Filter nezaželene elektronske pošte za akademski svet

Anja Vrečer

Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
Večna pot 113, 1000 Ljubljana
anja.vrecer@gmail.com

Povzetek

Nezaželena akademska elektronska sporočila so nezaželena sporočila, ki jih prejemajo predvsem profesorji, raziskovalci in drugi akademiki, in jih navadni filtri nezaželene elektronske pošte ne zaznavajo. V prispevku predstavimo izdelavo filtra nezaželene akademske elektronske pošte, pri čemer smo naredili primerjavo različnih metod filtriranja sporočil in različnih tehnik obdelave besedila. Za končni model smo uporabili nevronsko mrežo v kombinaciji z vektorskimi vložitvami besed ter ga povezali z izbranim odjemalcem elektronske pošte, in sicer z Gmailom. Filter smo testirali z 10-kratnim prečnim preverjanjem in dosegli tudi do 98% točnost.

1. Uvod

Elektronska pošta je v zadnjem času postala ena najbolj uporabljenih aplikacij za komunikacijo. Vsakodnevno jo uporablja na milijone ljudi, tako v službi kot v prostem času (Whittaker et al., 2005). Slabost vsesplošne uporabnosti elektronske pošte pa je vse večja količina elektronskih sporočil, ki jih prejemamo. Med njimi je tudi veliko nezaželenih elektronskih sporočil. Prebiranje vseh sporočil nam zato včasih vzame ogromno časa in energije. Ker želimo čim hitreje ločiti nezaželena sporočila od drugih, uporabnih sporočil, imajo mnogi poštni odjemalci že vgrajene filtre nezaželene elektronske pošte. Vendar pa takšni filtri ne zaznajo vseh vrst nezaželene elektronske pošte. V prispevku se osredotočimo na eno takšnih skupin nezaželene elektronske pošte, in sicer na nezaželeno akademsko elektronsko pošto.

Profesorji in drugi akademiki v svoj elektronski nabiralnik stalno dobivajo vabila k objavljanju člankov v različnih revijah, k sodelovanju na konferencah ali ponudbe odprtih delovnih mest. Takšne ponudbe se velikokrat ne navezujejo na prejemnikovo področje raziskovanja ali pa je takšnih ponudb preprosto preveč. Velik problem predstavljajo vabila k prispevanju člankov za manj znane ali predatorske revije. Akademiki, ki se strinjajo z objavo svojega članka v takšnih revijah, tvegajo, da je njihova kariera lahko oškodovana. Takšne revije namreč objavijo vsak članek, ki ga prejmejo in s tem razveljavijo akademsko vrednost objavljenih člankov, akademika pa zaznamujejo kot soavtorja predatorske revije (da Silva et al., 2020). Hkrati se nepazljivemu prejemniku lahko zgodi, da preko sporočila posreduje svoje osebne informacije osebam, ki imajo od tega finančno korist (Lin, 2013).

Ker je vsebina akademskih nezaželenih elektronskih sporočil pogosto precej drugačna od nezaželenih elektronskih sporočil, ki jih zazna večina navadnih filtrov nezaželene pošte, mora prejemnik sam ločevati uporabna in neuporabna sporočila. Raziskovalni prispevek našega dela je razvoj orodja za filtriranje nezaželenih akademskih elektronskih sporočil, ki za klasifikacijski model uporablja nevronsko mrežo in dosega primerljive ali celo boljše rezultate od nekaterih raziskovalcev, ki so se ukvarjali s podobnim problemom.

2. Namen članka

Obstoječi filtri nezaželene akademske elektronske pošte so v veliki večini samo "ročno" napisana pravila, ki izključujejo sporočila določenih prejemnikov ali z določenimi ključnimi besedami. Takšna pravila pa je za uspešno delovanje potrebno stalno posodabljati, saj se pošiljatelji, pa tudi vsebina oziroma besede v teh sporočil, ves čas spreminjajo. Zato smo v sklopu raziskave ustvarili filter nezaželene akademske elektronske pošte, ki temelji na modelu nevronske mreže v kombinaciji z vektorskimi vložitvami besed. Začetni model je naučen na množici 660 nezaželenih akademskih elektronskih sporočil, skupaj z 2.551 drugih sporočil. Model se lahko tudi prilagodi uporabniku, tako da upošteva uporabnikova nezaželena akademska elektronska sporočila v njegovem elektronskem nabiralniku.

3. Sorodna dela

3.1. Nezaželena akademska elektronska pošta

V tem razdelku opišemo ugotovitve o nezaželeni akademski elektronski pošti, povzete po različnih avtorjih. Pri pregledu značilnosti smo upoštevali tudi ugotovitve pri pregledu nezaželene akademske elektronske pošte iz naše testne zbirke sporočil.

Nezaželena vabila. Izkoriščevalske ali predatorske revije so revije, katerih glavni cilj ni širjenje znanja ali upoštevanje akademske kvalitete člankov, ampak nepošten zaslužek. Profesorje in druge akademike skušajo pretentati, da bi z njimi sodelovali, s tem da bi jim plačali za objavo svojih člankov. Glavne lastnosti (Wahyudi, 2017) teh revij so:

- za objavo članka je potrebno plačilo,
- revija se izdaja pogosto,
- za objavo je sprejeto nadpovprečno veliko člankov,
- čas obdelave in pregleda člankov sta nerealno hitra in
- kvaliteta objavljenih člankov je slaba ali zelo neenakomerna.

Leta 2014 je knjižničar iz Univerze v Koloradu, Jeffrey Beall, sestavil dva seznama, in sicer seznam vprašljivih

založnikov in seznam vprašljivih revij. Zapisal je, da obstajajo samo zato, da črpajo denar od avtorjev, ki morajo plačati za to, da so njihovi članki sprejeti v revijo (Wahyudi, 2017). Beallov seznam vprašljivih revij (angl. *Beall's list of predatory journals*) se najpogosteje uporablja pri identifikaciji izkoriščevalskih revij. Obstajajo tudi druge zbirke sumljivih revij, kot na primer *Alexa database* in baza lažnih spletnih strani *Phish Tank database* (Dadkhah et al., 2017). Tudi za pomoč pri identifikaciji pravih, strokovnih revij obstajajo baze, kot je na primer Direktorij odprto-dostopnih revij (angl. *Directory of Open Access Journals*) (Kozak et al., 2016).

Na podoben način so zasnovana tudi vabila na konference. V večini primerov se takšna vabila sploh ne navezujejo na prejemnikovo področje raziskovanja in ne obstajajo za širjenje znanja med podobno mislečimi akademiki, ampak je njihov namen oglaševati svoje revije in služiti (D. Cobey et al., 2017).

Zavajanje. Pri zavajanju oziroma ribarjenju (angl. *phishing attacks*) so spletne strani, na katere elektronsko sporočilo usmerja, ustvarjene z namenom, da prejemnik vanje vnese osebne podatke, kot so številka bančne kartice, gesla in podobno (da Silva et al., 2020). Te spletne strani so narejene tako, da so podobne dejanskim stranem resničnih organizacij, zato prejemnik velikokrat sploh ne ve, da gre za ponaredek (Dadkhah et al., 2017). Zavajajoča elektronska sporočila so torej podvrsta nezaželene elektronske pošte, v kateri se pošiljatelj pretvarja, da je predstavnik neke druge legitimne organizacije z namenom pridobivanja osebnih podatkov (Gupta et al., 2018). Sporočila te vrste so večinoma namenjena določeni skupini ljudi ali določeni organizaciji.

Še en način, kako delujejo zavajajoča elektronska sporočila, je s samo-izvršilno kodo. Ta način deluje tako, da se ob kliku na povezavo izvede skrit program in povzroči škodo na prejemnikovem računalniku z vgraditvijo virusa, ki uniči prejemnikove datoteke ali pa ukrade osebne informacije, gesla in druge podatke iz njega (da Silva et al., 2020).

3.2. Generična struktura nezaželenih akademskih sporočil

Wahyudi (2017) je v svojem članku natančno preučil strukturo nezaželene akademske elektronske pošte, zato v nadaljevanju opišemo glavne ugotovitve iz tega in drugih člankov.

Generična struktura nezaželene akademske elektronske pošte je sestavljena iz pozdrava, napovedi, uvoda, osrednjega dela in zaključka. Velikokrat so uporabljeni laskajoči pozdravi in nazivi, kot sta "ugledni profesor" ali "ste strokovnjak na tem področju" (Grey et al., 2016). V pozdravu sta lahko uporabljena tudi prejemnikova ime in priimek. Sporočilo velikokrat izraža hvaljenje, lažne spodbude in obljublja nagrade ali karierne priložnosti (Dadkhah et al., 2017; Soler in Cooper, 2019). Pošiljatelj velikokrat zatrjuje, da je prebral prejemnikov članek in da to sporočilo ni nezaželena pošta (da Silva et al., 2020). V veliki večini sporočilo govori o splošni temi, ki se ne navezuje na prejemnika (Grey et al., 2016; Moher in Srivastava, 2015). Še ena lastnost nezaželene akademske elektronske pošte je ta, da od prejemnika zahteva odgovor v nerealno kratkem času (Dadkhah et al., 2017). V nekaterih primerih se tudi zgodi, da če prejemnik ne odgovori na prvo sporočilo, sledijo nova (Grey et al., 2016).

Tudi pri pošiljateljih nezaželene akademske elektronske pošte so prisotne nekatere skupne značilnosti. Pošiljatelji se ponavljajo ali pošiljajo ponavljajoča sporočila več prejemnikom naenkrat (da Silva et al., 2020). Včasih je elektronski naslov zakrit, ponarejen ali pa se ne sklada s podpisom na koncu besedila (Soler in Cooper, 2019). Elektronski naslovi, ki niso zakriti imajo večinoma uradno domeno institucije, ki ji ukradejo reference (Dadkhah et al., 2017). Poleg tega je v veliko primerih lažno predstavljena lokacija sedeža pošiljatelja (Kozak et al., 2016). To pomeni, da pošiljatelj v sporočilo napiše drugo lokacijo, kot je dejanska lokacija, iz katere je bilo sporočilo poslano.

Opisane značilnosti so povzete iz ugotovitev različnih študij. Tudi pri pregledovanju nezaželene akademske elektronske pošte, ki smo jo uporabili za učno množico, smo opazili podobne značilnosti. Nekateri filtri nezaželene akademske elektronske pošte sicer upoštevajo najdene skupne lastnosti teh sporočil, vendar pa so to v večini "na roko" napisana pravila, ki jih je za dobro delovanje potrebno stalno spreminjati. Zato v nadaljevanju opišemo razvoj filtra nezaželene elektronske pošte, ki deluje na podlagi klasifikatorja, ki avtomatsko klasificira elektronska sporočila.

4. Razvoj filtra nezaželene pošte

V tem poglavju predstavimo zasnovo filtra nezaželene akademske elektronske pošte. Najprej opišemo učno množico sporočil in tehnike obdelave besedila, ki smo jih uporabili. Zatem predstavimo poenostavljen načrt filtra. Poglavje zaključimo z opisom povezave filtra z izbranim odjemalcem elektronske pošte.

4.1. Učna množica elektronskih sporočil

Učno množico elektronskih sporočil smo pridobili iz dveh različnih virov, saj ni bilo mogoče najti ustrezne zbirke akademskih sporočil, ki bi zajemala tako nezaželena kot tudi druga akademska sporočila. Uporabili smo nezaželena akademska sporočila od profesorjev z Univerze v Ljubljani in druga sporočila s spleta. Skupno smo zbrali 660 sporočil, označenih kot nezaželena akademska elektronska sporočila. Drugo skupino sporočil, ki niso nezaželena, smo našli na spletu, in sicer na spletni strani kaggle (van Lit, 2019). Omenjena spletna zbirka vsebuje nezaželeno in drugo elektronsko pošto, vendar pa ta sporočila nimajo akademske vsebine. Za potrebe izdelave našega sistema smo uporabili le sporočila, ki niso nezaželena. Iz omenjene spletne baze sporočil smo dobili 2.551 sporočil, ki smo jih uporabili kot učne primere sporočil, ki niso nezaželena. Ker je bila množica elektronskih sporočil sestavljena iz sporočil iz različnih virov, je bilo potrebno sporočila pretvoriti v enako obliko, primerno za nadaljnjo obdelavo. Poleg tega je bilo potrebno obdelati besedilo sporočila in ga ustrezno spremeniti. V nadaljevanju opišemo, kako smo se lotili tega problema.

Vsako sporočilo v učni množici smo spremenili v slovar s ključi `Subject` (zadeva), `Sender` (pošiljatelj), `Receiver` (prejemnik), `Date` (datum prejema) in `Body` (telo sporočila). Sporočila v skupini sporočil, ki

niso nezaželena, imajo isti vir in obliko. Zato smo vsa sporočila v tej skupini lahko pretvorili v slovar na isti način. Nezaželena sporočila pa smo dobili iz različnih virov in jih je bilo zato potrebno spremeniti v slovar na različne načine glede na končnico datoteke.

Naslednji korak obdelave sporočil je pretvorba slovarjev sporočil v obliko, primerno za model. Lai (2007) v svojem članku trdi, da je najbolj uporaben del za klasifikacijo nezaželenih sporočil zadeva sporočila in da samo telo sporočila ne klasificira tako dobro, kot če je v kombinaciji z zadevo. To smo tudi preizkusili in se odločili, da tudi mi uporabimo kombinacijo zadeve in telesa sporočila. Poleg tega Méndez et al. (2006) pravijo, da priloga, ki je lahko priložena sporočilu in jo spremenimo v besedilo, doda nepotrebne informacije, ki niso dobre za klasifikacijo. Zato priloge sporočila nismo uporabili.

Sledi opis obdelave besedila sporočil. V zadevi so bile v nekaterih primerih v oglatih oklepajih zapisane oznake sporočila (na primer `INBOX`). Zato smo iz besedila odstranili del, ki je v oglatih oklepajih. Predvsem v množici sporočil, ki niso nezaželena, je veliko sporočil, ki vsebujejo druga sporočila (nizi izmenjujočih odgovorov). Zato smo morali najti takšne dele sporočil in jih odstraniti. To smo naredili tako, da smo odstranili vrstice, ki se začnejo z določenim znakom ali nizom znakov, kot so na primer: "`To:`", "`From:`", "`Wrote:`" itd.

Nato smo zadevo in telo sporočila obdelali na enak način. Najprej smo odstranili velike začetnice in celotno sporočilo spremenili v male črke. Opazili smo, da se v nekaterih nezaželenih sporočilih pojavljajo znaki, ki izgledajo kot črke, vendar so v resnici drugi znaki in jih program zazna kot ločila. Primeri znakov, ki smo jih našli v naši zbirki sporočil, so prikazani na sliki 1. Pošiljatelji nezaželenih akademskih sporočil so s tem očitno želeli preprečiti filtrom nezaželene elektronske pošte razpoznavo nekaterih besed, ki nakazujejo na nezaželeno akademsko elektronsko pošto. Najdene znake smo zamenjali s pravo črko in na koncu besedila dodali značko "`specialchars`". V besedilu smo poiskali klicaje in jih zamenjali z značko "`exclamationmark`". Opazili smo namreč, da izrazito velika raba klicajev lahko nakazuje na to, da je sporočilo nezaželena akademska elektronska pošta. Poleg tega smo poiskali elektronske naslove, povezave in imena mesecev ter jih zamenjali z značkami "`emailwashere`", "`linkwashere`" in "`monthwashere`", saj struktura elektronskega naslova in povezave ter ime meseca niso pomembni. Poleg omenjenega smo iz sporočil odstranili ločila in nepotrebne besede, kot so vezniki, zaimki in vprašalnice. V angleščini se te pogoste in nepotrebne besede imenujejo *stop words*.

Da bi identiteta profesorjev, ki so prispevali nezaželena akademska sporočila za učno množico, ostala skrita, je bilo potrebno iz sporočil odstraniti imena prejemnikov. Poleg tega smo odstranili tudi imena pošiljatelja, saj je tudi ta podatek nepotreben pri klasifikaciji. Ime ali priimek smo zamenjali z značko `receivername` za prejemnika oziroma `sendername` za pošiljatelja.

Na opisani način smo sporočila spremenili iz seznama slovarjev v seznam besedil oziroma zbirko besedil (angl. *corpus*). Sporočila smo nato shranili s pomočjo knji-

Znak	Črka
α	a
a	
b	b
c	c
c	
c	
d	d
e	e
e	

Znak	Črka
l	i
i	
i	
ι	
l	
j	j
μ	m
o	o
p	p

Znak	Črka
ρ	p
r	r
s	s
s	
u	u
u	
υ	
v	v

Slika 1: Primeri znakov iz nezaželenih akademskih sporočil in ustrezne črke, ki so jih pošiljatelji nezaželenega akademskega sporočila zamenjali.

žnice, ki objekt serializira (angl. *serialize*) in s tem spremeni v binarni tok (angl. *byte stream*). Tako shranjenih sporočil ne moremo brati direktno iz datoteke, ampak jih je za branje potrebno pretvoriti nazaj v besedilo.

4.2. Tehnike obdelave sporočil

Štetje ponovitev besed. Najbolj enostavna tehnika obdelave sporočil je štetje ponovitev besed v posameznem sporočilu. Zbirke besedil smo spremenili v matriko, v kateri vsaka vrstica predstavlja sporočilo, stolpec pa besedo. Ker je vseh besed v vseh sporočilih lahko zelo veliko, smo se omejili na 2000 besed. Poskusili smo tudi z odstranitvijo besed, ki se pojavijo v manj kot treh sporočilih, tako kot so to opisali Sakkis in sodelavci (Sakkis et al., 2003).

Frekvenca besed z inverzno frekvenco v dokumentih (angl. *term frequency-inverse document frequency* - TF-IDF). Frekvenca besed (angl. *term frequency*) enostavno pomeni število besed v posameznem sporočilu. Inverzna frekvenca v dokumentih (angl. *inverse document frequency*) pa predstavlja informativnost besede, torej ali se beseda pogosto ali redko pojavlja v sporočilih (Hakim et al., 2014). TF-IDF besede izračunamo tako, da uporabimo enačbo (1), pri čemer je t têrmin in d dokument oziroma sporočilo. $TF(t, d)$ predstavlja frekvenco besede t v sporočilu d, $IDF(t)$ pa je inverzna frekvenca besede t v dokumentih. Izračuna se jo z enačbo (2), pri čemer je n število vseh sporočil, $DF(t)$ pa število sporočil v katerih se beseda t pojavi vsaj enkrat. V imenovalcu ulomka vrednosti $DF(t)$ dodamo še $+1$, da se izognemo deljenju z nič.

$$TF\text{-}IDF(t, d) = TF(t, d) * IDF(t) \tag{1}$$

$$IDF(t) = log\left(\frac{n}{DF(t) + 1}\right) + 1 \tag{2}$$

Prednost uporabe tehnike TF-IDF je, da se normalizira vpliv besed, ki se v dokumentih pojavljajo zelo pogosto in so zato manj informativne kot besede, ki se pojavijo manjkrat.

Medsebojna informacija. Za izbiro atributov smo preizkusili tudi odstranitev atributov, ki imajo premajhno medsebojno informacijo (angl. *mutual information*). Medse-

bojna informacija dveh naključnih spremenljivk je nenegativna vrednost, ki pove odvisnost med tema spremenljivkama (Kraskov et al., 2004). Z drugimi besedami, medsebojna informacija meri količino informacije, ki jo pridobimo o neki spremenljivki, če imamo podano neko drugo spremenljivko (Witten in Frank, 2000). Večja ko je medsebojna informacija dveh spremenljivk, bolj sta spremenljivki odvisni med sabo. Če pa je medsebojna informacija enaka nič, sta spremenljivki popolnoma neodvisni. Medsebojno informacijo med dvema naključnima spremenljivkama lahko izračunamo z enačbo (3), kjer $I(X;Y)$ predstavlja medsebojno informacijo za spremenljivki X in Y, $H(X)$ predstavlja entropijo spremenljivke X, $H(X|Y)$ pa je pogojna entropija za spremenljivko X, če imamo podano spremenljivko Y. Entropija je enaka povprečni lastni informaciji in prestavlja stopnjo negotovosti oziroma informacije. Izračunamo jo s pomočjo formule (4), kjer so možni rezultati $x_1, ..., x_n$ in $P(x_i)$ verjetnost rezultata x_i. Pogojno entropijo pa izračunamo s formulo (5).

$$I(X;Y) = H(X) - H(X|Y) \tag{3}$$

$$H(X) = -\sum_{i-1}^{n} P(x_i) log P(x_i) \tag{4}$$

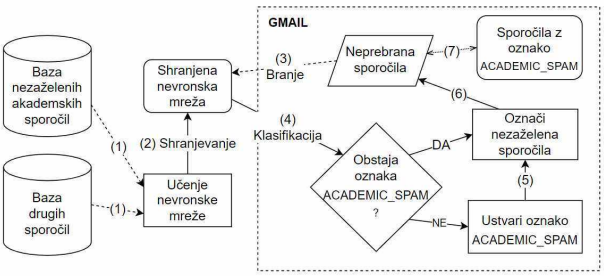
$$H(X|Y) = -\sum_{x \in X, y \in Y} p(x,y) log \frac{p(x,y)}{p(x)} \tag{5}$$

Vektorska vložitev besed (angl. *word to vector embedding*). Vektorska vložitev besed je tehnika predstavitve besed z vektorji, ki ohranjajo pomenske značilnosti besed. To pomeni, da so besede, ki so si pomensko bolj podobne, bolj blizu v vektorskem prostoru (Ghannay et al., 2016). Vektorje besed se sestavi glede na to, katere besede se v stavku nahajajo skupaj, saj se tako najlažje ugotovi pomen besede. Zaradi tega je za učinkovito sestavljanje besednih vektorjev potrebna velika učna množica besedil. Ker je to velikokrat težko pridobiti in ker je učenje vektorjev lahko precej zamudno, na spletu obstajajo baze besed in njihovih vektorjev, ki so naučeni na velikih množicah besedil. Primeri zbirk naučenih vektorjev so na primer `Googlova zbirka` in `zbirka GloVe` (*Global Vectors for Word Representation*) (Pennington, 2014).

Ker je zbirka vektorjev iz baze GloVe naučena na ogromni množici besedil in je prosto dostopna, smo se odločili, da jo bomo uporabili v našem sistemu. Na voljo ima več zbirk vektorjev iz različnih virov in velikosti. Zbirke smo preizkusili in ocenili njihovo uspešnost. Preizkusili smo tudi različno maksimalno število besed v posameznem sporočilu in maksimalno število unikatnih besed. V končnem sistemu smo uporabili 100-dimenzionalne vektorje, omejitev 2.000 besed na sporočilo in omejitev 500.000 različnih besed.

4.3. Zasnova filtra nezaželenih akademskih sporočil

Zgradili smo programsko rešitev, sestavljeno iz dveh programov. Prvi, glavni program, je namenjen klasifikaciji neprebranih sporočil. Drugi program pa je namenjen posodobitvi nevronske mreže glede na uporabnikova označena sporočila. Pri gradnji programske rešitve smo preizkusili več klasifikatorjev, in sicer smo preizkusili naivni Bayes, naključni gozd, metodo podpornih vektorjev, logistično regresijo in različne nevronske mreže. V končnem sistemu smo uporabili nevronsko mrežo, saj so bili rezultati testiranja pri tem klasifikacijskem modelu najboljši. Za odjemalec elektronske pošte pa smo izbrali Gmail (Google, 2022).

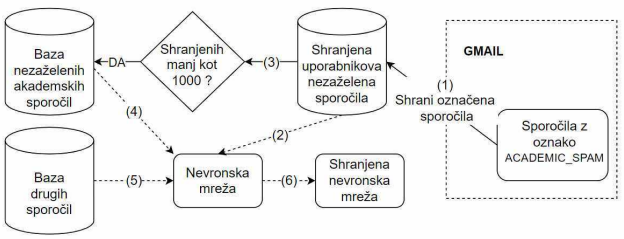


Slika 2: Načrt delovanja sistema ob prvem učenju nevronske mreže in ob klasifikaciji neprebranih sporočil.

Slika 2 prikazuje delovanje sistema ob zagonu programa za klasifikacijo neprebranih sporočil. Najprej program preveri, ali je nevronska mreža že shranjena na disku. Če ni, se izvede začetno učenje nevronske mreže. Za učenje nevronske mreže so potrebni označeni učni podatki, kar so v našem primeru nezaželena akademska elektronska sporočila in druga elektronska sporočila. Ker je vir sporočil lahko različen, se elektronska sporočila pretvori v enako obliko in obdela tako, da se odstrani nepotrebne atribute sporočila. Ta korak je na sliki označen s številko (1). Sledi učenje nevronske mreže in shranjevanje na disk (2). Nevronska mreža je tako ob naslednjem zagonu pripravljena na klasifikacijo in ni potrebno pri vsakem zagonu čakati na učenje nevronske mreže. Naslednji korak programa je branje neprebranih sporočil iz elektronskega nabiralnika (3). Prebrana sporočila se obdelajo na enak način kot pri koraku (1). Shranjena nevronska mreža nato klasificira neprebrana sporočila. Če je katero izmed neprebranih sporočil klasificirano kot nezaželena akademska elektronska pošta, program preveri, ali v uporabnikovem elektronskem nabiralniku že obstajajo sporočila z oznako ACADEMIC_SPAM. Če ne, program za uporabnika ustvari novo oznako ACADEMIC_SPAM in označi ustrezna sporočila (5). Če oznaka že obstaja, program samo označi ustrezna sporočila s to oznako. Oznaka se nato prikaže na neprebranih sporočilih v uporabnikovem elektronskem nabiralniku (6), hkrati pa nastane oziroma se dopolnjuje tudi mapa sporočil z oznako ACADEMIC_SPAM (7).

Slika 3 prikazuje drugi program, ki je namenjen posodobitvi nevronske mreže, tako da se čim bolj prilagodi uporabniku. Posodobitev deluje samo, če ima uporabnik v svojem elektronskem nabiralniku sporočila, označena z oznako ACADEMIC_SPAM.

Program najprej prebere sporočila, ki so označena z oznako ACADEMIC_SPAM. Nato ta sporočila doda k shranjenim uporabniškim sporočilom ali pa ustvari novo datoteko s shranjenimi uporabnikovimi nezaželenimi akademskimi sporočili (1). Ta sporočila se potem uporabi kot del učne množice pri učenju nevronske mreže (2). Če je shra-



Slika 3: Načrt delovanja sistema ob posodobitvi nevronske mreže.

njenih uporabnikovih sporočil več kot 1000, se sporočila razvrsti po datumu prejema in se jih izbere le zadnjih 1000. Če pa je shranjenih uporabnikovih sporočil manj kot 1000 (3), se množica nezaželenih akademskih sporočil dopolni s sporočili iz baze nezaželenih akademskih sporočil (4). Poleg nezaželenih akademskih sporočil nevronska mreža za učenje potrebuje tudi množico drugih sporočil. Te se pridobijo iz baze drugih sporočil (5). Nevronska mreža se nato nauči na podanih učnih podatkih in posodobljena mreža se shrani na disk (6), kjer je na voljo za naslednjo klasifikacijo neprebranih sporočil.

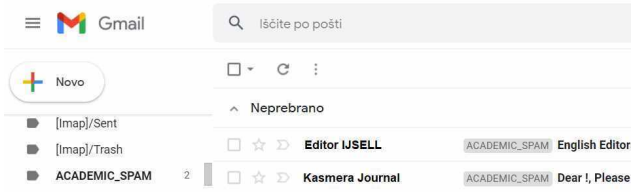
4.4. Povezava z odjemalcem elektronske pošte

Filter nezaželene akademske elektronske pošte smo povezali z brezplačno e-poštno storitvijo, ki jo ponuja Google, in sicer Gmail. Za povezavo tega spletnega odjemalca elektronske pošte s programom smo uporabili Gmail API. To je aplikacijski programski vmesnik, ki temelji na arhitekturi REST (angl. *RESTful API*) (Developers, 2021). Arhitektura REST (angl. *representational state transfer*) je arhitektura za izmenjavo podatkov med spletnimi storitvami, kjer je vsak vir dostopen z enoličnim identifikatorjem vira URL. Uporablja se za dostop do Gmail elektronskih nabiralnikov in pošiljanje elektronskih sporočil preko programa.

Program smo z Gmail API-jem povezali s pomočjo računalniškega okolja `Google Cloud`. Tam smo ustvarili nov projekt in v njem omogočili Gmail API ter dodali avtorizacijo in avtentikacijo za program. Uporabili smo `API Keys` in `OAuth 2.0 Client IDs` za omogočanje `Gmail APIja` v programu.

Če je povezovanje z `Gmail APIjem` uspešno, je program pripravljen na branje uporabnikovih neprebranih sporočil. V primeru, da neprebrana sporočila ne obstajajo, se izpiše sporočilo: `"No messages found."` in program se zaključi. V nasprotnem primeru pa se iz podatkov, pridobljenih z `Gmail APIjem`, generira slovar s ključi `Subject` (zadeva), `Sender` (pošiljatelj), `Receiver` (prejemnik), `Date` (datum prejema) in `Body` (telo sporočila). Sporočila v obliki slovarja je nato potrebno preurediti v obliko, primerno za klasifikator, podobno kot smo to naredili za sporočila v učni množici (glej razdelek 4.1.). Tako smo namesto seznamov slovarjev dobili seznam obdelanih besedil. Ta seznam smo nato s pomočjo shranjenih vektorjev spremenili v seznam vektorjev in ga pretvorili v matriko.

Naslednji korak je nalaganje shranjenega klasifikatorja in klasifikacija neprebranih sporočil. Če klasifikator označi katerega izmed sporočil kot nezaželeno akademsko pošto, se izvede del programa za posodobitev oznak. Najprej program preko `Gmail APIja` prebere vse oznake, ki obstajajo v uporabnikovem elektronskem nabiralniku, in preveri, ali je katera med njimi `ACADEMIC_SPAM`. Če oznaka že obstaja, se sporočilom, ki jih je klasifikator označil kot nezaželena, doda ta oznaka. Če oznaka še ne obstaja, pa se ustvari nova oznaka `ACADEMIC_SPAM`.



Slika 4: Izsek elektronskega nabiralnika v Gmailu, kjer sta bili dve neprebrani sporočili klasificirani kot nezažalena akademska pošta.

Rezultat zagona programa in klasifikacije neprebranih sporočil je oznaka `ACADEMIC_SPAM`, ki se prikaže na ustreznih sporočilih. Na sliki 4 je prikazan primer takšne klasifikacije v Gmailu. Pred zadevo sporočil, klasificiranih kot nezaželena akademska sporočila, se pojavi oznaka `ACADEMIC_SPAM`. Hkrati pa lahko na levi strani v seznamu vseh oznak opazimo oznako `ACADEMIC_SPAM`, kjer lahko najdemo vsa sporočila, ki so bila v preteklosti označena kot nezaželena akademska sporočila.

5. Testiranje in rezultati

V zadnjem poglavju predstavljamo način testiranja preizkušenih modelov klasifikacije in obdelave elektronskih sporočil. Primerjamo rezultate in prikažemo rezultate algoritma SHAP, ki poišče besede, ki so najbolj pripomogle h klasifikaciji nezaželenih akademskih sporočil.

5.1. Način testiranja

Za testiranje uspešnosti smo uporabili 10-kratno prečno preverjanje (angl. *10-fold cross-validation*). Pri tej metodi učno množico razdelimo na 10 približno enako velikih množic in za vsak model naredimo 10 ponovitev testiranja. V vsaki iteraciji vzamemo za testno množico eno izmed množic, ostale množice pa združimo v učno množico.

Na tak način bolj natančno preverimo uspešnost modelov, kot če bi iz množice naključno izbrali 10% primerov in le enkrat testirali model. Pri 10-kratnem prečnem preverjanju je namreč vsak primer v množici enkrat uporabljen kot testni. Tako lahko na koncu izračunamo povprečje in standardno deviacijo rezultatov iz vseh ponovitev testiranja ter dobimo bolj realne rezultate. Poleg tega smo lahko zaradi večkratne ponovitve testiranja za primerjavo modelov uporabili tudi statistične teste.

5.2. Rezultati

Del rezultatov testiranja različnih modelov je prikazan v tabeli 1. Preizkusili smo različne tehnike obdelave besedila, v tabeli pa so prikazani rezultati ob uporabi tehnike TF-IDF z odstranitvijo besed, ki se pojavijo v manj kot treh sporočilih in besed, ki imajo medsebojno informacijo manjšo kot 0.01 pri prvih petih modelih ter

vektorsko vložitvijo besed pri zadnjem modelu nevronske mreže. Uporabili smo celotno množico sporočil, in sicer 660 nezaželenih akademskih sporočil in 2.551 drugih sporočil. Kot lahko vidimo, so rezultati že pri teh modelih precej dobri, saj so pravilno klasificirana skoraj vsa sporočila iz testne množice sporočil.

Tabela 1: Povprečne vrednosti in standardna deviacija testiranja z 10-kratnim prečnim preverjanjem.

Model	Točnost	F1	AUC
Naivni Bayes	88.49% ± 2.40%	77.29% ± 5.27%	0.91 ± 0.02
Naključni gozd	98.32% ± 0.32%	95.88% ± 0.79%	0.96 ± 0.01
SVM	98.65% ± 0.68%	96.62% ± 1.79%	0.97 ± 0.01
Logistična regresija	98.82% ± 0.58%	97.02% ± 1.55%	0.97 ± 0.01
Nevronska mreža	98.98% ± 0.42%	97.49% ± 1.10%	0.98 ± 0.01
Nevronska mreža z GloVe	98.69% ± 0.62%	96.79% ± 1.47%	0.97 ± 0.01

Tabela 2: Povprečni rangi uspešnosti modelov glede na vrednost AUC.

Naivni Bayes	Naključni gozd	SVM	Logistična regresija	Nevronska mreža
5	2.9	2.65	2.15	2.3

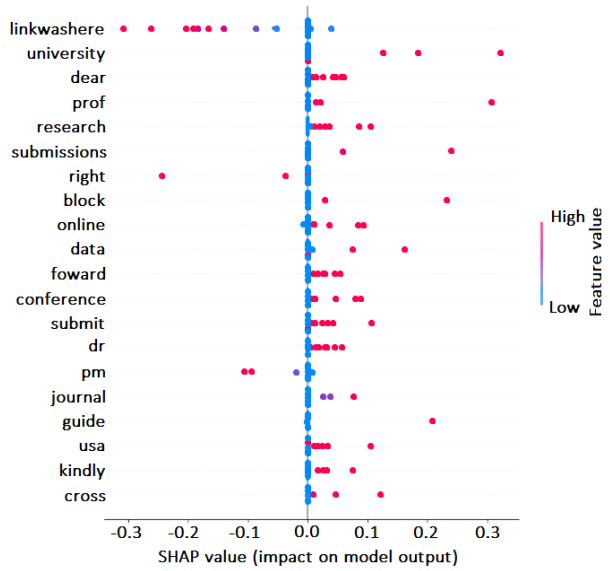
Za primerjavo modelov smo uporabili Friedmanov test (Friedman, 1937). Natančno razlago uporabe tega testa opisuje Demšar (2006). Najprej smo primerjali skupino klasifikacijskih modelov, na katerih smo uporabili prej omenjene tehnike obdelave besedila in so v tabeli na prvih petih mestih. S Friedmanovim testom pri $\alpha = 0.05$ na AUC smo preverili, ali lahko za kateri par modelov rečemo, da je eden izmed njiju izrazito boljši od drugega. Povprečni rangi uspešnosti modelov glede na AUC so razvidni v tabeli 2. Izračunali smo kritično razdaljo $CD = 1.93$ in jo primerjali z razlikami povprečnih vrstnih redov uspešnosti modelov ter ugotovili, da so vsi modeli izrazito boljši za klasifikacijo nezaželenih akademskih sporočil kot naivni Bayes. Za ostale pare modelov s Friedmanovim testom tega nismo mogli dokazati.

Čeprav so bili rezultati že pri teh modelih precej dobri, smo se vseeno odločili implementirati še različne modele nevronskih mrež v kombinaciji z vektorskimi vložitvami besed. Zgradili smo več različnih nevronskih mrež in jih med seboj primerjali. V tabeli 2 je na zadnjem mestu prikazan rezultat ene izmed teh nevronskih mrež. Čeprav je rezultat nekoliko slabši od zgoraj opisanih modelov, smo v končnem sistemu vseeno uporabili to nevronsko mrežo z vektorskimi vložitvami besed. Ta metoda namreč upošteva pomene besed in ne samo njihov zapis, tako kot ostale metode obdelave besedil.

Rezultati našega testiranja so nekoliko boljši od rezultatov nekaterih raziskovalcev, ki so se ukvarjali s podobnim problemom. Sicer nismo našli primerov, v katerih bi raziskovalci skušali klasificirati nezaželeno akademsko elektronsko pošto, vseeno pa lahko do neke mere primerjamo naše rezultate z rezultati navadnih filtrov nezaželenih elektronskih sporočil. Koprinska et al. (2007) so na eni izmed testnih množic z modelom naključnega gozda dosegli točnost 96.03%, natančnost 95.62%, priklic 95.62% in F1 mero 94.16%. Za obdelavo sporočil so uporabili posebno metodo izbire atributov, in sicer varianco frekvence têrma (angl. *term frequency variance*). Ostali modeli v njihovem primeru niso bili tako uspešni. Lai (2007) je v članku opisal preizkus modelov Naivnega Bayesa, k-najbližjih sosedov, SVM in kombinacijo TF-IDF z metodo SVM. Za najbolj uspešen model se je izkazala kombinacija TF-IDF z metodo SVM. S to metodo so v enem primeru dosegli točnost 93.43%.

5.3. Razlaga klasifikacije z algoritmom SHAP

Razumljivost in enostavna razlaga modela sta izjemno pomembni za interpretacijo rezultatov in možnost nadgradnje modela. To je velikokrat razlog, da se nekateri raziskovalci odločijo za uporabo enostavnih (linearnih) modelov namesto kompleksnejših, ki jih je težko razumeti. Vendar pa je zaradi naraščajoče količine podatkov, ki jih želimo obdelati, nujno, da uporabljamo tudi slednje. Za to obstajajo algoritmi, ki nam jih pomagajo razumeti in interpretirati rezultat njihove klasifikacije. Eden takšnih algoritmov je algoritem SHAP (Lundberg in Lee, 2017).



Slika 5: Slika prikazuje besede, ki najbolj vplivajo na rezultat klasifikacije nevronske mreže. Besede, ki imajo več pik na desni strani, so prispevale k temu, da je bilo sporočilo klasificirano kot nezaželeno.

Algoritem SHAP (*SHapley Additive exPlanations*) oziroma Shapleyjeve aditivne razlage je algoritem, ki za podane primere razloži, zakaj jih je model klasificiral tako, kot jih je. Z drugimi besedami, algoritem SHAP nam pove, kako posamezen atribut vpliva na napoved modela. V primeru klasifikacije sporočil s tem algoritmom torej lahko ugotovimo, katere besede najbolj vplivajo na rezultat klasifikacije. Na sliki 5 je prikazan rezultat algoritma SHAP na eni izmed ustvarjenih nevronskih mrež. Zaradi zahtevnosti algoritma smo uporabili manjšo podmnožico testnih

sporočil. Graf na sliki od spodaj navzgor prikazuje katere besede naj bi najbolj vplivale na klasifikacijo nezaželenih sporočil. Beseda `linkwashere`, s katero smo nadomestili vse url povezave, očitno najbolje nakazuje na to, da sporočilo ni nezaželeno. Besede, ki močno nakazujejo na to, da je sporočilo nezaželeno akademsko elektronsko sporočilo pa so `university` (univerza), `dear` (dragi oz. spoštovani), `prof` (kratica za profesor), `research` (raziskava) in `submissions` (oddaje).

6. Zaključek

Za cilj smo si zadali izdelavo filtra nezaželene akademske elektronske pošte, ki bi med neprebranimi elektronskimi sporočili v uporabnikovem elektronskem nabiralniku, čim bolj učinkovito poiskal nezaželena akademska elektronska sporočila in jih označil. Za dosego tega cilja smo morali preučiti strukturo in skupne značilnosti nezaželene akademske elektronske pošte ter preiskati obstoječe načine filtriranja nezaželene elektronske pošte. S testiranjem smo določili, da je model nevronske mreže najbolj učinkovit pri filtriranju nezaželene akademske elektronske pošte, zato smo ga tudi uporabili v končnem sistemu.

Ugotovili smo, da obstaja zelo malo rešitev za filtriranje nezaželene elektronske pošte, katerih osrednji cilj bi bil filtriranje nezaželenih akademskih elektronskih sporočil. Velika večina teh rešitev uporablja le prepoznavanje znanih pošiljateljev nezaželenih akademskih sporočil, vendar pa je za učinkovitost tega načina filtriranja potrebno stalno posodabljanje seznama. Zato smo implementirali sistem, ki neprebrana elektronska sporočila klasificira kot nezaželeno akademsko elektronsko pošto, glede na pomen besed v sporočilih. To smo dosegli z vektorsko vložitvijo besed v kombinaciji z modelom nevronske mreže. Poleg tega smo izdelali program, ki lahko klasifikacijski model posodobi glede na uporabnikovo nezaželeno akademsko elektronsko pošto. Na tak način se model lahko prilagodi uporabnikovemu elektronskemu nabiralniku in še bolj natančno označuje nezaželena akademska elektronska sporočila.

Ena izmed večjih pomanjkljivosti opisane rešitve je nadomestitev akademskih sporočil, ki niso nezaželena, z navadnimi nezaželenimi sporočili. Zaradi varovanja osebnih podatkov namreč nismo mogli uporabiti sporočil profesorjev, pa tudi na spletu ni bilo mogoče najti zbirk s takšnimi akademskimi sporočili. Tudi profesorji in drugi akademiki sicer dobivajo takšna navadna sporočila in so zato tudi ta sporočila do neke mere ustrezna za učno množico. Vseeno pa bi bilo potrebno preveriti, da klasifikator zaradi pomanjkanja akademskih sporočil, ki niso nezaželena, ne označi kar vseh akademskih sporočil, kot nezaželena.

Sistem bi lahko izboljšali še tako, da bi ob posodobitvi modela upoštevali ne samo uporabnikova nezaželena akademska sporočila, ampak tudi druga sporočila. Poleg tega sistem trenutno dobro deluje le za angleška sporočila, saj je naša množica učnih sporočil bila sestavljena le iz angleških sporočil. Možna izboljšava bi torej lahko bila prepoznavanje jezikov in prilagajanje filtra nanje. Dodali bi lahko tudi uporabniški vmesnik, ki bi uporabniku olajšal uporabo sistema.

7. Zahvala

Zahvaljujem se prof. dr. Zoranu Bosniću za vodenje, nasvete in mentorstvo med raziskavo ter profesorjem Fakultete za računalništvo in informatiko, Univerze v Ljubljani, ki so prispevali nezaželena akademska elektronska sporočila za učno množico sporočil.

8. Literatura

Kelly D. Cobey, Miguel de Costa e Silva, Sasha Mazzarello, Carol Stober, Brian Hutton, David Moher in Mark Clemons. 2017. Is this conference for real? navigating presumed predatory conference invitations. *Journal of oncology practice*, 13(7):410–413.

Jaime A Teixeira da Silva, Aceil Al-Khatib in Panagiotis Tsigaris. 2020. Spam emails in academia: issues and costs. *Scientometrics*, 122(2):1171–1188.

Mehdi Dadkhah, Glenn Borchardt in Tomasz Maliszewski. 2017. Fraud in academic publishing: researchers under cyber-attacks. *The American journal of medicine*, 130(1):27–30.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30.

Google Developers. 2021. Gmail api overview. `https://developers.google.com/gmail/api/guides`.

Milton Friedman. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701.

Sahar Ghannay Ghannay, Benoit Favre, Yannick Esteve in Nathalie Camelin. 2016. Word embedding evaluation and combination. V: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, str. 300–305. European Language Resources Association (ELRA).

Google. 2022. Gmail: Brezplačna, zasebna in varna e-pošta. `https://www.google.com/intl/sl/gmail/about/`, pridobljeno: 2022-01-08.

Andrew Grey, Mark J. Bolland, Nicola Dalbeth, Greg Gamble in Lynn Sadler. 2016. We read spam a lot: prospective cohort study of unsolicited and unwanted academic invitations. *BMJ*, 355.

Brij B. Gupta, Nalin AG Arachchilage in Kostas E. Psannis. 2018. Defending against phishing attacks: taxonomy of methods, current issues and future directions. *Telecommunication Systems*, 67(2):247–267.

Ari Aulia Hakim, Alva Erwin, Kho I Eng, Maulahikmah Galinium in Wahyu Muliady. 2014. Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (tf-idf) approach. V: *2014 6th international conference on information technology and electrical engineering (ICITEE)*, str. 1–4. IEEE.

Irena Koprinska, Josiah Poon, James Clark in Jason Chan. 2007. Learning to classify e-mail. *Information Sciences*, 177(10):2167–2187.

Marcin Kozak, Olesia Iefremova in James Hartley. 2016. Spamming in scholarly publishing: A case study. *Jour-*

nal of the Association for Information Science and Technology, 67(8):2009–2015.

Alexander Kraskov, Harald Stögbauer in Peter Grassberger. 2004. Estimating mutual information. *Physical review E*, 69(6):066138.

Chih-Chin Lai. 2007. An empirical study of three machine learning methods for spam filtering. *Knowledge-Based Systems*, 20(3):249–254.

Songqing Lin. 2013. Why serious academic fraud occurs in China. *Learned Publishing*, 26(1):24–27.

Scott M Lundberg in Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

José Ramon Méndez, Florentino Fdez-Riverola, Fernando Díaz, Eva Lorenzo Iglesias in Juan Manuel Corchado. 2006. A comparative performance study of feature selection methods for the anti-spam filtering domain. V: *Industrial Conference on Data Mining*, str. 106–120. Springer.

David Moher in Anubhav Srivastava. 2015. You are invited to submit.... *BMC medicine*, 13(1):1–4.

Jeffrey Pennington. 2014. Glove: Global vectors for word representation. `https://nlp.stanford.edu/projects/glove/`, pridobljeno: 2022-07-15.

Georgios Sakkis, Ion Androutsopoulos, Georgios Paliouras, Vangelis Karkaletsis, Constantine D. Spyropoulos in Panagiotis Stamatopoulos. 2003. A memory-based approach to anti-spam filtering for mailing lists. *Information retrieval*, 6(1):49–73.

Josep Soler in Andrew Cooper. 2019. Unexpected emails to submit your work: Spam or legitimate offers? the implications for novice english l2 writers. *Publications*, 7(1):7.

Wessel van Lit. 2019. Email spam Kaggle. https://www.kaggle.com/veleon/ham-and-spam-dataset.

Ribut Wahyudi. 2017. The generic structure of the call for papers of predatory journals: A social semiotic perspective. V: *Text-based research and teaching*, str. 117–136. Springer.

Steve Whittaker, Victoria Bellotti in Paul Moody. 2005. Introduction to this special issue on revisiting and reinventing e-mail. *Human–Computer Interaction*, 20(1-2):1–9.

Ian H Witten in Eibe Frank. 2000. *Data mining: practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann.