poruke. Kolekcija je sakupljena s dvije internetske stranice u Velikoj Britaniji gdje korisnici mogu ostavljati svoje pritužbe. Svaka poruka sadrži datum kada je prikupljena, a kolekcija je sakupljena u periodu od kasne 2003. g do sredine 2010. g. Sve poruke u kolekciji su napisane na engleskom jeziku. Kolekcija je napisana u XML (engl. *Extensible Markup Language*) datoteci [23].

U ovome radu koristiti će se *SMS Spam Corpus v.0.1* kolekcija SMS poruka, te će i samo programsko rješenje biti napisano na poseban način kako bi sustav mogao, pri čitanju datoteke, saznati koja je poruka prava, a koja neželjena.

3. PROGRAMSKO RJEŠENJE ZA RASPOZNAVANJE NEŽELJENIH PORUKA

U programskom dijelu rada će se prikazati program koji će pomoću pet različitih klasifikatora pokušati prepoznati neželjene SMS poruke. Program će se istrenirati na trening skupu, pri čemu će znati koje su poruke prave, a koje neželjene. Nakon toga, program će se testirati tako što će se u program učitati ostatak poruka iz podatkovnog skupa iz kojeg je sastavljen testni skup i pokušati će se odrediti klase porukama.

3.1. Specifikacija zahtjeva na sustav za prepoznavanje

Sustav za prepoznavanje neželjenih poruka je podjeljen u dvije faze. U prvoj fazi, sustav će se istrenirati na trening skupu, koji je izlučen iz podatkovnog skupa koji se koristi u ovom radu, kako bi stekao dojam što je neželjena poruka, a što prava poruka. Sustav će izlučiti sve opisne značajke iz poruka koje pripadaju trening skupu. Opisne značajke koje će se koristiti u ovom sustavu su djelomično preuzete iz [6]. Brojat će se koliko je znakova u poruci iskorišteno za slanje poruke jer je poznato da jedna SMS poruka sadrži najviše 160 znakova. Provjeravati će se i broj nealfanumeričkih znakova u poruci, odnosno broj točki, zareza i slično. Potrebno je i provjeriti koliko znakova u poruci su brojevi jer će neželjene poruke često sadržavati brojeve mobitela ili telefona. Također, kao opisna značajka će se promatrati i postoje li u poruci posebni znakovi koji se ne mogu prikazati koristeći *Unicode Latin-1* standard. Ova opisna značajka će se prikazivati kao nula ili jedan, odnosno je li posebni znak prisutan ili ne. Na kraju, sustav će provjeriti i koliko je SMS poruka potrebno za slanje jedne poruke jer jedna SMS poruka sadrži najviše 160 znakova. U drugoj fazi, sustav će pokušati odrediti klase poruka iz testnog skupa koji ima približno jednak omjer pravih i neželjenih poruka kao i trening skup.Program će biti ugrađen u *Python* programskom jeziku. Svaka učitana poruka će biti zapisana u matricu sa šest stupaca. U pet stupaca će biti zapisane vrijednosti opisnih značajki dok će u šesti stupac biti upisana klasa poruke, odnosno *ham* ili *spam*. Za izradu programa, bilo je potrebno uvesti nekoliko biblioteka kako bi se prilikom izrade programa mogle koristiti dodatne funkcije.

3.2. Opis korištenih python biblioteka

Na slici 3.1. se može vidjeti dio programa gdje se u program uvoze biblioteke korištene prilikom izrade programa. Kako bi se omogućilo crtanje prilikom izvođenja programa, uvežena je biblioteka *matplotlib.pyplot*. Biblioteka *numpy* se koristi kako bi se omogućilo *pythonu* da radi s matricama, odnosno da se mogu obavljati sve operacije s matricama, poput množenja matrica, zbrajanja

matrica i slično. Biblioteka *re* se koristi za pisanje regularnih izraza, odnosno izraza koji opisuju neki tekst jer svaki tekstualni izraz možemo zapisati kao niz nekih posebnih znakova. Biblioteka *sklearn.metrics* sadrži funkcije koje se koriste pri strojnom učenju. Iz ove biblioteke, koristi se samo *confusion_matrix* funkcija koja služi za računanje matrice zabune. Biblioteka *nltk* je biblioteka za obradu prirodnog jezika (engl. *Natural Language ToolKit*), a iz ove biblioteke se uvozi samo funkcija *word_tokenize* koja služi kako bi se poruka pretvorila u listu riječi. Biblioteka *collections* se uvozi kako bi se koristio brojač (engl. *counter*) koji se koristi za sortiranje i zbrajanje istih riječi koji se mogu pojaviti u poruci.

```
import matplotlib.pyplot as plt
import numpy as np
import re
from sklearn.metrics import confusion_matrix
from nltk import word_tokenize as tkn
from sklearn.neighbors import KNeighborsClassifier as KNN
from collections import Counter as Cnt
```

Slika 3.1 *Korištene biblioteke*

Na slici 3.2. prikazane su biblioteke koje su korištene za rad s klasifikatorima. Iz biblioteke *sklearn.metrics* uvežena je funkcija koja se koristi pri izračunu mjere F.

```
from sklearn.neighbors import KNeighborsClassifier as KNN
from sklearn.neural_network import MLPClassifier as MLP
from sklearn.ensemble import RandomForestClassifier as RF
from sklearn.svm import SVC
from sklearn.naive_bayes import GaussianNB as GNB
from sklearn.metrics import f1_score as f1
```

Slika 3.2. *Biblioteke s klasifikatorima*

3.3. Korištene metode dohvaćanja značajki i klasifikacije

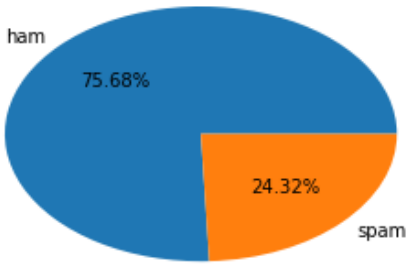
Kako bi se prepoznale neželjene poruke, implementirana je metoda klasifikacije pomoću pet različitih klasifikatora. Klasifikatori se koriste kako bi porukama dodjelili klasu, a u slučaju ovog rada klase su *spam* i *ham*, odnosno neželjena poruka i prava poruka. Svrha klasifikatora je

1. *Učitati podatke trening skupa koji će se koristiti za predviđanje daljnih podataka te izlučiti iz njih opisne značajke i učitavanje u m-dimenzijonalni prostor gdje je m broj značajki*
2. *Učitavanje testnih podataka, izlučivanje opisnih značajki i učitavanje podataka u m-dimenzijonalni prostor*
3. *Svaki klasifikator obrađuje unesene rečenice iz testnog skupa*
4. *Svaki klasifikator dodjeljuje klasu rečenici iz testnog skupa*
5. *Rješenja svakog klasifikatora se zapisuju u matricu zabune i računaju se vrijednosti pomoću kojih se može vidjeti uspješnost klasifikatora*

Blok 3.1 *Pseudokod programa*

klasificiranje podataka pomoću opisnih značajki koje će program izlučiti iz poruka. U bloku 3.2 se može vidjeti pseudokod programa.

Opisne značajke, koje će se koristiti u ovom radu, opisane su u poglavlju 3.1. Koristit će se broj znakova u poruci, broj nealfanumeričkih znakova, broj numeričkih znakova u poruci, postojanost posebnih znakova koja će se označavati s nula ili jedan te broj SMS poruka potrebnih za slanje jedne poruke. Slika 3.3. prikazuje izgled podatkovnog skupa koji se koristi u programu. U skupu se nalaze 1002 prave poruke i 322 neželjene poruke.



Slika 3.3. *Udio pravih i neželjenih poruka u skupu*

3.4. Prikaz ključnih dijelova programa

U ovom dijelu rada su prikazani najvažniji dijelovi programa i ti dijelovi su ukratko objašnjeni. Program se sastoji od 2 dijela. Prvi dio služi za učitavanje poruka i stvaranje podataka koji se koriste pri prepoznavanju poruka tijekom klasifikacije. Drugi dio programa služi za prepoznavanje pravih i neželjenih poruka na temelju dobivenih podataka. Na slici 3.4. je prikazan dio koda koji

služi za učitavanje poruka i stvaranje podataka, odnosno izlučivanje opisnih značajki nužnih za klasifikaciju. Funkcija *importAndPreproces* će učitati datoteku s porukama, pregledati koliki je broj poruka u datoteci, stvoriti će matricu sa šest stupaca i pozvati će funkciju *getFeatures*. Funkcija *getFeatures* se koristi za izlučivanje opisnih značajki svake rečenice i te značajke se zapisuju u matricu. Prvih pet stupaca služe za upisivanje vrijednosti svake značajke dok šesti stupac služi za upisivanje klase poruke

```
def getFeatures(line):
    temp=line.lower().decode('latin1')
    last=len(tkn(temp)[-1])
    noChar=(len(line)-last-1)
    noSMS=np.ceil(noChar/160.0)
    noNum=sum([len(word) for word in tkn(temp) if re.match(r'.*[0-9]+.*',word)])
    no_nonAlpNum=0
    unknown=0;
    for i in xrange(noChar):
        if not temp[i].isalnum() and not temp[i].isspace():
            no_nonAlpNum=no_nonAlpNum+1
        if temp[i]==u'\xa3':
            unknown=1
    return noChar,noSMS,no_nonAlpNum,noNum,unknown,last

def importAndPreproces(dat):
    f=open(dat,'r')
    MailStats=np.zeros((1,6))
    for line in f:
        noChar,noSMS,no_nonAlpNum,noNum,unknown,last=getFeatures(line)
        MailStats=np.vstack((MailStats,[noChar/(160.0*noSMS),
                                float(no_nonAlpNum)/noChar,
                                float(noNum)/noChar,
                                unknown,
                                noSMS,
                                last-3]))
    f.close()
    return np.random.permutation(MailStats[1:,:])
```

Slika 3.4. *Funkcije getFeatures i importAndPreproces*

Program je odradio 10 ponavljanja za svaki klasifikator. Tijekom svakog ponavljanja, program je skup s porukama podjelio na trening i testni skup te je klasifikator ponovio treniranje i klasifikaciju testnog skupa 30 puta. Nakon odrađene klasifikacije, podaci se unose u matricu zabune i računaju se vrijednosti pomoću kojih se može vidjeti upsješnost svakog klasifikatora.

3.5. Prikaz rezultata programa

U ovom poglavlju prikazani su rezultati klasifikacije za svaki klasifikator. Vrijednosti su izračunate iz matrice zabune. Način na koji se vrijednosti računaju objašnjen je u poglavlju 2.3. Svaka vrijednost u tablici je prosječna vrijednost od 30 klasifikacija koji svaki klasifikator odradi tijekom jednog ponavljanja. Za svaku vrijednost izračunata je i standardna devijacija, odnosno prosječno odstupanje od prosjeka. Tablica 3.1. prikazuje uspješnost klasifikacije pomoću neuronskih mreža. Iz tablice 3.1. se može vidjeti da je ANN klasifikator ostvario prosječnu točnost pri klasifikaciji od 96.45%. Također, može se primjetiti i da je klasifikator uspješniji pri prepoznavanju pravih poruka što se može iščitati iz prosječnog odziva od 97.98%, dok iz prosječne specifičnosti se može iščitati uspješnost prepoznavanja neželjenih poruka koja iznosi 93.64%.

Tablica 3.1. Uspješnost klasifikacije za ANN klasifikator

#	Mjera F	Točnost	Preciznost	Odziv	Specifičnost
1	0,927 ± 0,021	96,566 ± 1,073	98,102 ± 1,080	97,410 ± 1,389	94,041 ± 3,236
2	0,933 ± 0,024	96,750 ± 1,102	98,123 ± 1,145	97,561 ± 1,242	94,423 ± 3,318
3	0,928 ± 0,025	96,625 ± 1,425	98,290 ± 1,021	97,292 ± 1,500	94,552 ± 2,991
4	0,924 ± 0,025	96,147 ± 1,222	97,409 ± 0,885	97,428 ± 1,444	92,499 ± 2,359
5	0,927 ± 0,029	96,415 ± 0,977	97,923 ± 1,021	97,318 ± 1,324	93,780 ± 2,944
6	0,928 ± 0,024	96,608 ± 1,480	97,981 ± 1,539	97,569 ± 1,101	93,621 ± 2,989
7	0,926 ± 0,024	96,533 ± 1,224	98,066 ± 1,301	97,409 ± 1,234	93,901 ± 4,065
8	0,932 ± 0,022	96,725 ± 1,479	98,161 ± 1,280	97,509 ± 1,505	94,348 ± 3,550
9	0,932 ± 0,029	96,776 ± 1,201	98,201 ± 1,080	97,560 ± 1,250	94,399 ± 3,558
10	0,926 ± 0,027	96,441 ± 1,326	97,990 ± 1,276	97,314 ± 1,118	94,814 ± 3,505
$\mu \pm \sigma$	0,928 ± 0,025	96,448 ± 1,253	97,975 ± 1,163	97,352 ± 1,310	93,644 ± 3.257

Tablica 3.2. prikazuje uspješnost klasifikacije za kNN klasifikator. Prosječna točnost klasifikacije iznosi 96,52%. Iz tablice 3.2. može se zaključiti da je kNN klasifikator uspješniji pri prepoznavanju pravih poruka s prosječnim odzivom od 97,54%, dok je pri prepoznavanju neželjenih poruka ostvario prosječni rezultat od 93,41%. Kroz svih 10 ponavljanja programa, nema većih odstupanja mjere F pri čemu prosječno odstupanje mjere F je nisko i iznosi 0,020.

Tablica 3.2. Uspješnost klasifikacije za kNN klasifikator

#	Mjera F	Točnost	Preciznost	Odziv	Specifičnost
1	0,929 ± 0,019	96,516 ± 0,856	97,669 ± 1,058	97,722 ± 0,890	92,862 ± 2,881
2	0,924 ± 0,017	96,323 ± 1,005	97,649 ± 0,813	97,477 ± 1,081	92,752 ± 2,357
3	0,925 ± 0,022	96,482 ± 0,954	97,745 ± 1,099	97,641 ± 0,934	92,754 ± 3,075
4	0,932 ± 0,019	96,608 ± 1,035	97,941 ± 0,864	97,544 ± 1,172	92,852 ± 2,564
5	0,933 ± 0,022	96,918 ± 0,838	98,180 ± 0,987	97,799 ± 0,886	92,995 ± 3,133
6	0,928 ± 0,017	96,499 ± 0,947	97,846 ± 0,821	97,526 ± 0,970	92,352 ± 2,404
7	0,928 ± 0,022	96,616 ± 1,173	98,146 ± 1,294	97,433 ± 1,079	92,904 ± 3,632
8	0,929 ± 0,021	96,441 ± 0,775	97,702 ± 0,744	97,555 ± 0,968	92,226 ± 2,301
9	0,927 ± 0,021	96,382 ± 0,984	97,726 ± 0,842	97,471 ± 0,994	93,147 ± 2,926
10	0,925 ± 0,023	96,348 ± 1,053	97,756 ± 0,999	97,428 ± 1,184	92,979 ± 3,006
$\mu \pm \sigma$	0,927 ± 0,020	96,516 ± 0,962	97,905 ± 0.952	97,536 ± 1,016	93,412 ± 2,828

Tablica 3.3. prikazuje uspješnost klasifikacije za klasifikator Naivni Bayes. Prosječna točnost klasifikacije iznosi 91,57%. Naivni Bayesov klasifikator ostvario je bolji rezultat pri prepoznavanju neželjenih poruka što se može vidjeti u prosječnoj specifičnosti koja iznosi 93,53%. Iz odziva, koji iznosi 91.24%, može se vidjeti uspješnost klasifikatora pri prepoznavanju pravih poruka. Iz tablice 3.3. može se vidjeti i veće standardno odstupanje od prosjeka kod odziva i specifičnosti.

Tablica 3.3. Uspješnost klasifikacije za klasifikator Naivni Bayes

#	Mjera F	Točnost	Preciznost	Odziv	Specifičnost
1	0,816 ± 0,085	92,588 ± 2,896	98,439 ± 1,569	92,475 ± 3,659	93,400 ± 4,810
2	0,804 ± 0,103	92,295 ± 3,164	98,551 ± 1,272	92,120 ± 3,909	93,870 ± 4,523
3	0,826 ± 0,088	93,090 ± 4,065	98,212 ± 1,241	93,232 ± 4,860	93,126 ± 5,406
4	0,752 ± 0,074	89,623 ± 3,697	98,662 ± 1,566	88,682 ± 4,796	94,750 ± 5,068
5	0,792 ± 0,111	91,817 ± 3,311	98,504 ± 1,391	91,554 ± 4,054	93,745 ± 6,071
6	0,778 ± 0,108	91,064 ± 4,265	98,457 ± 1,443	90,714 ± 4,963	93,937 ± 5,132
7	0,780 ± 0,114	91,173 ± 3,538	97,957 ± 1,090	91,284 ± 4,090	91,933 ± 4,312
8	0,768 ± 0,111	90,620 ± 3,504	98,384 ± 1,301	90,278 ± 4,321	93,616 ± 4,645
9	0,816 ± 0,091	92,370 ± 3,596	98,492 ± 1,511	92,100 ± 4,150	93,900 ± 5,220
10	0,802 ± 0,087	92,035 ± 2,708	98,503 ± 1,415	91,724 ± 3,473	93,801 ± 5,767
$\mu \pm \sigma$	0,792 ± 0,097	91,569 ± 3,474	98,380 ± 1,382	91,343 ± 4.228	93,525 ± 5,095

Tablica 3.4. prikazuje uspješnost klasifikacije za klasifikator šuma odluke. Prosječna točnost klasifikacije iznosi 99,24%. Klasifikator šuma odluke pokazao se kao najtočniji klasifikator pri

prepoznavanju klasa poruka. Uspješniji je pri prepoznavanju pravih poruka što se vidi iz odziva gdje prosječni odziv iznosi 99.47%, dok je pri prepoznavanju neželjenih poruka ostvario prosječni rezultat od 98.17% što se može vidjeti iz prosječne specifičnosti. Također, sve vrijednosti iz tablice 3.4. imaju mala standardna odstupanja osim specifičnosti gdje prosječno standardno odstupanje iznosi 2.27%.

Tablica 3.4. Uspješnost klasifikacije za klasifikator šuma odluke

#	Mjera F	Točnost	Preciznost	Odziv	Specifičnost
1	0,987 ± 0,012	99,405 ± 0,610	99,605 ± 0,535	99,567 ± 0,637	98,808 ± 2,738
2	0,979 ± 0,014	99,087 ± 0,550	99,328 ± 0,484	99,440 ± 0,797	97,911 ± 1,725
3	0,983 ± 0,017	99,313 ± 0,484	99,543 ± 0,554	99,534 ± 0,415	98,401 ± 2,639
4	0,983 ± 0,015	99,146 ± 0,549	99,458 ± 0,558	99,351 ± 0,511	98,617 ± 1,603
5	0,979 ± 0,021	99,087 ± 0,617	99,321 ± 0,621	99,443 ± 0,661	97,448 ± 1,956
6	0,977 ± 0,020	99,171 ± 0,806	99,373 ± 0,923	99,538 ± 0,808	97,649 ± 2,270
7	0,978 ± 0,020	99,162 ± 0,501	99,989 ± 0,546	99,488 ± 0,437	97,046 ± 2,229
8	0,985 ± 0,013	99,330 ± 0,819	99,536 ± 0,432	99,540 ± 0,849	98,480 ± 2,010
9	0,979 ± 0,021	99,171 ± 0,650	99,467 ± 0,748	99,411 ± 0,616	98,246 ± 3,294
10	0,983 ± 0,021	99,380 ± 0,588	99,626 ± 0,611	99,560 ± 0,566	98,559 ± 2,200
$\mu \pm \sigma$	0,980 ± 0,017	99,235 ± 0,617	99,519 ± 0,602	99,469 ± 0,629	98,171 ± 2,267

Tablica 3.5. prikazuje uspješnost klasifikacije za SVM klasifikator. Prosječna točnost klasifikacije iznosi 94,56%. SVM klasifikator pokazao se uspješnijim pri klasifikaciji pravih poruka s prosječnim odzivom od 95.06%, dok je prosječno točno klasificirao 94.27% neželjenih poruka što se može vidjeti iz prosječne specifičnosti. Iz tablice 3.5. može se vidjeti da SVM klasifikator ima veća standardna odstupanja od prosjeka od ostalih klasifikatora.

Tablica 3.5. Uspješnost klasifikacije za SVM klasifikator

#	Mjera F	Točnost	Preciznost	Odziv	Specifičnost
1	0,875 ± 0,070	95,008 ± 2,837	98,220 ± 1,486	95,352 ± 3,665	94,284 ± 4,175
2	0,868 ± 0,084	94,665 ± 3,381	98,065 ± 2,536	95,143 ± 4,163	94,158 ± 6,290
3	0,879 ± 0,053	95,092 ± 2,177	98,922 ± 2,434	94,864 ± 3,091	96,175 ± 6,460
4	0,878 ± 0,079	94,430 ± 3,063	98,146 ± 2,597	94,594 ± 3,352	93,870 ± 6,227
5	0,864 ± 0,083	94,079 ± 2,706	97,334 ± 1,150	95,035 ± 3,137	92,592 ± 3,716
6	0,845 ± 0,093	94,439 ± 3,305	98,697 ± 3,853	94,383 ± 3,826	95,333 ± 7,261
7	0,854 ± 0,065	94,045 ± 2,576	98,116 ± 1,557	94,284 ± 2,905	94,469 ± 4,960
8	0,875 ± 0,087	94,489 ± 3,521	98,446 ± 2,209	94,560 ± 4,076	95,779 ± 4,778
9	0,872 ± 0,099	95,017 ± 3,577	98,319 ± 3,659	95,359 ± 3,639	94,891 ± 7,920
10	0,884 ± 0,080	95,008 ± 3,198	97,813 ± 3,209	95,727 ± 3,849	94,455 ± 7,930
$\mu \pm \sigma$	0,869 ± 0,079	94,599 ± 3,034	98,051 ± 2,469	95,056 ± 3,570	94,272 ± 5,972

U tablicama 3.6. i 3.7. prikazana je usporedba klasifikatora prema najvećim i najmanjim dobivenim vrijednostima tijekom klasifikacije. Iz tablica se može zaključiti da je klasifikator šuma odluke najuspješniji pri određivanju klasa poruka. Također, vidljivo je i da je klasifikator šuma odluke sa svojim najmanjim rezultatima bolji od ostalih klasifikatora pri određivanju klasa. Naivni Bayesov klasifikator je jedini klasifikator koji je točnije klasificirao neželjene poruke nego prave poruke.

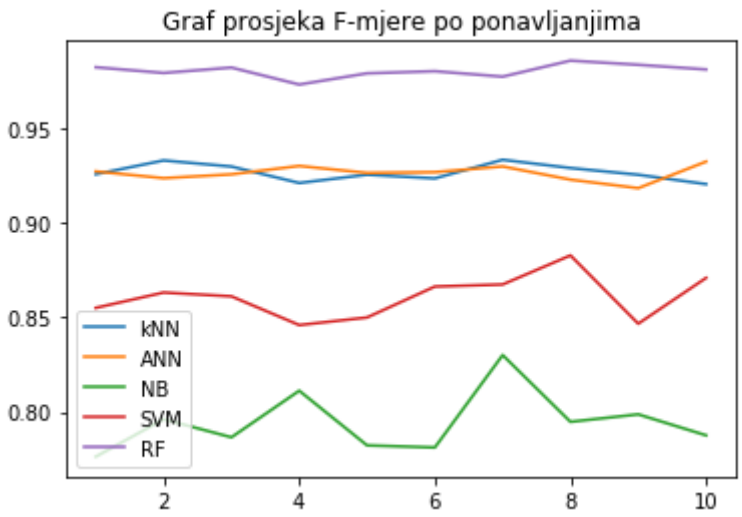
Tablica 3.6. Usporedba klasifikatora prema najvećim dobivenim vrijednostima

Klasifikator	Mjera F	Točnost	Preciznost	Odziv	Specifičnost
ANN	0,933	96,776	98,290	97,561	94,814
kNN	0,933	96,918	98,180	97,799	93,147
Naivni Bayes	0,826	93,090	98,662	93,232	94,750
Šuma odluke	0,987	99,405	99,989	99,567	98,808
SVM	0,884	95,092	98,992	95,727	96,175

Tablica 3.7. Usporedba klasifikatora prema najmanjim dobivenim vrijednostima

Klasifikator	Mjera F	Točnost	Preciznost	Odziv	Specifičnost
ANN	0,926	96,147	97,409	97,292	92,499
kNN	0,924	96,323	97,649	97,428	92,226
Naivni Bayes	0,752	89,623	97,957	88,682	91,933
Šuma odluke	0,977	99,087	99,321	99,351	97,046
SVM	0,845	94,045	97,334	94,284	92,592

Na slici 3.6. prikazana je usporedba klasifikatora prema mjeri F. Graf prikazuje kako je klasifikator šuma odluke najuspješniji pri određivanju klasa porukama. Također, iz grafa je vidljivo da klasifikatori šuma odluke, ANN i kNN imaju znatno manja standardna odstupanja od prosjeka dok SVM klasifikator i naivni Bayesov klasifikator imaju veća odstupanja što se vidi kako njihov graf više „skače“ po ponavljanjima programa.



Slika 3.6. *Usporedba klasifikatora prema mjeri F*

4. ZAKLJUČAK

U ovom radu objašnjen je problem neželjenih poruka. Neželjene poruke pokazale su se kao značajan problem u današnjem svijetu, posebno poslovnim krugovima. U začetcima interneta, neželjene poruke pojavljivale su se samo pri elektroničkoj pošti, no danas su prisutne u gotovo svim načinima digitalne komunikacije. Elektronička pošta, SMS poruke, instant poruke i društvene mreže su samo neki od digitalnih komunikacijskih kanala koji su danas meta neželjenih poruka. Glavna odlika takvih poruka je da se u što kraćem vremenu pošalju što većem broju korisnika. Osim što se preko neželjenih poruka najčešće šalje razni reklamni sadržaj, u zadnje vrijeme se pojavljuje sve veći broj poruka koje u sebi imaju skrivene viruse ili neke druge sigurnosne prijetnje, s ciljem da oštete uređaj krajnjeg korisnika. Sve se više pojavljuju i poruke preko kojih se korisnicima nude lažne nagrade, a jedina svrha je da im korisnik prosljedi svoj broj kartice i PIN. Kao odgovor na sve veći broj neželjenih poruka, pokrenuta su razna istraživanja kako otkriti takve poruke i razvijeni su mnogi klasifikatori koji se koriste pri prepoznavanju neželjenih poruka. Klasifikatori koji se u praksi najčešće koriste su kNN, naivni Bayes, SVM i šuma odluke. U praktičnom dijelu ovog rada, osmišljen je program koji uz pomoć pet klasifikatora klasificira SMS poruke i tako pokušava otkriti koje su poruke prave a koje neželjene. Pri izradi programa, potrebno je skup poruka pri svakom ponavljanju podjeliti na testni i trening skup kako bi se dobili što bolji rezultati pri svakom ponavljanju. Upravo je to bio jedan od početnih problema pri izradi programa. Također, program bi se mogao unaprijediti dodatnim funkcijama poput omogućavanja korisniku da unosi svoje poruke ili stvaranjem korisničkog sučelja koje bi olakšalo korištenje programa. Program se pokazao uspješnim pri čemu je klasifikator šuma odluke ostvario najveći uspjeh pri prepoznavanju klasa poruka.

LITERATURA

[1] E. Allman, The Economics of Spam, Queue, br. 9, str. 78-79, Prosinac 2003.

[2] G. V. Cormack, J. M. Gomez, E. P. Sanz, Spam Filtering for Short Messages, Conference on information and knowledge, Lisbon, Portugal, 6-10 Studeni 2007.

[3] The Spamhaus Project, The Definition of Spam, https://www.spamhaus.org/consumer/definition/, pristupljeno 4.6.2016.

[4] Spam Laws, Spam Statistics and Facts, http://www.spamlaws.com/spam-stats.html, pristupljeno 4.6.2016.

[5] Groupmail, What is spam, http://group-mail.com/email-marketing/what-is-spam/, pristupljeno 28.6.2016.

[6] O. Stipetić, G.Valentić, V. Vazdar, Prepoznavanje spam SMS poruka, https://web.math.pmf.unizg.hr/nastava/su/index.php/download_file/-/view/145/, pristupljenno 12.6.2016.

[7] J. M. G. Hidalgo, G . C. Bringas, E .P . Sanz, F. C. Garcia, Content Based SMS Spam Filtering, ACM Symposium on Document engineering (DocEng '06), str. 107-144, Amsterdam, The Netherlands, 10-13 Listopad 2006.

[8] R. Broida, How to block text message spam on your iPhone(AT-T), http://www.cnet.com/how-to/how-to-block-text-message-spam-on-your-iphone-at-t/, pristupljeno 28.6.2016.

[9] C .Biever, Spam being rapidly outpaced by spim, https://www.newscientist.com/article/dn4822-spam-being-rapidly-outpaced-by-spim/, pristupljeno 10.6.2016.

[10] S. Basu, Instant Messenger Hacks: 10 Security Tips to Protect Yourself, https://www.makeuseof.com/tag/10-security-tips-to-protect-yourself-from-instant-messenger-hacks/, pristupljeno 30.1.2018.

[11] D. Tynan, Social spam is taking over the Internet, http://www.itworld.com/article/2832566/it-management/social-spam-is-taking-over-the-internet.html, pristupljeno 17.6.2016.

[12] D. Bisson, A Guide on 5 Common Twitter Scams, https://www.tripwire.com/state-of-security/security-awareness/a-guide-on-5-common-twitter-scams/, pristupljeno 24.1.2018.

[13] S. Polamuri, Supervised and unsupervised learning, http://dataaspirant.com/2014/09/19/supervised-and-unsupervised-learning/, pristupljeno 15.2.2018.

[14] A. Santacruz, Why it is important to work with balanced classification dataset, http://amsantac.co/blog/en/2016/09/20/balanced-image-classification-r.html, pristupljeno 15.2.2018.

[15] M. N. Murty i V. S. Devi, Pattern Recognition An Algorithmic Approach, Springer, SAD, 2011.

[16] J. J. Eberhardt, Bayesian Spam Detection, Scholarly Horizons: University of Minnesota, Morris Undergraduate Journal, br. 1, sv. 2, Ožujak 2015.

[17] L. Baoli, Y. Shiwen, L. Qin, An Improved k-Nearest Neighbour Algorithm for Text Categorization, Lipanj 2003.

[18] Technology of computing blog, A Short Introduction to K-Nearest Neighbours Algorithm, https://helloacm.com/a-short-introduction-to-k-nearest-neighbors-algorithm/, pristupljeno 17.2.2018.

[19] K. Markham, Simple guide to confusion matrix terminology, http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/, pristupljeno 27.8.2017.

[20] S. J. Delany, M. Buckley, D. Greene, SMS spam filtering: Methods and data, Expert Systems and Applications, br. 10, sv. 39, str. 9899-9908, Kolovoz 2012.

[21] SMS Spam Collection v. 1, http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/, pristupljeno 24.7.2017.

[22] SMS Spam Corpus v.0.1, http://www.esp.uem.es/jmgomez/smsspamcorpus/, pristupljeno 25.7.2017.

[23] DIT SMS spam dataset, http://www.dit.ie/computing/research/resources/smsdata/, pristupljeno 25.7.2017.

SAŽETAK

U ovome radu objašnjen je problem neželjenih SMS poruka i prikazan je program koji koristi pet klasifikatora za prepoznavanje neželjenih i pravih SMS poruka. Za klasificiranje poruka koristili su se sljedeći klasifikatori: kNN, SVM, naivni Bayes, šuma odluke i umjetne neuronske mreže. U teorijskom dijelu rada opisana je problematika neželjenih poruka i nekoliko metoda kojima se takva problematika rješava. U praktičnom dijelu rada osmišljen je program koji klasificira SMS poruke pomoću pet klasifikatora koristeći opisne značajke poruka. Podaci u skupu se prvo dijele na trening i testni skup poruka u kojima postoji približno jednak omjer između pravih i neželjenih SMS poruka, a nakon toga se svaka poruka obrađuje na temelju zadanih opisnih značajki. U sljedećem koraku se program istrenira na trening skupu koje su poruke prave a koje neželjene i onda na testnom skupu pokušava prepoznati klase poruka. Klasifikator šuma odluke se pokazao kao najtočniji klasifikator pri prepoznavanju poruka.

Ključne riječi: kNN, naivni Bayes, neželjene poruke, prave poruke, SMS

ABSTRACT

This thesis researches the problem of spam and it presents a program which uses five classifiers in order to decipher between spam and real messages. The following classifiers were used to classify text messages: kNN, SVM, Naive Bayes, Random Forest and artificial neural netowork.The theoretical part of the thesis describes the problem of spam and several methods used to solve this problem. For the practical part, a program which classifies text messages with regard to their descriptive attributes was created. The program does this by the means of five classifiers.The data in the data set are first divided into the training and testing categories of messages where there is an approximately equal number of real and spam messages, and afterwards each SMS is processed on the basis of predefined descriptive attributes. In the next step, the program learns, on the basis of a training set, which messages are real and which are spam, and then it tries to recognize the classes text messages belong to in a testing set.The research conducted showed the random forest classifier to have been the most accurate one in deciphering between different messages.

Keywords: kNN, naive bayes, spam messages, real messages, SMS

Životopis

Eugen Tubić rođen je 6. kolovoza 1994. godine u Slavonskom Brodu, Hrvatska. Nakon završene osnovne škole, školovanje nastavlja u Gimnaziji Matija Mesić, smjer prirodoslovno – matematička gimnazija. Akademske godine 2013./2014. upisuje Fakultet elektrotehnike, računarstva i nformacijskih tehnologija, smjer računarstvo.

_____

Eugen Tubić

PRILOZI:

1. „Prepoznavanje neželjenih SMS poruka“ u .docx formatu
2. „Prepoznavanje neželjenih SMS poruka“ u .pdf formatu
3. Izvorni kod