UNIVERZA V LJUBLJANI

FAKULTETA ZA DRUŽBENE VEDE

Matej Gorenšek

Bayesove metode razvrščanja nezaželene elektronske pošte

Diplomsko delo

Ljubljana, 2013

UNIVERZA V LJUBLJANI

FAKULTETA ZA DRUŽBENE VEDE

Matej Gorenšek

Mentor: doc. dr. Damjan Škulj

Bayesove metode razvrščanja nezaželene elektronske pošte

Diplomsko delo

Ljubljana, 2013

ZAHVALA

Hvala mentorju doc. dr. Damjanu Škulju za vso strokovno pomoč,
nasvete in moralno podporo pri pisanju diplomske naloge.

Hvala staršema, ker sta me vedno podpirala in mi stala ob strani.
Hvala, ker sta prenašala vse moje muhe in bila z mano potrpežljiva.
In hvala, ker nista izgubila upanja vame.
Rad vaju imam!

Hvala babici in dedku, ker sta poskrbela zame, ko sem potreboval pomoč.

Hvala vsem prijateljem. Skupaj smo se vedno imeli lepo.
Najboljši ste!

Bayesove metode razvrščanja nezaželene elektronske pošte

Komunikacija preko elektronske pošte je v zadnjih nekaj letih postala sestavni del našega vsakdana. Zaradi enostavne uporabe, učinkovitosti, hitrosti in nizkih stroškov je kmalu pridobila velik krog uporabnikov. V njej so priložnost videli tudi oglaševalci in jo začeli uporabljati za pošiljanje komercialnih ponudb, kar je vodilo v pojav nezaželene elektronske pošte. Količina poslane nezaželene elektronske pošte se je tako začela izjemno hitro povečevati in je v letu 2013 predstavljala kar 64 % vse poslane elektronske pošte. Tako je nezaželena elektronska pošta postajala vedno večji problem, ki uporabnikom preprečuje oziroma omejuje normalno uporabo elektronske pošte in ovira njihovo produktivnost. Zato so se pojavili različni ukrepi, s katerimi bi se zmanjšala količina poslanih nezaželenih elektronskih sporočil. Vendar ti niso dosegli večjega uspeha, saj so nezaželena elektronska sporočila še vedno nemoteno prihajala do uporabnikov elektronskih poštnih predalov. Ker je bil poskus omejitve zmanjšanja števila poslane nezaželene elektronske pošte neuspešen, se je pojavila ideja po njenem razvrščanju, s čimer bi dosegli, da ne bi motila uporabnikov pri uporabi elektronske pošte. Pojavili so se filtri nezaželene elektronske pošte, ki na osnovi analize vsebine in elementov elektronskega sporočila prepoznajo, ali je neko elektronsko sporočilo nezaželeno ali ne. Večina filtrov nezaželene elektronske pošte danes temelji na Bayesovi formuli o pogojnih verjetnostih, na podlagi katere delujejo njihovi algoritmi strojnega učenja. To jim omogoča, da so se zmožni sproti učiti in tako avtomatsko prilagajati raznim situacijam, zaradi česar jih pošiljatelji nezaželene elektronske pošte težje zaobidejo. V diplomski nalogi sem se osredotočil na štiri najpogostejše Bayesove metode, ki se uporabljajo pri razvrščanju elektronskih sporočil, in jih podrobneje opisal. Nato sem izvedel test primera učenja algoritmov z eno izmed opisanih metod na vzorcu elektronskih sporočil in pridobljene rezultate ocenil z ustreznimi merami uspešnosti.

Ključne besede: nezaželena elektronska pošta, Bayesove metode razvrščanja, verjetnost, mere uspešnosti, algoritmi strojnega učenja, razvrščanje.

Bayesian spam filtering techniques

Email correspondence has become an integral part of our daily communication in recent years. It has gained a wide following due to its ease of use, speed and low costs. Email was soon seen as an opportunity by advertisers, who started using it for sending commercial offers, which led to the emergence of unsolicited commercial email, better known as spam. The volume of spam started to increase rapidly and in 2013 accounted for an estimated 64 % of all sent email. It has thus become a growing problem which prevents or at least limits the normal use of email and hinders the users' productivity. As a result, various measures were developed with the aim of reducing the amount of spam. However, with little success, since unwanted email messages were still reaching email users. Due to the failure to limit the amount of spam, the idea to filter it and in that way limit its impact on the email users emerged. Hence, spam filters based on content and element analysis were created; they rcognize whether or not an email is spam. Nowadays, most spam filters are based on the Bayesian formula for conditional probability which forms the basis for their machine learning algorithms. It enables them to learn continuously and thus automatically adapt to various situations, which makes them more difficult to bypass by the spammers. This diploma thesis focuses on the four most commonly used Bayesian techniques for spam filtering; further on, they are described in detail. We also provide a case study where one of the most commonly used algorithms is tested on a sample or real email messages and the gathered results are graded by using appropriate performance measures.

Key words: spam, Bayesian anti-spam filtering techniques, probability, performance measures, computer learning algorithms, classification.

KAZALO VSEBINE

KAZALO PONAZORITEV

KAZALO SLIK

KAZALO TABEL

KAZALO GRAFOV

1 UVOD

Z nastankom interneta leta 1969 so se začele pojavljati nove oblike komunikacije med ljudmi. Eden izmed najpogostejših in najbolj priljubljenih načinov komuniciranja v zadnjem času je nedvomno postala elektronska pošta. Njene prednosti so hitrost, učinkovitost, je pa tudi izjemno poceni in enostavna za uporabo. Zato ni čudno, da je danes na svetu več kot 3.3 milijarde elektronskih naslovov, iz katerih je dnevno poslanih in prejetih okoli 144 milijard elektronskih sporočil. Po napovedih raziskovalcev naj bi se do leta 2016 število uporabnikov elektronske pošte povečalo še za dodatnih 6 % (Radicati in Hoang 2012).

Zaradi velikega števila uporabnikov je elektronska pošta postala zanimivo področje za oglaševalce, ki so preko nje začeli oglaševati svoje izdelke. To je vodilo v pojav nezaželene elektronske pošte ali spama kot hitro rastočega glavnega problema. Njena težava je predvsem ta, da uporabnikom preprečuje oziroma omejuje normalno uporabo elektronske pošte in ovira njihovo produktivnost.

Junija 2013 je nezaželena elektronska pošta predstavljala kar 64 % vseh poslanih elektronskih sporočil (Symantec Corporation 2013). Razlog za tako velik delež se skriva v izredno nizkih stroških pošiljatelja nezaželene elektronske pošte ali »spamerja« (*ang. spammer*), saj večino stroškov nosijo lastniki strežnikov, preko katerih je nezaželena elektronska pošta poslana. Zaradi tega stroški niso vezani na količino poslanih nezaželenih elektronskih sporočil (Molan in Dečman 2005). Edini strošek, ki ga ima pošiljatelj, je cena dostopa do širokopasovne povezave. Po oceni Mednarodne telekomunikacijske zveze (International Telecommunication Union) iz leta 2004 je strošek vsakega poslanega nezaželenega elektronskega sporočila manjši kot 0.0005 USD. Nezaželena elektronska pošta je zato postala priljubljena oglaševalska pot, saj lahko oglaševalci z njo poceni in enostavno dostopajo do velike množice potencialnih kupcev. Hkrati omogoča tudi oglaševanje izdelkov, ki jih je po običajnih oglaševalskih poteh težko prodajati (npr. viagro ali spolne pripomočke).

Da bi omejili količino poslane nezaželene elektronske pošte, so se začeli pojavljati razni ukrepi. Eden izmed njih je zakonodaja. Države so zaradi ekonomskih učinkov, ki jih ima nezaželena elektronska pošta na gospodarstvo, začele omejevati njeno pošiljanje. Vendar ima ta ukrep eno omejitev – omenjeni zakoni veljajo zgolj v državi, v kateri so bili sprejeti. Ker večina nezaželene elektronske pošte izhaja iz drugih držav, kot so npr. Združene države

Amerike, Finska, Španija, Brazilija, Indija, Argentina itd. (Symantec Corporation 2013), ti ukrepi nimajo želenega učinka. Prav tako lahko omejitve, povezane s sledenjem pošiljateljev nezaželene elektronske pošte, omejujejo izvajanje teh zakonov.

Drugi, učinkovitejši ukrep se nanaša na uničenje glavnih virov nezaželene elektronske pošte. To so predvsem zlorabljeni strežniki in omrežja računalnikov pod nadzorom pošiljateljev nezaželene elektronske pošte ali tako imenovani »bootneti«. Ta ukrep je v zadnjem času uspešno vplival na trend upadanj števila poslane nezaželene elektronske pošte, saj naj bi po nekaterih ocenah iz teh omrežij izhajalo tudi do 90 % vse poslane nezaželene elektronske pošte (Porenta in drugi 2013, 52-53). Žal pa ta ukrep ni uspel preprečiti, da nezaželena elektronska pošta ne bi prispela do uporabnikov elektronskih poštnih predalov.

Rešitev za to težavo so postali filtri nezaželene elektronske pošte, ki so se v komercialne namene začeli pojavljati okoli leta 2000. Ti temeljijo na analizi vsebine in elementov elektronskega sporočila ter tako skušajo prepoznavati in ločiti nezaželeno elektronsko pošto od legitimne. Uporabniki zato prejmejo manj nezaželenih elektronskih sporočil, kar jim prihrani čas, ki ga porabijo za pregledovanje elektronske pošte, hkrati pa filtri poskrbijo za njihovo zaščito. Zaradi njihove učinkovitosti in preprostosti, ki omogoča enostavno vgraditev v poštne predale oz. odjemalce, so v kratkem postali zelo razširjeni in priljubljeni v boju proti nezaželeni elektronski pošti. Njihova dobra lastnost je tudi ta, da so se zmožni avtomatsko prilagajati raznim situacijam, saj temeljijo na algoritmih strojnega učenja (*ang. computer learning algorithms*), zaradi česar jih je težko zaobiti. Najbolj razširjeni in učinkoviti filtri nezaželene elektronske pošte danes temeljijo na Bayesovi formuli o pogojnih verjetnostih.

Diplomska naloga je sestavljena iz štirih delov. V prvem delu sem podrobneje opredelil pojem nezaželene elektronske pošte in raziskal njeno zgodovino. V drugem delu sem skozi teorijo verjetnosti prikazal potek nastanka Bayesove formule, ki predstavlja osnovo najbolj razširjenemu klasifikatorju besedil – Naivnemu Bayesovemu klasifikatorju. V nadaljevanju sem opisal postopek strojnega učenja in mere učinkovitosti klasifikatorjev. Nato sem na kratko predstavil stopnje razvrščanja in na koncu še štiri najpogostejše Bayesove metode razvrščanja nezaželene elektronske pošte.

V empiričnem delu sem na podlagi zbranih nezaželenih in legitimnih elektronskih sporočil v slovenskem jeziku, s programom SpamBayes, izvedel test primera razvrščanja nezaželene

elektronske pošte. S tem sem prikazal delovanje izbrane Bayesove metode in algoritmov strojnega učenja. Test sem izvedel na treh mejnih vrednostih ($t = 0.5$, $t = 0.9$ in $t = 0.999$), ki določajo verjetnost pri kateri je elektronsko sporočilo razvrščeno kot nezaželeno. Rezultate testa sem ocenil z ustreznimi merami uspešnosti razvrščanja in cenovno občutljivimi merami, ter v sklepu podal pomembnejše ugotovitve, do katerih sem prišel.

2 NEZAŽELENA ELEKTRONSKA POŠTA

Nezaželena elektronska pošta ali »spam« je vsako nenaročeno ali nezaželeno komercialno elektronsko sporočilo, »ki je poslano večjemu številu prejemnikov, z namenom vsiljevanja vsebine, ki se je naslovniki sami ne bi odločili prejemati« (ARNES 2012). V večini primerov takšna sporočila oglašujejo izdelke ali plačljive storitve dvomljive kvalitete, hitre zaslužke, finančne ugodnosti, erotične in pornografske vsebine itd. in so večkrat povezana z goljufijami ali prevarami. Lahko vključujejo tudi viruse, ki se razmnožujejo po elektronski pošti. V literaturi se za nezaželeno elektronsko pošto uporabljajo tudi izrazi, kot so nenaročena komercialna elektronska sporočila (*ang. unsolicited commerial e-mail*), nenaročena masovna elektronska sporočila (*ang. unsolicited bulk e-mail*) in reklamna elektronska sporočila (*ang. junk mail*).

Neodvisni francoski urad, ki se ukvarja z zaščito podatkov, je definiral nezaželeno elektronsko pošto kot »dejanje množičnega pošiljanja nenaročenih elektronskih sporočil, večinoma komercialne narave, posameznikom, s katerimi pošiljatelj ni imel nobenega prejšnjega stika, in katerega elektronski naslov se lahko najde na javnem mestu na internetu, kot npr. na novičarskih skupinah, poštnih seznamih, imenikih ali spletnih straneh« (ITU Strategy and Policy Unit 2004, 15).

Združene države Amerike so leta 2003 v zakonu CAN-SPAM (Controlling the Assault of Non-Solicited Pornography And Marketing Act of 2003) nezaželeno elektronsko pošto opredelile kot elektronska sporočila, katerih glavni namen je oglaševanje ali promocija komercialnih izdelkov ali storitev poleg vsebine na spletnem mestu. Z zakonom so prav tako opisali zahteve, ki jih morajo pošiljatelji reklamnih elektronskih sporočil izpolniti, ter kazni, ki jih lahko doletijo v primeru kršitve omenjenega zakona (ITU Strategy and Policy Unit 2004, 15).

Evropska unija je leta 2002 z Direktivo 2002/58/ES o obdelavi osebnih podatkov in varstvu zasebnosti na področju elektronskih komunikacij uporabila tehnološko nevtralno definicijo nezaželene elektronske pošte, kjer je uporabila termina »nenaročene komercialne komunikacije« (*ang. unsolicited commercial communications*) in »elektronska sporočila za namen neposrednega trženja« (*ang. electronic mail for the purpose of direct marketing*).

Elektronsko sporočilo namreč pomeni »vsako besedno, govorno, zvočno ali slikovno sporočilo, poslano prek javnega komunikacijskega omrežja, ki se lahko shrani v omrežju ali v prejemnikovi terminalski opremi, dokler ga prejemnik ne prevzame« (Direktiva 2002/58/ES, člen 2). S tem so želeli zajeti nezaželena elektronska sporočila, poslana tako preko elektronskih predalov, kot SMS sporočil, takojšnjega sporočanja (*ang. instant messaging*) itd. Pri tem so nezaželena elektronska sporočila definirali kot vsa elektronska sporočila, ki so poslana z namenom neposrednega trženja, torej tudi tista, pri katerih primarna korist ni le finančna. Po definiciji Evropske komisije bi med neželena elektronska sporočila lahko uvrstili tudi tista, ki oglašujejo politično ali religiozno miselnost.

2.1 Zgodovina nezaželene elektronske pošte

Prvo nezaželeno elektronsko sporočilo je bilo poslano leta 1978 po omrežju ARPANET, ameriškem vojaškem predhodniku interneta. Po omrežju ga je poslalo računalniško podjetje Digital Equipment Corporation, znano tudi pod okrajšavo DEC, za svoj nov računalnik DEC-20. Sporočilo so poslali na vse naslove na zahodni ameriški obali. Za to dejanje so bili kasneje kaznovani zaradi kršitve pravil uporabe ARPANET-a, drugim uporabnikom omrežja pa so prepovedali takšna dejanja.

Po Schwartz in Ganfinkel (1998) so bili pri razvoju nezaželene elektronske pošte pomembni predvsem trije mejniki. Prvi je bil leta 1994, ko sta odvetnika Laurence A. Canter in Martha S. Siegel poslala oglas za plačljiv pravni nasvet vsem imigrantom, ki so v Združenih državah Amerike zaprosili za delovno zeleno karto. Problem oglasa je bil, da je bila pridobitev delovne karta brezplačna, ter da so elektronsko pošto prejeli vsi, ne glede na izraženo željo po njenem prejemu. Ker se jima je hkrati pritožilo več tisoč uporabnikov, so ti onemogočili delovanje njunega internetnega ponudnika, zaradi česar jima je ta preprečil dostop do interneta.

Drugi mejnik je bila zamisel o »spam marketingu« Jeffa Slatona oz. »Spam kinga«, leta 1995. Slaton je zbiral elektronske naslove posameznikov, ki so mu služili kot osnova za pošiljanje komercialnih sporočil. Nato je na te naslove poslal elektronsko sporočilo, v katerem je prodajal načrte za atomsko bombo in s tem hitro obogatel. Ustanovil je svoje podjetje Unix/Eunuchs Etc. in začel svoje usluge ponujati tudi drugim podjetjem. Postal je pionir

»spam marketinga«. Hkrati je uvedel številne novosti, ki jih pošiljatelji nezaželene elektronske pošte uporabljajo še danes: neresnični elektronski naslovi pošiljateljev, v elektronsko sporočilo je vključil izmišljeno telefonsko številko telefonske tajnice, izkoriščal je druge strežnike za pošiljanje nezaželene elektronske pošte in ponujal lažno možnost odjave (»opt-out«). Po robu so se mu postavili »branilci interneta«, ki so mu hoteli onemogočiti pošiljanje nezaželenih elektronskih sporočil. Zato so sestavili seznam, v katerega so vpisovali njegove podatke, da bi jih filtri elektronske pošte lažje prepoznali in mu tako skušali onemogočiti masovno pošiljanje.

Zadnji mejnik je bil leta 1996, ko je Sanford Wallace iz svoje domene začel pošiljati več deset tisoč izvodov elektronskih sporočila hkrati. Elektronske naslove mu je uspelo pridobiti od največjega ameriškega ponudnika storitev elektronske pošte AOL (American Online). Dnevno je tako vsakemu uporabniku poslal približno pet nezaželenih elektronskih sporočil. Zaradi številnih pritožb uporabnikov je bil AOL s filtrirnim sistemom prisiljen svoje uporabnike zaščititi, da jih Wallaceova nezaželena sporočila niso dosegla.

3 RAZVRŠČANJE NEZAŽELENE ELEKTRONSKE POŠTE

3.1 Teorija verjetnosti

Teorija verjetnosti predstavlja osnovo Bayesovem klasifikatorju, ki ga uporabljajo filtri nezaželene elektronske pošte. Verjetnost lahko interpretiramo kot predviden delež poskusov, pri katerih se zgodi določen dogodek.

Poskus je dejanje, ki ga opravimo v natančno določenih pogojih, pri katerih opazujemo enega ali več dogodkov. Dogodek je pojav, ki ne spada v množico skupaj nastopajočih dejstev in se lahko v določenem poskusu zgodi ali ne. Poznamo tri vrste dogodkov:

- gotov dogodek (G), ki se zgodi ob vsaki ponovitvi poskusa,
- nemogoč dogodek (N), ki se ne zgodi nikoli, in
- slučajen dogodek, ki se včasih zgodi, včasih ne (Jurišič 2010).

Verjetnost dogodka (Možina 2007) je število, ki je prirejeno dogodku glede na njegovo pogostost v več ponovitvah poskusa. Gre za poskus, kjer nastopa dogodek A, ki ga ponovimo n-krat. S k označimo, kolikokrat se je dogodek A zgodil, čemur pravimo frekvenca dogodka A. Relativno frekvenco (pogostost) dogodka izračunamo po naslednji formuli:

$$f(A) = \frac{k}{n}.$$

Verjetnost dogodka se označuje z veliko črko P. Za določitev verjetnostne funkcije morajo biti izpolnjeni trije minimalni pogoji, ki jim pravimo aksiomi Kolmogorova (Škulj 2009):

1. **Nenegativnost:** verjetnost dogodka A je vedno nenegativno realno število:

$$P(A) \geq 0.$$

2. **Normiranost:** verjetnost gotovega dogodka (G) znaša 1 in verjetnost nemogočega dogodka (N) 0:

$$P(G) = 1, P(N) = 0.$$

3. **Aditivnost:** če sta dogodka A in B paroma nezdružljiva (se ne moreta zgoditi hkrati), je relativna frekvenca vsote dogodkov enaka vsoti relativnih frekvenc:

$$P(A \cup B) = P(A) + P(B).$$

Iz aksiomov Komogorova lahko izpeljemo ostale lastnosti verjetnostne funkcije (Hladnik 2002, 12):

1. Če je dogodek A način dogodka B, potem velja:

$$P(A) \leq P(B).$$

2. Verjetnost prazne množice dogodkov je 0:

$$P(N) = 0.$$

3. Verjetnost dogodka A vedno zavzame vrednost med 0 in 1:

$$0 \leq P(A) \leq 1.$$

4. Verjetnost, da se zgodi dogodek A ali B, je enaka vsoti verjetnosti dogodka A in dogodka B minus verjetnosti dogodka, da se hkrati zgodi A in B:

$$A(A \cup B) = P(A) + P(B) - P(A \cap B).$$

5. Verjetnost, da se bo zgodil dogodek B, je enaka razliki med 1 in verjetnostjo da se bo zgodil njemu nasproten dogodek A:

$$P(B) = 1 - P(A).$$

Pogojna verjetnost je verjetnost, da se zgodi dogodek A, ob pogoju, da se je zgodil nek drug dogodek B, kar označimo s $P(A|B)$. Temu pravimo produkt dogodkov. Za dva dogodka pogojno verjetnost izračunamo po naslednji formuli:

$$P(A|B) = \frac{P(AB)}{P(B)}.$$

Pri tem pa mora veljati $P(B) > 0$. Z oznako $P(AB)$ je označeno hkratno ponavljanje dogodka.

Velja tudi:

$$P(B|A) = \frac{P(BA)}{P(A)}.$$

Dogodka A in B sta neodvisna če velja:

$$P(A|B) = P(A) \text{ in } P(B|A) = P(B),$$

in nezdružljiva, če velja:

$$P(A|B) = 0 \quad in \quad P(B|A) = 0.$$

Poskusi lahko potekajo tudi v več fazah, čemur pravimo relejni poskusi. V primeru dvofaznega poskusa, imamo v prvi fazi popoln sistem dogodkov $H_1, \ldots, H_n$, ki jim pravimo tudi hipoteze. Od tega, kateri izmed dohodkov H_n se je pripetil v prvi fazi, so odvisni pogoji druge faze poskusa, v kateri opazujemo dogodek A.

Od tod sledi formula za popolno verjetnost, ki nam pove, kako izračunamo brezpogojno verjetnost dogodka A:

$$P(A) = P(H_1)P(A|H_1) + \cdots + P(H_n)P(A|H_n) =$$
$$= \sum_{i=1}^{n} P(H_i) \cdot P(A|H_i).$$

Če nas zanimajo pogojne verjetnosti hipotez glede na opaženo drugo fazo poskusa $P(H_i|A)$ (zgodil se je dogodek A), velja po definiciji pogojne verjetnosti:

$$P(H_i|A) = \frac{P(A|H_i) \cdot P(H_i)}{P(A)}.$$

Ko vstavimo formulo o popolni verjetnosti v imenovalec, dobimo **Bayesovo formulo**, ki predstavlja osnovo vsem modernim sistemom za verjetnostno sklepanje:

$$P(H_i|A) = \frac{P(A|H_i) \cdot P(H_i)}{\sum_{i=1}^{n} P(H_i) \cdot P(A|H_i)}.$$

3.2 Strojno učenje

»Strojno učenje je področje umetne inteligence, ki se ukvarja z razvojem tehnik, ki omogočajo računalnikom oz. strojem, da se lahko učijo. Strojno učenje je v bistvu metoda za kreiranje računalniških programov na podlagi podatkov (vzorcev)« (Polanec 2006). Močno je povezano s statistiko, ki se prav tako ukvarja s podatki, vendar se v nasprotju z njo strojno učenje bolj ukvarja z samimi algoritmi in računskimi operacijami.

Ena izmed najbolj uporabljenih metod strojnega učenja je klasifikacija ali razvrščanje besedil oz. tekstovnih dokumentov (*ang. text categorization*) v mape, med katere spada tudi razvrščanje elektronske pošte. Tako matematični algoritmi na podlagi vsebine in elementov elektronskega poročila določijo, ali je elektronsko sporočilo nezaželeno ali ne.

Najbolj pogosto uporabljene metode razvrščanja besedil so: Naivni Bayesov klasifikator, Bayesove verjetnostne mreže (*ang. Bayesian probability networks*), metoda K najbližjih sosedov (*ang. K-nearest neighbor*), metoda podpornih vektorjev (*ang. support vector*

machine), logistična regresija, odločitvena drevesa, odločitvena pravila, linearna diskriminantna funkcija in usmerjene večnivojske nevronske mreže.

3.3 Mere uspešnosti razvrščanja (ocena stroškov)

Pri ocenjevanju uspešnosti razvrščanja nezaželene elektronske pošte je pomemben vidik, na katerega je potrebno opozoriti, asimetrija stroškov in napačno porazdeljene nezaželene elektronske pošte. Nezaželena elektronska sporočila, razvrščena med legitimna (*ang. false negative*; FN) predstavljajo relativno majhen problem, saj jih lahko uporabnik enostavno odstrani ali označi kot nezaželena. Veliko večji problem predstavljajo legitimna sporočila, ki so razvrščena med nezaželena (*ang. false pozitive*; FP), zaradi katerih se lahko izgubijo pomembne informacije, še posebej, če se nezaželena sporočila avtomatsko izbrišejo. Pojav napačno razvrščenih legitimnih sporočil je zato nesprejemljiv in zmanjšuje zaupanje uporabnikov v filter nezaželene elektronske pošte. V tem primeru je bolje, da dovolimo nekaj napačno razvrščenih nezaželenih elektronskih sporočil kot legitimnih.

Pri uspešnosti klasifikatorja torej ne moremo govoriti zgolj o relativnem številu pravilno razvrščenih sporočil (klasifikacijska natančnost), saj to predpostavlja enake stroške napačne razvrstitve legitimnih in nezaželenih elektronskih sporočil. Zato je v realni situaciji, kjer vedno obstaja verjetnost nepravilne razvrstitve legitimnih elektronskih sporočil, pomembno sprejeti kompromis med željami uporabnikov in kazalci uspešnosti.

Binarni klasifikatorji razvrščajo elektronsko pošto v eno od dveh skupin: pozitivno (spam) ali negativno (legitimna). Da lahko to počnejo, jih je najprej potrebno naučiti, katera sporočila so pozitivna in negativna. To storimo v stopnji učenja na podlagi testnih primerov (elektronskih sporočil), za katera vemo, kateri skupini pripadajo. Tej stopnji sledi stopnja razvrščanja, kjer preverimo, ali je klasifikator sposoben pravilno razvrstiti prejeto elektronsko pošto. Stopnje razvrščanja so podrobneje razložene v točki 3.5. tega poglavja.

3.3.1. Kontingenčna tabela

Pri razvrščanju nezaželene elektronske pošte lahko klasifikator stori tako imenovano klasifikacijsko napako, kar pomeni, da elektronskih sporočil ne razvrsti v pravilne skupine. To

najlažje prikažemo z matriko razvrstitev (*ang. confusion matrix*) ali kontingenčno tabelo (Tabela 3.1), v kateri prikažemo število pravilno in napačno razvrščenih nezaželenih ter legitimnih elektronskih sporočil.

Tabela 0.1: Kontingenčna tabela

		dejanski razred +	dejanski razred -
razvrščen razred	+	TP	FP
	-	FN	TN

Pri binarnemu razvrščanju obstajajo štirje možni izidi razvrščanja elektronskih sporočil:

- pravilna pozitivna (*ang. true pozitive*; TP) so pravilno razvrščena nezaželena elektronska sporočila,
- napačna pozitivna (*ang. false negative*; FN) so napačno razvrščena nezaželena elektronska sporočila,
- pravilna negativna (*ang. true negative*; TN) so pravilno razvrščena legitimna elektronska sporočila,
- napačna negativna (*ang. false pozitive*; FP) so napačno razvrščena legitimna elektronska sporočila.

Pravilne razvrstitve klasifikatorja ležijo na glavni diagonali kontingenčne tabele in napačne na stranski.

Ker klasifikatorji kot rezultat vrnejo številsko in ne binarno vrednost (0, 1), je potrebno določiti še mejo (*ang. threshold*), pri kateri se neko elektronsko sporočilo razvrsti kot nezaželeno (0) ali legitimno (1). Če je vrednost večja od določene meje, je takšno elektronsko sporočilo nezaželeno, in če je manjša, je legitimno. Če določimo višjo mejo, s čimer povečamo občutljivost, je verjetnost napačne razvrstitve nezaželenih elektronskih sporočil manjša. In če določimo nižjo mejo ter tako zmanjšamo občutljivost, povečamo verjetnost napačno razvrščenih legitimnih elektronskih sporočil.

3.3.2. Mere uspešnosti razvrščanja

Iz kontingenčne tabele razvrstitev primerov je moč izračunati mere uspešnosti razvrščanja. Ena izmed njih je TPR (*ang. true pozitive rate*) oz. delež pravilno razvrščenih pozitivnih primerov med vsemi pozitivnimi primeri. Tej meri drugače pravimo tudi priklic ali občutljivost (*ang. recall, sensitivity*). Izračunamo jo po naslednji formuli:

$$Priklic = Občutljivost = TPR = \frac{TP}{TP + FN}.$$

Mera FNR (*ang. false negative rate*) je nasprotje priklica in kot takšna predstavlja delež napačno razvrščenih pozitivnih primerov med vsemi pozitivnimi primeri:

$$FNR = \frac{FN}{FN + TP} = 1 - TPR.$$

FPR (*ang. false pozitive rate*) nam pove delež napačno razvrščenih negativnih primerov med vsemi negativnimi primeri:

$$FPR = \frac{FP}{FP + TN}.$$

Delež pravilno razvrščenih negativnih primerov med vsemi negativnimi primeri TNR (*ang. true negative rates*) ali specifičnost se izračuna po formuli:

$$Specifičnost = TNR = \frac{TN}{TN + FP} = 1 \quad FPR.$$

Preciznost (*ang. precision*) nam pove delež pravilno razvrščenih pozitivnih primerov med vsemi napovedanimi pozitivnimi primeri:

$$Preciznost = \frac{TP}{TP + FP}.$$

Če združimo priklic in preciznost, dobimo mero F *(ang. F-measure*), kjer lahko po želji utežimo priklic ($\beta > 1$) ali preciznost ($\beta < 1$). Pri razvrščanju nezaželene elektronske pošte se praviloma močno uteži preciznost:

$$F = (1 + \beta^2) \frac{priklic \cdot preciznost}{\beta^2 \cdot priklic + preciznost}.$$

Klasifikacijska točnost ACC (*ang. accuracy*) predstavlja delež vseh pravilno razvrščenih primerov med vsemi primeri. Torej delež nezaželenih elektronskih sporočil pravilno razvrščenih med nezaželena elektronska sporočila in legitimnih elektronskih sporočil pravilno razvrščenih med legitimna, med vsemi prejetimi elektronskimi sporočili. Točnost je pogosto uporabljena mera v primeru ocenjevanja uspešnosti razvrščanja. Njeno nasprotje je stopnja napake ERR (*ang. error rate*), ki jo definiramo kot delež vseh napačno razvrščenih primerov med vsemi primeri.

$$Točnost = ACC = \frac{TN + TP}{TN + FP + FN + TP}$$

$$Stopnja\ napake = ERR = 1 - Točnost = \frac{FN + FP}{TN + FP + FN + TP}$$

3.3.3. Cenovno občutljive mere (ang. cost-sensitive evaluation measures)

Klasifikacijska točnost in stopnja napake sta najpogosteje uporabljeni meri za ocenjevanje uspešnosti razvrščanja. Problem omejenih mer je, da nista cenovno občutljivi, saj ne predpostavljata višjih stroškov napačno razvrščenega legitimnega elektronskega sporočila v primerjavi z napačno razvrščenim nezaželenim elektronskim sporočilom. Da bi meri postali cenovno občutljivi, je potrebno vsako legitimno elektronsko sporočilo šteti kot λ elektronskih sporočil. Če je torej legitimno elektronsko sporočilo napačno razvrščeno, šteje kot λ napak, če pa je pravilno, pa kot λ uspehov. Upoštevanje te predpostavke nas pripelje do formule za uteženo točnost (W_{ACC}) in uteženo stopnjo napake ($W_{ERR} = 1 - W_{ACC}$):

$$W_{ACC} = \frac{\lambda TN + TP}{\lambda TN + FN + TP + \lambda FP}, \qquad W_{ERR} = 1 - W_{ACC} = \frac{FN + \lambda FP}{\lambda TN + FN + TP + \lambda FP}.$$

Da bi bolje razumeli uspešnost razvrščanja, si poglejmo enostaven »osnovni« pristop Potamiasa in drugih (2000), ki predpostavlja, da ne uporabljamo filtra nezaželene elektronske pošte oz. ta ni aktiven. Tako se izognemo napačni interpretaciji pogosto visokih vrednosti točnosti in nizkih vrednosti stopnje napake. V osnovnem pristopu vsa nezaželena elektronska sporočila pomotoma vedno zaobidejo filter in legitimna elektronska sporočila praviloma niso nikoli zablokirana. Potem sta uteženа točnost in stopnja napake v osnovi:

$$W_{ACC}{}^{b} = \frac{\lambda TN + \lambda FP}{\lambda TN + FN + TP + \lambda FP}, \qquad W_{ERR}{}^{b} = \frac{FN + TP}{\lambda TN + FN + TP + \lambda FP}.$$

Iz formule za uteženo stopnjo napake in utežene osnovne stopnje napake nastane formula za skupno razmerje med stroški TCR (*ang. total cost ratio*):

$$TCR = \frac{W_{ERR}{}^{b}}{W_{ERR}} = \frac{FN + TP}{FN + \lambda FP}.$$

Višja kot je vrednost skupnega razmerja med stroški, boljša je uspešnost razvrščanja. Če znaša TCR < 1, je osnovi pristop boljša izbira, saj predvideva odsotnost filtra nezaželene elektronske pošte, ki daje boljše rezultate kot njegova uporaba. Če so stroški sorazmerni s porabljenim časom, potem skupno razmerje med stroški meri čas, porabljen za ročno brisanje vseh nezaželenih elektronskih sporočil ob odsotnosti filtra ($TP + FP$), v primerjavi s časom, porabljenim za ročno brisanje vseh nezaželenih elektronskih sporočil, ki so zaobšla filter (FN), ter časom porabljenim za obnovitev vseh napačno razvrščenih legitimnih sporočil (λFP).

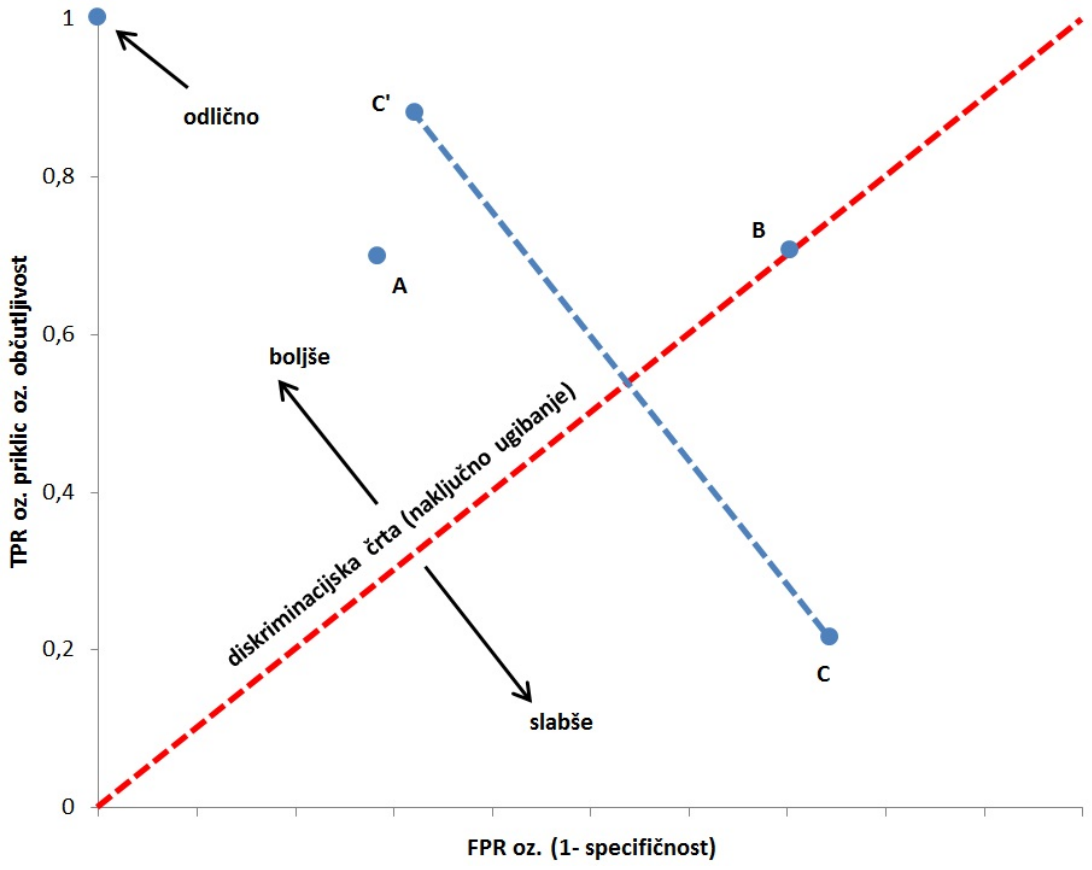
3.3.4. ROC krivulja

Na podlagi deleža pravilno razvrščenih pozitivnih primerov (občutljivost) v odvisnosti od deleža napačno razvrščenih negativnih primerov (specifičnost-1) lahko izrišemo ROC (*ang. Receiver Operating Characteristic*) krivuljo, ki predstavlja razmerje med omenjenima dvema merama. Ta je v zadnjih letih zaradi svojega enostavnega in preglednega prikaza postala vse bolj priljubljena za ocenjevanje učinkovitosti algoritmov strojnega učenja.

Vsaka točka na krivulji grafa predstavlja presečišče med specifičnostjo in občutljivostjo, ki pripada določenemu odločitvenemu pragu. Bolj ko se krivulja približuje zgornjemu levemu kotu (točki 1,0), večja je natančnost določenega klasifikatorja. Če bi imeli popoln klasifikator, bi njegova krivulja potekala navpično od točke 0,0 do točke 0,1 in vodoravno do točke 1,1. Vendar takšnih rezultatov na resničnih podatkih ni mogoče dobiti. Krivulja, ki se ne obnaša nič bolje od naključnega ugibanja, poteka od točke 0,0 do točke 1,1 (diskriminacijska črta). Večina krivulj resničnih testov leži med tema dvema ekstremoma, kar pomeni, da je takšen klasifikator uporaben, saj je njegova učinkovitost večja od naključnega ugibanja.

ROC krivulje so same po sebi izredno uporabne pri ocenjevanju uspešnosti klasifikatorjev, vendar z njimi ne moremo natančno povzeti, kakšna je uspešnost nekega klasifikatorja. ROC indeks, s katerim lahko to storimo, se imenuje AUC (*ang. area under the ROC curve*) ali področje pod ROC krivuljo. Statistično vrednost AUC interpretiramo kot verjetnost, da bo imelo naključno izbrano nezaželeno elektronsko sporočilo večjo vrednost kot naključno izbrano legitimno elektronsko sporočilo (Fawcett 2006). AUC nam omogoča tudi primerjavo dveh ali več klasifikatorjev med sabo, tako da med njimi primerjamo AUC vrednosti.

Graf 0.1: ROC graf



3.4 Naivni Bayesov klasifikator

Filtriranje nezaželene elektronske pošte je ena izmed vej razvrščanja besedil. Eden izmed najbolj razširjenih klasifikatorjev, ki se uporablja za razvrščanje nezaželene elektronske pošte, je Naivni Bayesov klasifikator. Njegovo uporabo za namen filtriranja so v članku z naslovom A Bayesian Approach to Filtering Junk E-mail (1998) prvič podrobneje raziskali avtorji M. Sahami, S. Dumais, D. Heckerman in E. Horvitz.

Problema razvrščanja se niso lotili zgolj na podlagi analize vsebine elektronske pošte, ampak so ugotovili, da igrajo pri razvrščanju pomembno vlogo tudi ostali elementi, ki znatno povečajo uspešnost klasifikatorja. To so zlasti določene fraze (npr. »private message«, »earn money«, »lose weight«), pretirana uporaba ločil (npr. »!!!!!!!!!!!!!!!« »…«), veliko število nebesednih značilnosti, kot je tip domene pošiljatelja (npr. .edu, .com, .si), čas poslanega sporočila (večina nezaželene elektronske pošte je poslane ponoči), prisotnost priponke (večina nezaželene elektronske pošte ne vsebuje priponke) in ali je bilo sporočilo poslano posamezniku ali preko »mailing« liste.

V Naivni Bayesov klasifikator so tako enostavno vključili dodatne funkcije za razvrščanje nezaželene elektronske pošte, in sicer s tem, da so dodali nove spremenljivke, ki so označevale prisotnost ali odsotnost omenjenih elementov v vektor vsakega elektronskega sporočila. S tem so dosegli, da je lahko pri razvrščanju elektronskih sporočil hkrati enakomerno vključenih več vrst dokazov, ne da bi bilo potrebno spremeniti algoritme strojnega učenja.

Naivni Bayesov klasifikator temelji na Bayesovem izreku. Vsako elektronsko sporočilo je predstavljeno z vektorjem $\vec{x} = [x_1, x_2, \ldots, x_n]$, kjer so $x_1, x_2, \ldots, x_n$ vrednosti atributov $X_1, \ldots, X_n$. Atributi so binarni, kar pomeni, da $X_i = 1$ predstavlja prisotnost določenih značilnosti in $X_i = 0$ njihovo odsotnost. V primeru razvrščanja nezaželene elektronske pošte atributi predstavljajo posamezne besede (npr. »free«) in njihovo prisotnost/odsotnost v določeni elektronski pošti.

Iz Bayesovega izreka in izreka o popolni verjetnosti pri danem vektorju $\vec{x} = [x_1, \ldots, x_n]$ dokumenta d je verjetnost, da d pripada kategoriji c:

$$P(C = c|\vec{X} = \vec{x}) = \frac{P(C = c) \cdot P(\vec{X} = \vec{x}|C = c)}{\sum_{k \in \{spam, legitimno\}} P(C = k) \cdot P(\vec{X} = \vec{x}|C = k)}.$$

Ker je verjetnosti $P(\vec{x}|C)$ skoraj nemogoče neposredno oceniti (saj obstaja preveč verjetnih vrednosti $\vec{x}$ in problem razpršenosti podatkov), naivni Bayesov klasifikator predpostavi, da so atributi $X_1, \ldots, X_N$ pogojno neodvisni glede na dano kategorijo C. Zato lahko $P(C = c|\vec{X} = \vec{x})$ zapišemo kot:

$$P(C = c|\vec{X} = \vec{x}) = \frac{P(C = c) \cdot \prod_{i=1}^{n} P(X_i = x_i|C = c)}{\sum_{k \in \{spam, legitimno\}} P(C = k) \cdot \prod_{i=1}^{n} P(X_i = x_i|C = k)},$$

kjer lahko $P(X_i|C)$ in $P(C)$ brez težav ocenimo kot relativne frekvence v stopnji učenja.

Ker je napaka napačno razvrščenih legitimnih sporočil (FP) veliko večja kot napačno razvrščenih nezaželenih elektronskih sporočil (FN), predpostavimo še, da je strošek napačno razvrščenega legitimnega elektronskega sporočila $\lambda - krat$ večji kot strošek napačno razvrščenega nezaželenega elektronskega sporočila. Tako lahko razvrstimo elektronsko sporočilo kot nezaželeno, če velja:

$$\frac{P(C = spam|\vec{X} = \vec{x})}{P(C = legitimno|\vec{X} = \vec{x})} > \lambda.$$

V našem primeru je $P(C = spam|\vec{X} = \vec{x}) = 1 - P(C = legitimno|\vec{X} = \vec{x})$, kar vodi do preoblikovanja formule:

$$P(C = spam|\vec{X} = \vec{x}) > t, \text{ kjer je } t = \frac{\lambda}{1+\lambda}, \lambda = \frac{t}{1-t}.$$

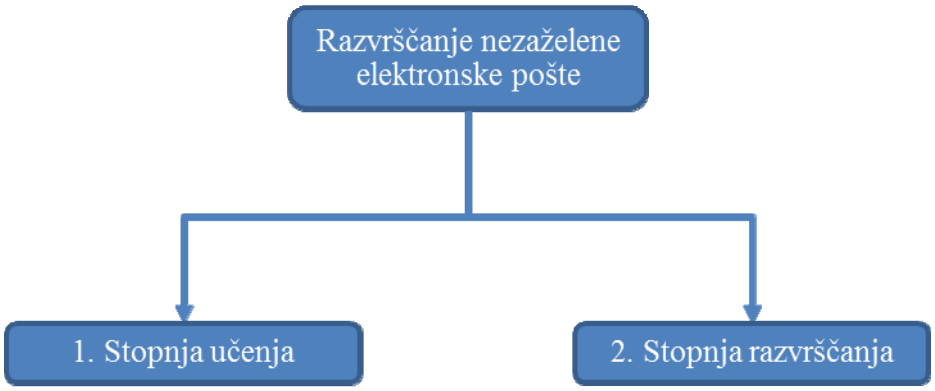
Sahami in drugi (1998) so nastavili vrednost *t* na 0.999 ($\lambda = 999$), kar pomeni, da je eno napačno razvrščeno legitimno elektronsko sporočilo enako 999 napačno razvrščenim nezaželenim elektronskim sporočilom. Visoko vrednost λ so opravičili s tem, da bi večina uporabnikov štela izgubo legitimnega elektronskega sporočila za nesprejemljivo, še posebej v

primeru, ko je filter nastavljen tako, da vsa nezaželena elektronska sporočila avtomatsko izbriše. Seveda obstajajo tudi druge alternative, pri katerih so sprejemljive nižje vrednosti λ. Tako lahko npr. filter nastavimo tako, da pošiljatelju napačno razvrščenega sporočila pošlje obvestilo o zavrnjenem sporočilu, kjer mu predlaga ponovno pošiljanje na drug nefiltriran osebni elektronski naslov. V tem primeru lahko uporabimo vrednost $\lambda = 9$ ($t = 0.9$), kjer je napačno razvrščeno legitimno elektronsko sporočilo 9-krat slabše kot napačno razvrščeno nezaželeno elektronsko sporočilo. Slabost takšne nastavitve filtra nezaželene elektronske pošte je, da od pošiljatelja v primeru zavrnitve elektronske pošte zahteva več dodatnega dela. Če prejemnik prejme vsa elektronska sporočila in filter med njimi posebej označi nezaželena, bi bilo smiselno nastaviti vrednost na $\lambda = 1$ ($t = 0.5$), kjer so napačno razvrščena legitimna in nezaželena elektronska enakovredna. Ta vrednost se zdi primerna tudi v primeru, ko filter nezaželeno elektronsko pošto razvrsti v posebno mapo, namenjeno nezaželenim elektronskim sporočilom, kjer jih lahko nato uporabnik pregleduje in napačno razvrščena legitimna elektronska sporočila ročno prestavi v drugo mapo.

3.5 Stopnje razvrščanja nezaželene elektronske pošte

Razvrščanje elektronske pošte je proces, sestavljen iz dveh stopenj: stopnje učenja in stopnje razvrščanja (Rajput in Toshniwal 2012).

Graf 0.2: Stopnje razvrščanja nezaželene elektronske pošte



Stopnja učenja je sestavljena iz petih podstopenj:

- Zbiranje poznanih elektronskih sporočil, za katera vemo, v katero skupino spadajo (nezaželena in legitimna). Priporočljivo je, da zberemo elektronska sporočila iz več različnih virov, saj je tako učenje bolj uspešno.
- Priprava elektronske pošte, tako da iz besedila odstranimo veznike, stavčne člene itd., saj te besede nimajo uporabne vrednosti pri razvrščanju elektronske pošte. Priporočljivo je, da sestavimo in dopolnjujemo seznam pošiljateljev, v katerem shranjujemo informacije pošiljateljev nezaželene in legitimne elektronske pošte.
- Ustvarjanje zbirne tabele (*ang. hash map*) besed in števila ponavljanja določenih besed v elektronskih sporočilih. Če beseda že obstaja v zbirni tabeli, potem se njeno število ob ponovnem pojavu poveča, v nasprotnem primeru se beseda vanjo doda. Prav tako je potrebno upoštevati različne oblike besed, npr. ednino - množino in glagolske oblike.
- Računanje verjetnosti pojava določene besede v nezaželeni ali legitimni elektronski pošti, na podlagi katere lahko nato izračunamo verjetnost, da se določena beseda pojavi v nezaželeni elektronski pošti:

$$Sp = \frac{f_1}{f_1 + f_2}.$$

  f_1 predstavlja frekvenco, da se beseda pojavi v nezaželeni elektronski pošti, in f_2 frekvenco, da se pojavi v legitimni.
- Urejanje besed po vrstnem redu verjetnosti.

Stopnji učenja sledi stopnja razvrščanja, ki je prav tako sestavljena iz petih podstopenj:

- Priprava sklopa elektronskih sporočil, ki bo služil testiranju filtra nezaželene elektronske pošte.
- Priprava elektronske pošte (npr. odstranimo veznike, stavčne člene itd.).
- Priprava seznama besed v zbirni tabeli z zelo visoko ali nizko verjetnostjo pojava v nezaželeni elektronski pošti.
- Iskanje splošne verjetnosti pojava nezaželene elektronske pošte in preverjanje pošiljatelja s seznama pošiljatelja.
- Razvrščanje elektronskih sporočil.

3.6 Metode razvrščanja nezaželene elektronske pošte

Ko Naivni Bayesov filter nezaželene elektronske pošte naučimo razlikovati med nezaželeno in legitimno elektronsko pošto, je ta pripravljen za razvrščanje novo prejetih elektronskih sporočil. Obstaja več metod, na podlagi katerih lahko to storijo. V nadaljevanju sem predstavil štiri najpogostejše, ki so jih v članku z naslovom An Evaluation of Naive Bayesian Anti-Spam Filtering Techniques (2007) raziskali avtorji Deshpande in drugi.

3.6.1 Metoda uporabe vseh besed v postopku razvrščanja

Ta metoda pri razvrščanju uporabi vse besede v besedilu prejetega elektronskega sporočila. Ker vsaka beseda predstavlja verjetnost, da je lahko elektronsko sporočilo nezaželeno, se vse besede uporabijo pri računanju skupne verjetnosti, ki dodeli končno oceno elektronski pošti. V primeru pojava nove besede, ki je še ni v bazi besed, se elektronskemu sporočilu dodeli verjetnost 0.4. Ta predpostavka je vgrajena v Naivni Bayesov klasifikator in se je pri razvrščanju izkazala za zelo uspešno. Kaže na to, da vsako novo besedo obravnavamo kot del legitimne elektronske pošte in ne kot del nezaželene. To nakazuje na pozitiven pristop filtrov nezaželene elektronske pošte, saj je strošek napačno razvrščenih legitimnih elektronskih sporočil (FP) veliko večji kot napačno razvrščenih nezaželenih elektronskih sporočil (FN).

Metodo je smiselno uporabiti v fazi učenja, saj lahko na njeni podlagi zgradimo bazo besed, ki jo uporabimo za razvrščanje. Logično je, da isto metodo uporabimo tudi v fazi razvrščanja. Pri tem je potrebno poudariti, da je faza razvrščanja kritična zaradi visokih stroškov napačno razvrščenih legitimnih elektronskih sporočil (FP) v primerjavi s fazo učenja, kjer točno vemo, ali je neko elektronsko sporočilo nezaželeno ali ne.

Slabost te metode so nezaželena elektronska sporočila, ki vsebujejo veliko število dobrih besed (dolga nezaželena elektronska sporočila), saj obstaja velika verjetnost, da bodo razvrščena kot legitimna. Vendar nezaželena elektronska sporočila praviloma ne vsebujejo veliko besedila. Prav tako obstaja majhna verjetnost, da bodo prejemniki takšnih nezaželenih elektronskih sporočil dejansko prebrali besedilo iz neznanega naslova. To slabost opisane metode so hitro izkoristili pošiljatelji nezaželene elektronske pošte. Začeli so pošiljati elektronska sporočila, v katerih so uporabili veliko število dobrih besed, vendar so jih zakrili tako, da bralcu niso bile vidne (npr. besede na beli podlagi so obarvali z belo barvo).

3.6.2 Metoda uporabe fiksnega števila besed v postopku razvrščanja

Metoda uporabe fiksnega števila besed upošteva točno določeno število besed elektronskega sporočila, na podlagi katerih izračuna končno oceno. Število uporabljenih besed se praviloma giblje med 15 in 25. Zanje se predvideva, da so najboljši pokazatelji, ali je neko elektronsko sporočilo nezaželeno ali legitimno. Pri takih besedah se verjetnost najbolj odmika od vrednosti 0.5. Na podlagi skupne verjetnosti teh besed se dodeli končna ocena prejetega elektronskega sporočila.

Pri razvrščanju so tako uporabljene le najbolj učinkovite besede. Metoda je neposredno usmerjena na besede, ki jih v večini primerov najdemo v nezaželenih ali legitimnih elektronskih sporočilih. Ocena razvrščanja se večinoma giblje okoli 1, če je elektronsko sporočilo spam, in 0, če je legitimno. Na ta način metoda blaži dvom razvrščanja elektronske pošte, kjer se ocena giblje blizu 0.5.

To metodo so predlagali Sahami in drugi (1998), ko so izračunali njeno učinkovitost s pomočjo matematične formule vzajemne informacije (*ang. mutual information*). Priporočljivo je, da se ista beseda pri računanju končne ocene uporabi samo enkrat. S tem dosežemo nepristransko odločitev filtra, saj bi ga v nasprotnem primeru pri odločitvah motile besede, ki se v besedilu elektronske pošte pojavijo večkrat. Odločitev o številu besed, ki bodo vključene v razvrščanje, temelji predvsem na učinkovitosti filtra nezaželene elektronske pošte osebnih elektronskih sporočil.

Prednost te metode je, da se pri njej izognemo problemu napačno razvrščenih legitimnih elektronskih sporočil (FP), saj lahko mejno vrednost dvignemo od 0.5 tudi do 0.9. Prav tako je ta metoda v primerjavi z drugimi metodami v primeru prejemanja velike količine elektronskih sporočil hitrejša.

3.6.3 Metoda prilagajanja meje standardnega odklona (ang. Standard Deviation Threshold Filter)

Ta metoda tako kot Metoda uporabe fiksnega števila besed za razvrščanje uporablja samo učinkovite besede. Razlika med njima je, da namesto števila besed poudarja verjetnost pojava besed v nezaželeni elektronski pošti. Če je meja standardnega odklona (σ_T) vrednost x, potem

so vse besede z verjetnostjo pojava v nezaželeni elektronski pošti v območju od $0.5 - x$ do $0.5 + x$ izbrisane. Tako ostanejo le učinkovite besede, ki se uporabijo pri računanju skupne verjetnosti in dodelijo skupno oceno prejetemu elektronskemu sporočilu. Vrednost σ_T je odvisna od učinkovitosti filtra nezaželene elektronske pošte osebnih elektronskih sporočil. Za trenutno najbolj učinkovito vrednost se je izkazala 0.4, pri verjetnosti besed, ki je manjša od 0.1 in večja od 0.9.

Posebnost te metode je, da dodeli oceno elektronskega sporočila ne glede na njegovo velikost. Glede na dolžino vsebine elektronske pošte se lahko občasno pojavi le 10 učinkovitih besed ali tudi več kot 100. Vendar se v vsakem primeru pri razvrščanju uporabijo le besede, ki imajo verjetnost večjo od 0.9 ali manjšo od 0.1. Prav tako se pri tej metodi ocena elektronske pošte večinoma giblje blizu 1 (nezaželena) ali 0 (legitimna). Manj verjetno je tudi, da se bo ocena gibala okoli 0.5 in povečala možnost pojava napačno razvrščenih legitimnih elektronskih sporočil (FP).

Metoda vsako besedo v elektronskem sporočilu upošteva samo enkrat, tudi če se pojavi večkrat. Tudi mejo lahko, kot pri prejšnji metodi, dvignemo do 0.9, s čimer zmanjšamo možnost pojava napačno razvrščenih legitimnih elektronskih sporočil (FP). Čas razvrščanja elektronskih sporočil za to metodo je pogojen z velikostjo posameznega prejetega elektronskega sporočila.

3.6.4. Metoda vključitve relativnega števila besed v postopek razvrščanja

To metodo so posebej razvili Deshpande in drugi (2007), ki so hoteli raziskati vpliv uporabe relativnega števila besed v primerjavi z uporabo fiksnega števila besed, ki ga uporabljajo prej opisane metode. Pri tej metodi se tako izbere določen delež (npr. 30 %) učinkovitih besed izmed vseh besed, vključenih v besedilo elektronskega sporočila. Te besede se nato uporabijo pri računanju skupne verjetnosti, na podlagi katere se dodeli končna ocena elektronskemu sporočilu. Izbrani delež besed, ki se ga izbere za razvrščanje, je odvisen od učinkovitosti filtra nezaželenih elektronskih sporočil osebnih elektronskih sporočil.

Ta metoda je kombinacija zgoraj opisanih metod: uporabe fiksnega števila besed in standardnega odklona. Pri razvrščanju združuje učinkovitost kot tudi število besed. Če

elektronsko sporočilo vsebuje npr. 100 besed, potem bo za razvrščanje uporabljenih 30 najbolj učinkovitih. Vendar obstaja možnost, da jih bo veliko od teh 30-ih padlo v območje zavračanja zaradi meja standardnega odklona (σ_T).

Takšen način razvrščanja elektronskih sporočil združuje prednosti vseh zgoraj opisanih metod. Ker ta metoda tako kot vse ostale temelji na pregledovanju vsebine elektronskih sporočil, obstajajo možnosti, da končna ocena elektronske pošte pade blizu vrednosti 0.5. Da bi se izognili pojavu napačno razvrščenih legitimnih elektronskih sporočil (FP), lahko mejno vrednost tudi zvišamo.

4 TEST PRIMERA RAZVRŠČANJA NEZAŽELENE ELEKTRONSKE POŠTE

V empiričnem delu sem izvedel test primera razvrščanja nezaželene elektronske pošte z metodo uporabe vseh besed v postopku razvrščanja na podlagi zbranih nezaželenih in legitimnih elektronskih sporočil v slovenskem jeziku. V nadaljevanju je opisan postopek zbiranja podatkov, opis programa SpamBayes, ki sem ga uporabil pri testu, opis poteka raziskave in pridobljenih rezultatov, ocenjenih z merami uspešnosti razvrščanja in cenovno občutljivimi merami.

4.1 Zbiranje podatkov

Zbiranje podatkov za prikaz delovanja algoritmov strojnega učenja je zelo zahteven in dolgotrajen postopek, ki se ga je potrebno lotiti čim bolj sistematično in natančno. Od testnih podatkov je namreč odvisno, kakšna bo učinkovitost klasifikatorja. Le dobri testni podatki bodo omogočili, da bodo elektronska sporočila pravilno razvrščena. Problem pridobivanja dobrih testnih podatkov je ta, da je do njih izredno težko priti, saj lahko pri tem storimo več napak, ki ne bodo dale pravih rezultatov razvrščanja.

Najpogostejše napake, do katerih lahko pride pri zbiranju podatkov za testiranje algoritmov strojnega učenja na primeru slovenskih legitimnih in nezaželenih elektronskih sporočil, so:

- Legitimna in nezaželena elektronska sporočila imajo različen datum prejetja (npr. legitimna elektronska sporočila imajo novejši datum prejetja, medtem ko imajo nezaželena starejšega).
- Legitimna in nezaželena elektronska sporočila niso v istem jeziku (npr. legitimna so v slovenskem in nezaželena v angleškem).
- Legitimna in nezaželena elektronska sporočila izhajajo iz različnih elektronskih naslovov (npr. legitimna izhajajo iz elektronskega naslova legitimna@gmail.com, nezaželena pa so bila prejeta na elektronski naslov spam@gmail.com).
- Legitimna elektronska sporočila vsebujejo priponke, medtem ko jih nezaželena ne.
- Nesorazmerna količina legitimnih in nezaželenih elektronskih sporočil (npr. velika količina legitimnih in nekaj nezaželenih).

Filtri nezaželene elektronske pošte so danes postali že tako izpopolnjeni, da na podlagi najmanjše napake, ki jo storimo pri zbiranju, v stopnji učenja zaznajo vzorce, ki so jim v pomoč pri razvrščanju. Zato lahko pride do navidezne učinkovitosti klasifikatorja, katerega dejanska učinkovitost razvrščanja je manjša od prikazane. Če naši testni podatki npr. vsebujejo zgolj slovenska legitimna in angleška nezaželena elektronska sporočila, bo klasifikator to v stopnji učenja prepoznal in bo v stopnji razvrščanja pravilno vsa angleška elektronska sporočila razvrstil kot nezaželena in slovenska kot legitimna. To pomeni, da njegova klasifikacijska točnost znaša 100 %. Učinkovitost razvrščanja takšnega klasifikatorja je zato zgolj navidezna, saj bi klasifikator odpovedal v primeru prejetih slovenskih nezaželenih elektronskih sporočil, ki bi jih razvrstil kot legitimna.

Za testiranje algoritmov strojnega učenja je pomembno, da uporabimo dobre podatke, s katerimi bomo prišli do zanesljivih in preverljivih rezultatov. Vendar je danes do takšnih podatkov težko priti, saj veliko nezaželene elektronske pošte - poleg filtrov nezaželene elektronske pošte na strani uporabnikov - odstranijo že sami ponudniki storitve elektronske pošte (npr. Gmail, Outlook, Yahoo, Najdi.si). Prav tako je danes večina nezaželene elektronske pošte v angleškem jeziku, kar še dodatno otežuje zbiranje podatkov.

Zato sem se zaradi pomanjkanja slovenskih nezaželenih elektronskih sporočil odločil, da jih bom zbral sam. Da bi bili podatki čim bolj primerljivi realni situaciji, sem za zbiranje uporabil svoj osebni Gmail elektronski naslov, s katerim sem se prijavil na več slovenskih spletnih straneh, ki po elektronski pošti pošiljajo uporabnikom komercialne ponudbe. Te sem v nadaljevanju definiral kot nezaželene. Takšne spletne strani so npr.:

- Nagradne igre in kuponi – Promplac: http://promplac.si/.
- Brezplacno.com - brezplačne in zastonj stvari na internetu: http://www.brezplacno.com/.
- Hudo poceni: http://hudopoceni.si/.
- SKUPONI: https://www.skuponi.si/.
- Vsi kuponi: www.vsikuponi.si/.
- Vroči popusti: http://popusti.ceneje.si/.
- Vsi skupinski popusti na enem mestu: http://vsipopusti.si/.

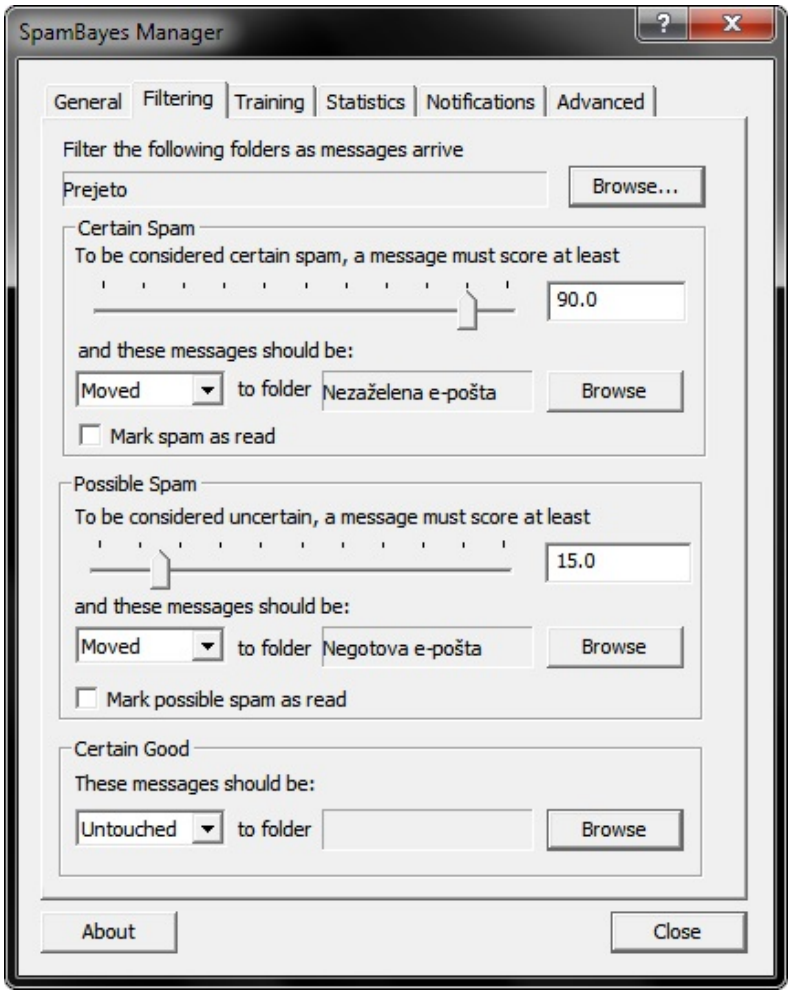
Kasneje so v moj elektronski predal začela samodejno prihajati še druga nezaželena elektronska sporočila. Tako sem uspel v treh tednih (od 28. julija do 11. avgusta) pridobiti 459 elektronskih sporočil, med katerimi je bilo 167 nezaželenih in 292 legitimnih.

4.2 SpamBayes

Odprto kodni program oz. dodatek k programu Microsoft Outlook 2010, ki sem ga uporabil za razvrščanje nezaželene elektronske pošte je SpamBayes verzija 1.16a (http://spambayes.sourceforge.net/). Program je napisan v programskem jeziku Python in za razvrščanje uporablja Bayesovo formulo o pogojni verjetnosti. Uporabil sem ga v kombinaciji s programom za pošiljanje in prejemanje elektronske pošte Microsoft Outlook 2010 (http://office.microsoft.com/sl-si/outlook/). Za SpamBayes sem se odločil, ker za razliko od drugih podobnih programov elektronsko pošto razvršča v tri in ne dve kategoriji: legitimna elektronska sporočila (»ham«), nezaželena elektronska sporočila (»spam«) in negotova elektronska sporočila (»unsure«). Med negotova elektronska sporočila razvrsti tista, ki ne morejo biti zanesljivo uvrščena kot nezaželena ali legitimna. Prav tako SpamBayes omogoča podrobnejšo analizo posameznih nezaželenih elektronskih sporočil, tako da prikaže njegovo splošno verjetnost pojava nezaželene elektronske pošte in posamezne besede, ki jih je uporabil pri razvrščanju.

Program je prednastavljen tako, da mora prejeto elektronsko sporočilo doseči mejo 90.0 točk ali več ($t = 0.9$), da je razvrščeno kot nezaželeno. To pomeni, da je strošek napačno razvrščenega legitimnega elektronskega sporočila 9-krat večji kot strošek napačno razvrščenega nezaželenega elektronskega sporočila ($\lambda=9$). Prav tako je nastavljena mejna vrednost za negotova elektronska sporočila, ki znaša 15.0 točk. Da sem lahko začel z razvrščanjem elektronske pošte, je bilo potrebno nastaviti še mapo, v katero bodo prihajala nova elektronska sporočila (Prejeto), in mapo, v katero naj se razvrstijo nezaželena (Nezaželena e-pošta) ter negotova (Negotova e-pošta) elektronska sporočila.

Slika 0.1: SpamBayes nastavitve razvrščanja elektronske pošte

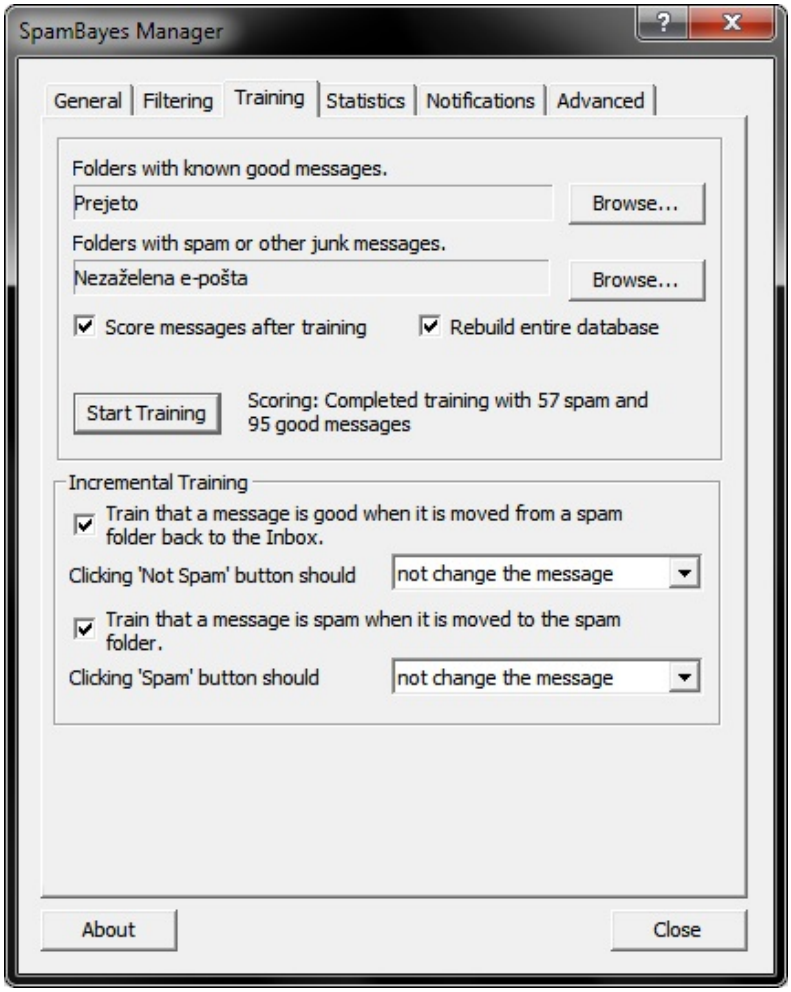


4.3 Potek raziskave

Raziskava je potekala po stopnjah razvrščanja, opisanih v poglavju 3.5. V prvi stopnji sem program za razvrščanje nezaželene elektronske pošte naučil ločevati med legitimnimi in nezaželenimi elektronskimi sporočili (glej Sliko 4.2). To sem storil tako, da sem uporabil prvo tretjino prejetih nezaželenih (57) in legitimnih (95) elektronskih sporočil ter jih iz Gmail elektronskega predala uvozil v program Microsoft Outlook 2010. Sporočila sem ročno razvrstil v mapi »Prejeta« in »Nezaželena e-pošta« ter programu SpamBayes pokazal pot do map, v katerih se nahajajo. Program je obdelal elektronska sporočila in v svojo bazo shranil 57 nezaželenih in 95 legitimnih elektronskih sporočil, ki mu bodo v pomoč v stopnji razvrščanja.

Program omogoča še dodatne nastavitve učenja, ki so uporabne v primeru, da ga redno uporabljamo za razvrščanje elektronskih sporočil. Te nastavitve mu omogočajo učenje iz storjenih napak. To pomeni, da program elektronsko sporočilo upošteva kot legitimno v primeru, da je bilo razvrščeno med nezaželena in ga je uporabnik ročno prestavil med legitimna, ter kot nezaželeno, če je je bilo razvrščeno med legitimna in ga je uporabnik prestavil med nezaželena.

Slika 0.2: SpamBayes prikaz stopnje učenja



V drugi stopnji sem nato skupaj uvozil še preostalih 307 elektronskih sporočil, med katerimi jih je bilo 110 nezaželenih in 197 legitimnih ter preveril, kako jih je filter razvrstil. Postopek razvrščanja sem ponovil na treh različnih mejnih vrednostih. Na prednastavljeni $t = 0.9$ $(\lambda = 9)$, ter na blažji $t = 0.5$ $(\lambda = 1)$, kjer sta napačno razvrščeno legitimno in nezaželeno elektronsko sporočilo enakovredna in strožji $t = 0.999$ $(\lambda = 999)$, ki enači eno napačno

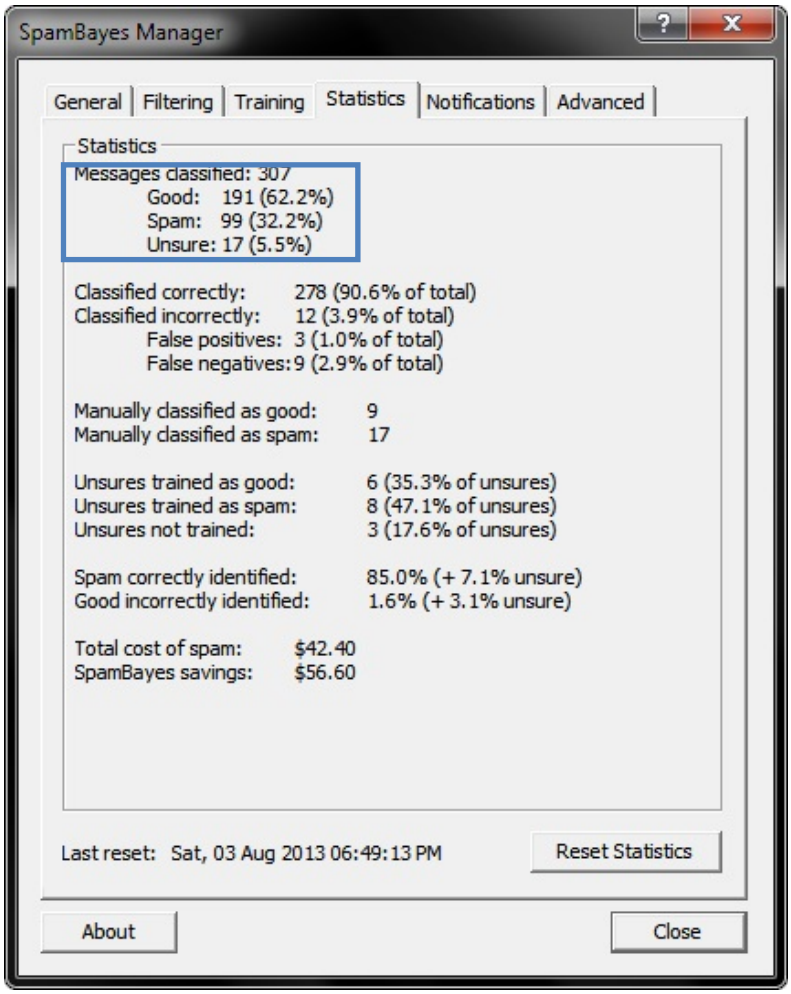
razvrščeno legitimno elektronsko sporočilo z 999 napačno razvrščenimi nezaželenimi elektronskimi sporočili.

4.4 Rezultati razvrščanja

V nadaljevanju so podrobneje opisani in predstavljeni rezultati razvrščanja z ustreznimi merami učinkovitosti in cenovno občutljivimi merami razvrščanja elektronskih sporočil. V prvem delu sem podrobneje prikazal rezultate mer uspešnosti razvrščanja in cenovno občutljivih mer za prednastavljeno mejno vrednost programa $t = 0.9$ $(\lambda = 9)$. V drugem delu pa strnjene rezultate razvrščanja za vrednost $t = 0.5$ $(\lambda = 1)$ in $t = 0.999$ $(\lambda = 999)$ ter primerjavo rezultatov razvrščanja za različne vrednosti *t*.

4.4.1. Rezultati razvrščanja za mejno vrednost t=0.9 ($\lambda = 9$)

Slika 0.3: Rezultati razvrščanja klasifikatorja



Iz Slike 4.3 je razvidno, kako je klasifikator razvrstil elektronska sporočila. 191 elektronskih sporočil je bilo razvrščenih med legitimna in 99 med nezaželena. 17 elektronskih sporočil ni bilo razvrščenih v nobeno od omenjenih kategorij.

V mapi negotovih elektronskih sporočil je bilo 16 nezaželenih elektronskih sporočil in eno legitimno. Vsa nezaželena elektronska sporočila so bila od istega pošiljatelja. Delež negotovih nezaželenih elektronskih sporočil znaša med vsemi nezaželenimi elektronskimi sporočili, vključenimi v stopnjo razvrščanja, 14.5 %. S podrobnim pregledom poteka raziskave in elektronskih sporočil, vključenih v stopnjo učenja in stopnjo razvrščanja, sem ugotovil, da so bila vsa negotova nezaželena elektronska sporočila prejeta v zadnjem tednu zbiranja podatkov. To je hkrati vzrok, da program ni mogel avtomatsko razvrstiti teh nezaželenih elektronskih sporočil, saj ni bilo v fazo učenja vključeno nobeno podobno elektronsko sporočilo.

Napaka je posledica postopka zbiranja podatkov in malega števila elektronskih sporočil, vključenih v raziskavo. Pri nadaljnjemu raziskovanju Bayesovih metod razvrščanja nezaželenih elektronskih sporočil in algoritmov strojnega učenja bi bilo potrebno pridobiti obsežnejšo bazo elektronskih sporočil, kar bi zmanjšalo pojav negotovih elektronskih sporočil, kot tudi povečalo uspešnost razvrščanja. Za lažji prikaz rezultatov razvrščanja sem sprejel odločitev, da vsa negotova elektronska sporočila ročno prestavim v ustrezno kategorijo, oz. jih obravnavam kot pravilno razvrščena.

Med 191 razvrščenimi legitimnimi elektronskimi sporočili je bilo eno nezaželeno elektronsko sporočilo. Torej je program pravilno razvrstil 190 legitimnih elektronskih sporočil. Ko k njim prištejemo še eno negotovo legitimno elektronsko sporočilo in šest napačno razvrščenih legitimnih elektronskih sporočil, pridemo do skupnega števila vseh legitimnih elektronskih sporočil, vključenih v drugo stopnjo raziskave (197).

Med 99 razvrščenimi nezaželenimi elektronskimi sporočili je bilo šest legitimnih, kar pomeni, da je bilo pravilno razvrščenih 93 nezaželenih elektronskih sporočil. Ko k njim prištejemo še 16 negotovih nezaželenih elektronskih sporočil in eno nezaželeno elektronsko sporočilo uvrščeno med legitimna, dobimo 110 vseh nezaželenih elektronskih sporočil, kar se ponovno ujema s številom nezaželenih elektronskih sporočil, vključenih v stopnjo učenja.

4.4.1.1 Kontingenčna tabela

Tabela 0.2: Rezultati razvrščanja klasifikatorja

		dejanski razred		**SKUPAJ**
		nezaželeno	legitimno	
razvrščen razred	nezaželeno	109	6	**115**
	legitimno	1	191	**192**
SKUPAJ		**110**	**197**	**307**

Ko sem preveril vsa razvrščena elektronska sporočila in negotova elektronska sporočila prestavil v ustrezne kategorije, sem podatke združil v kontingenčno tabelo, iz katere so razvidni končni rezultati razvrščanja (glej Tabelo 4.2). Iz tabele lahko razberemo, da je bilo:

- 109 pravilno razvrščenih nezaželenih elektronskih sporočil (TP),
- 1 napačno razvrščeno nezaželeno elektronsko sporočilo (FN),
- 191 pravilno razvrščenih legitimnih elektronskih sporočil (TN) in
- 6 napačno razvrščenih legitimnih elektronskih sporočil (FP).

Ko sem imel rezultate razvrščanja zbrane v tabeli dimenzije 2x2, sem na njihovi podlagi izračunal mere uspešnosti razvrščanja.

4.4.1.2 Mere uspešnosti razvrščanja

$$Priklic = Senzitivnost = TPR = \frac{TP}{TP + FN} = \frac{109}{109 + 1} = 0.991$$

Priklic nam pove delež pravilno razvrščenih pozitivnih primerov med vsemi pozitivnimi primeri. V mojem primeru znaša delež 99.1 %, kar pomeni, da je bilo pravilno razvrščenih 99.1 % nezaželenih elektronskih sporočil med vsemi nezaželenimi elektronskimi sporočili.

$$FNR = \frac{FN}{FN + TP} = 1 - TPR = 0.009$$

Mera FNR je nasprotje priklica in nam pove, da je delež napačno razvrščenih nezaželenih elektronskih sporočil med vsemi nezaželenimi elektronskimi sporočili 0.9 %.

$$FPR = \frac{FP}{FP + TN} = \frac{6}{6 + 191} = 0.03$$

Mera FPR nam pove delež napačno razvrščenih negativnih primerov med vsemi negativnimi primeri. Delež napačno razvrščenih legitimnih elektronskih sporočil med vsemi legitimnimi elektronskimi sporočili znaša 3.0 %.

$$Specifičnost = TNR = \frac{TN}{TN + FP} = 1 - FPR = 1 - 0.03 = 0.97$$

Specifičnost znaša 97.0 %. To pomeni, da je 97.0 % pravilno razvrščenih legitimnih elektronskih sporočil med vsemi legitimnimi elektronskimi sporočili.

$$Preciznost = \frac{TP}{TP + FP} = \frac{109}{109 + 6} = 0.948$$

Preciznost znaša 94.8 % in nam pove, da je med vsemi napovedanimi elektronskimi sporočili 94.8 % pravilno razvrščenih nezaželenih elektronskih sporočil.

$$Točnost = ACC = \frac{TN + TP}{TN + FP + FN + TP} = \frac{191 + 109}{191 + 2 + 3 + 111} = 0.977$$

Iz klasifikacijske točnosti (ACC) izvemo, kakšen je delež pravilno uvrščenih nezaželenih in legitimnih elektronskih sporočil med vsemi elektronskimi sporočili. V mojem primeru je delež 97.7 %, kar pomeni, da je bilo 97.7 % vseh elektronskih sporočil vključenih v razvrščanje pravilno razvrščenih.

$$Stopnja\ napake = ERR = 1 - Točnost = 1 - 0.977 = 0.033$$

Stopnja napake (ERR) je nasprotje klasifikacijske točnosti. Pove nam, da je bilo 3.3 % napačno razvrščenih nezaželenih in legitimnih elektronskih sporočil med vsemi elektronskimi sporočili. Ker nobena izmed izračunanih mer uspešnosti ne upošteva različnih stroškov

napačno razvrščenih legitimnih elektronskih sporočil v primerjavi z napačno razvrščenimi nezaželenimi elektronskimi sporočili, sem v nadaljevanju upošteval še to predpostavko.

4.1.1.3 Cenovno občutljive mere

Pri cenovno občutljivih merah sem upošteval nastavljeno mejno vrednost klasifikatorja za razvrščanje nezaželenih elektronskih sporočil na $t = 0.9$ ($\lambda = 9$), ki predpostavlja, da je strošek napačno razvrščenega legitimnega elektronskega sporočila 9-krat večji kot strošek napačno razvrščenega nezaželenega elektronskega sporočila. Potem znašata utežena točnost (W_{ACC}) in utežena stopnja napake ($W_{ERR} = 1 - W_{ACC}$):

$$W_{ACC} = \frac{\lambda TN + TP}{\lambda TN + FN + TP + \lambda FP} = \frac{9 \cdot 191 + 109}{9 \cdot 191 + 1 + 109 + 9 \cdot 6} = 0.971,$$

$$W_{ERR} = 1 - W_{ACC} = \frac{FN + \lambda FP}{\lambda TN + FN + TP + \lambda FP} = 0.029.$$

Uteženo točnost in stopnjo napake sem primerjal z enostavnim osnovnim pristopom. Njegova predpostavka je, da filter nezaželene elektronske pošte ni aktiven in vsa nezaželena elektronska sporočila pomotoma vedno zaobidejo filter ter da legitimna elektronska sporočila praviloma niso nikoli zablokirana. Tako dobimo uteženo točnost in stopnjo napake osnovnega pristopa:

$$W_{ACC}{}^{b} = \frac{\lambda TN + \lambda FP}{\lambda TN + FN + TP + \lambda FP} = \frac{9 \cdot 191 + 9 \cdot 6}{9 \cdot 191 + 1 + 109 + 9 \cdot 6} = 0.942,$$

$$W_{ERR}{}^{b} = \frac{FN + TP}{\lambda TN + FN + TP + \lambda FP} = \frac{1 + 109}{9 \cdot 191 + 1 + 109 + 9 \cdot 6} = 0.058.$$

Na podlagi formule za uteženo stopnjo napake in uteženo stopnjo napake osnovnega pristopa sem izračunal skupno razmerje med stroški TCR:

$$TCR = \frac{W_{ERR}{}^{b}}{W_{ERR}} = \frac{FN + TP}{FN + \lambda FP} = \frac{0.058}{0.029} = 2.0.$$

Vrednost skupnega razmerja med stroški (TCR) za $\lambda = 9$ znaša 2.0. Ker je $TCR < 1$, je uporaba filtra nezaželene elektronske pošte smiselna, saj da njegova uporaba boljše rezultate kot če filter ne bi bil aktiven (osnovni pristop) in bi moral uporabnik preveriti vsa prejeta elektronska sporočila in iz njih ročno izbrisati nezaželena.

4.4.2. Primerjava rezultatov za različne mejne vrednosti

Postopek razvrščanja sem ponovil še za mejni vrednosti $t = 0.5$ $(\lambda = 1)$ in $t = 0.999$ $(\lambda = 999)$. V tabeli 4.2 in 4.3 so prikazani rezultati razvrščanja za različne mejne vrednosti λ in mere uspešnosti razvrščanja.

Tabela 0.3: Rezultati razvrščanja klasifikatorja za različne vrednosti λ

λ	TP	FP	TN	FN
1	101	10	187	9
9	109	6	191	1
999	110	5	192	0

Iz tabele rezultatov razvrščanja klasifikatorja za različne vrednosti λ (glej Tabelo 4.2) lahko razberemo, da se s nižanjem mejne vrednosti od 0.999 do 0.5 povečuje tveganje za pojava napačno razvrščenih legitimnih elektronskih sporočil (FP) in napačno razvrščenih nezaželenih elektronskih sporočil (FN). Predvsem število napačno razvrščenih nezaželenih elektronskih sporočil se je pri $\lambda = 1$ $(t = 0.5)$ znatno povečalo v primerjavi z drugima dvema mejnima vrednostma. Uporaba mejne vrednosti $t = 0.5$ je v praksi zelo malo verjetna, saj predvideva, da bi moral uporabnik preveriti vsa nezaželena elektronska sporočila, preden jih izbriše.

Tabela 0.4: Mere uspešnosti razvrščanja glede na različne vrednosti λ

λ	Priklic (TPR)	Preciznost	Utežena točnost (W_{ACC})	Utežena točnost v osnovi ($W_{ACC}{}^{b}$)	Skupno razmerje med stroški (TCR)
1	99.1 %	91.6 %	96.4 %	64.2 %	9.9
9	99.1 %	94.8 %	97.1 %	94.2 %	2.0
999	100 %	95.7 %	97.5 %	99.9 %	0.04

Rezultati tabele 4.3 kažejo, da se vrednost skupnega razmerja med stroški (TCR) z nižanjem mejne vrednosti povečujejo. To pomeni, da se povečanje števila napačno razvrščenih legitimnih in nezaželenih sporočil pri nižjih mejnih vrednostih izkaže kot večji strošek. Če je strošek enak času, ki ga uporabnik porabi za pregledovanje elektronskih sporočil in brisanje napačno razvrščenih nezaželenih ter ročno razvrščanje napačno razvrščenih legitimnih elektronskih sporočil, to pomeni, da mora pri nižji mejni vrednosti za to porabiti več časa kot pri višji mejni vrednosti.

5 SKLEP

Filtri nezaželene elektronske pošte so v zadnjih letih precej napredovali in postajajo vedno bolj uspešni pri razlikovanju med nezaželeno in legitimno elektronsko pošto. Največ koristi imamo od tega predvsem uporabniki elektronske pošte, ki v svoje poštne predale prejemamo vedno manj nezaželenih elektronskih sporočil, čeprav je na svetu dnevno v obtoku še vedno okoli 30 milijard nezaželene elektronske pošte (Symantec Corporation 2013). Delo z elektronsko pošto je tako postalo bolj produktivno, hkrati pa se je povečalo tudi zaupanje v njeno uporabo.

V nalogi sem raziskoval delovanje filtrov nezaželene elektronske pošte oz. najpogostejše Bayesove metode za razvrščanje nezaželene elektronske pošte, na katerih temeljijo. Te v kombinaciji z algoritmi strojnega učenja, ki so se glede na dano situacijo sposobni samodejno prilagajati in učiti, pošto uspešno razvrščajo in ločujejo nezaželeno elektronsko pošto od legitimne. Da bi kar se da temeljito preučil postopek razvrščanja, sem v ta namen zbral bazo slovenskih nezaželenih in legitimnih elektronskih sporočil in izvedel test primera razvrščanja elektronske pošte s programom SpamBayes, ki deluje kot dodatek k Microsoft Outlook 2010. Za SpamBayes sem se odločil zato, ker za razliko od ostalih filtrov nezaželene elektronske pošte sporočila razvršča v tri kategorije: legitimna, nezaželena in negotova, kamor uvrsti tista elektronska sporočila, ki jih ne more z gotovostjo uvrstiti med nezaželena ali legitimna.

Test je potekal v dveh stopnjah. V prvi stopnji sem program na tretjini zbranih elektronskih sporočil naučil ločevati med nezaželenimi in legitimnimi elektronskimi sporočili ter nato v drugi stopnji preveril, kako je program opravil razvrščanje. Postopek sem ponovil trikrat za različne mejne vrednosti ($t = 0.5$, $t = 0.9$ in $t = 0.999$). Rezultate razvrščanja sem nato ocenil tako z najpogosteje uporabljenimi merami uspešnosti razvrščanja kakor tudi s cenovno občutljivimi merami, ki predpostavljajo, da je strošek napačno razvrščenega legitimnega elektronskega sporočila večji od napačno razvrščenega nezaželenega. Na koncu sem naredil še medsebojno primerjavo rezultatov razvrščanja za različne mejne vrednosti.

Rezultati testa so pokazali, da filtri nezaželene elektronske pošte uspešno opravijo postopek razvrščanja tudi pri manjšemu številu legitimnih in nezaželenih elektronskih sporočil, vključenih v stopnjo učenja. Tako je pri mejni vrednosti $t = 0.9$ znašal delež pravilno

razvrščenih nezaželenih elektronskih sporočil 99.1 %. Prav tako je bilo 97.0 % pravilno razvrščenih legitimnih elektronskih sporočil. Točnost razvrščanja je bila 97.7 %, kar pomeni, da je bilo pravilno razvrščenih 97.7 % vseh elektronskih sporočil. Vrednost skupnega razmerja med stroški (TCR), ki predpostavlja, da je strošek napačno razvrščenega legitimnega elektronskega sporočila večji kot strošek napačno razvrščenega nezaželenega elektronskega sporočila, je znašala 2.0. Uporabnik bi tako imel od uporabe filtra nezaželene elektronske pošte pri vrednosti $t = 0.9$ korist, saj bi porabil manj časa za brisanje napačno razvrščenih nezaželenih elektronskih sporočil in ročno razvrščanje napačno razvrščenih legitimnih elektronskih sporočil kot v primeru, če filtra ne bi uporabljal.

Medsebojna primerjava rezultatov razvrščanja za mejne vrednosti $t = 0.5$, $t = 0.9$ in $t = 0.999$ je pokazala, da se z manjšanjem mejne vrednosti povečuje verjetnost pojava napačno razvrščenih legitimnih elektronskih sporočil (FP) in napačno razvrščenih nezaželenih elektronskih sporočil (FN). Skladno s to ugotovitvijo se zmanjšujejo tudi vrednosti mer uspešnosti razvrščanja, kot sta npr. priklic in preciznost. Z nižanjem meje vrednosti se povečuje tudi skupno razmerje med stroški TCR. Uporabnik bi tako imel z nastavljeno nižjo mejno vrednostjo filtra več dodatnega dela s preverjanjem razvrščanja prejete elektronske pošte, kot če mejno vrednost zviša. Potrebno je tudi poudariti, da uporaba mejne vrednosti $t = 0.5$ v praksi ni smiselna, saj bi moral uporabnik vsa elektronska sporočila preveriti, preden izbriše nezaželena.

Sama izvedba testa primera razvrščanja elektronske pošte je razkrila nekaj težav, povezanih z bazo zbranih elektronskih sporočil. Zaradi izredno natančnih filtrov nezaželene elektronske pošte je težko pridobiti dobre testne podatke, ki bodo pokazali realne rezultate učinkovitosti razvrščanja. Hkrati sem želel test izvesti še na slovenskih elektronskih sporočilih, kar je zbiranje še dodatno otežilo, saj je do pravih nezaželenih elektronskih sporočil v slovenskem jeziku zelo težko priti. Zato sem za nezaželena uporabil vsa komercialna elektronska sporočila, do katerih sem prišel tako, da sem se prijavil na novice spletnih strani, ki po elektronski pošti pošiljajo obvestila o raznih akcijah, popustih, znižanjih itd. Največja težava baze elektronskih sporočil, ki se je pojavila kasneje pri izvedbi testa, je, da je filter 16 nezaželenih elektronskih sporočil razvrstil v kategorijo negotova, saj so nekatere ponudbe po elektronski pošti začele prihajati z zamikom. Prav tako bi bilo potrebno uporabiti večjo bazo elektronskih sporočil, kar bi dodatno vplivalo na uspešnost razvrščanja.

Razvoj filtrov nezaželene elektronske pošte bo tudi v prihodnje igral pomembno vlogo pri uporabi elektronske pošte. Pošiljatelji nezaželene elektronske pošte namreč vedno več pozornosti posvečajo ugotavljanju načinov, kako jih zaobiti, zaradi česar jim proizvajalci programske opreme le s težavo sledijo. Filtri res pomagajo, da v elektronske predale prejmemo vedno manj nezaželene elektronske pošte, vendar nas pred njo ne morejo povsem zaščititi, če sami nič ne storimo za to. Zato je pomembno, da svojih elektronskih naslovov ne objavljamo na javno dostopnih spletnih straneh in forumih, kjer bodo lahko do njih prišli pošiljatelji nezaželenih elektronskih sporočil. Prav tako ne smemo odgovarjati na prejeta nezaželena elektronska sporočila, saj bodo tako spoznali, da je na elektronski naslov aktiven in bodo na njega začeli pošiljati še več nezaželene elektronske pošte. Hkrati moramo biti pozorni na spletne strani, ki od nas zahtevajo osebne podatke in elektronski naslov, saj lahko pride do njihove zlorabe, razen v primeru, da spletno stran poznamo in ji zaupamo. Tako bo naše delo z elektronsko pošto veliko bolj varno in učinkovito.

6 LITERATURA

1 Androutsopoulos, Ion, John Koutsias, Konstantinos V. Chandrinos, George Paliouras in Constantine D. Spyropoulos 2000. An Evaluation of Naive Bayesian Anti-Spam Filtering. *Proceedings of the workshop on Machine Learning in the New Information Age.* Dostopno prek: http://arxiv.org/pdf/cs/0006013.pdf (25. avgust 2013).

2 Arnes. 2012. *Neželena elektronska pošta (spam).* Dostopno prek: http://www.arnes.si/pomoc-uporabnikom/varnostna-priporocila/nezelena-elektronska-posta-spam.html (25. avgust 2013).

3 Davis, James in Mark Goadrich. 2006. *The Relationship between Precision-Recall and ROC Curves.* Dostopno prek: https://lirias.kuleuven.be/bitstream/123456789/295592/1/d. (25. avgust 2013).

4 Deshpande, Vikas P., Robert F. Erbacher in Chris Harris. 2007. *An Evaluation of Naive Bayesian Anti-Spam Filtering Techniques.* Dostopno prek: http://digital.cs.usu.edu/~erbacher/publications/Bayes-Vikas2.pdf (25. avgust 2013).

5 Evropski parlament in Svet Evropske unije. 2002. *Direktiva 2002/58/ES EV o obdelavi osebnih podatkov in varstvu zasebnosti na področju elektronskih komunikacij.* Dostopno prek: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CONSLEG:2002L0058:20091219:SL:PDF (25. avgust 2013).

6 Fawcett, Tom. 2003. *ROC Graphs: Note and Practical Considerations for Data Mining Researchers.* Dostopno prek: http://www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf (25. avgust 2013).

7 Fawcett, Tom. 2006. An introductionto ROC analysis. *Pattern Recognition Letters, special issue on ROC analysis* 27 (8): 861–874.

8 Guzella, Tiago S. in Walmir M. Camihas. 2009. A review of machine learning approaches to Spam filtering. *Expert Systems with Applications* (36): 10206–10222.

9 Hladnik, Milan. 2002. *Verjetnost in statistika.* Ljubljana: Založba FE in FRI.

10 ITU Strategy and Policy Unit. 2004. *ITU WSIS thematic meeting on Countering Spam.* Ženeva: The ITU World Summit on the Information Society (WSIS) Thematic Meeting on Countering Spam.

11 Juričić, Aleksander. 2010. *Prosojnice: Statistika in analiza podatkov.* Dostopno prek: http://lkrv.fri.uni-lj.si/~ajurisic/stat10/folije/sap.pdf (25. avgust 2013).

12 Kononenko Igor in Marko Robnik Šikonja. 2010. *Inteligentni sistemi.* Ljubljana: Založba FE in FRI.

13 Majnik, Matjaž. 2011. *Nadgradnja mere AUC pri analizi klasifikatorjev s krivuljami ROC.* Diplomsko delo. Dostopno prek: http://eprints.fri.uni-lj.si/1560/1/Majnik1.pdf (25. avgust 2013).

14 Molan, Marko in Lina Dečman. 2005. *Kakšni so stroški nezaželene e-pošte in kdo jih plačuje.* Dostopno prek: http://profesor.gess.si/marjana.pograjc/%C4%8Dlanki_VIVID/Arhiv2005/Prispevki/23Molan2005.pdf (25. avgust 2013).

15 Možina, Bojan. 2007. *Osnove verjetnostnega računa.* Dostopno prek: http://www.fmf.uni-lj.si/~skreko/Pouk/Seminar2/Seminar2_BojanMozina_1.pdf (25. avgust 2013).

16 Porenta, Jernej, Damjan Harisc in Arnes. 2013. Spam – »Mesni narezek« v vašem e-poštnem nabiralniku. *Moj Mikro* 29 (7/8): 52–53.

17 Potamias, George, Vassilis Moustakis in Maarten van Someren. 2000. Machine Learning in the New Information Age. *Proceedings of the Workshop on Machine Learning in the New Information Age.*

18 Press, William H. 2008. *Computational Statistics with Application to Bioinformatics. Unit 17: Classifier Performance: ROC, Precision-Recall, and All That.* Dostopno prek: http://www.nr.com/CS395T/lectures2008/17-ROCPrecisionRecall.pdf (25. avgust 2013).

19 Radicati, Sara in Quoc Hoang. 2012. *Email Statistics Report, 2012-2016.* Dostopno prek: http://www.radicati.com/wp/wp-content/uploads/2012/04/Email-Statistics-Report-2012-2016-Executive-Summary.pdf (25. avgust 2013).

20 Rajput, Arun in Durga Toshniwal. 2012. Adaptive spam filtering based on bayesian algorithm. *International Journal on Advanced Computer Engeneering and Communication Tehnology* 1 (1): 8–11.

21 Sahami Mehran, Susan Dumais, David Heckerman, in Eric Horvitz. 1998. A Bayesian Approach to Filtering Junk E-Mail. *Learning for Text Categorization: Papers from the 1998 Workshop, Madison, Wisconsin, AAAI Technical Report WS-98-05.* Dostopno prek: ftp://ftp.research.microsoft.com/pub/ejh/junkfilter.pdf (25. avgust 2013).

22 Schwartz, Alan in Simon Garfinkel. 1998. *Stopping spam.* Sebastopol: O'Reilly & Associates, Inc.

23 *SpamBayes: Bayesian anti-spam classifier written in Python.* Dostopno prek: http://spambayes.sourceforge.net/ (25. avgust 2013).

24 Symantec Corporation. 2013. *Symantec Intelligence Report: June 2013*. Dostopno prek: http://www.symantec.com/security_response/publications/monthlythreatreport.jsp (25. avgust 2013).

25 Škulj, Damjan. 2009. *Matematika I (zapiski)*. Dostopno prek: http://mat.fdvinfo.net/ (25. avgust 2013).