
**Laboratorij za analizu teksta i inženjerstvo znanja – TakeLab**

Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
Zavod za elektroniku, mikrolektroniku, računalne i inteligentne sustave
Unska 3, 10000 Zagreb, Hrvatska

SVEUČILIŠTE U ZAGREBU
**FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA**

DIPLOMSKI RAD br. 508

# Klasifikacija važnosti poruka elektroničke pošte temeljem govornih činova

Tin Franović

Zagreb, veljača 2013.

**SVEUČILIŠTE U ZAGREBU**
**FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA**
ODBOR ZA DIPLOMSKI RAD PROFILA

Zagreb, 10. listopada 2012.

# DIPLOMSKI ZADATAK br. 508

Pristupnik: **Tin Franović**
Studij: Računarstvo
Profil: Računarska znanost

Zadatak: **Klasifikacija važnosti poruka elektroničke pošte temeljem govornih činova**

Opis zadatka:

Razmjena informacija putem poruka elektroničke pošte zauzima sve veći udio u poslovnoj i osobnoj komunikaciji. Automatska klasifikacija poruka prema važnosti korisniku omogućava kvalitetniju obradu poruka i uštedu vremena. Tipični sustavi za klasifikaciju važnosti poruka temelje se na modelu tzv. vreće riječi. S komunikacijskog aspekta, veći značaj od samih riječi imaju tzv. govorni činovi, odnosno radnje koje su izražene govorom ili pismom (zahtjev, izmjena, isporuka i sl). Budući da govorni činovi neupitno utječu na važnost poruke, pretpostavlja se da klasifikacija važnosti poruke temeljena na govornim činovima može dati bolje rezultate od uobičajene klasifikacije temeljene na riječima.
U okviru diplomskoga rada potrebno je proučiti postojeće postupke za određivanje važnosti poruka elektroničke pošte temeljene na metodama strojnog učenja. Proučiti teoriju govornih činova i pristupe za automatsku klasifikaciju govornih činova metodama nadziranog strojnog učenja. Razraditi postupak za označavanje poruka elektroničke pošte na hrvatskome jeziku govornim činovima. Razraditi postupak za klasifikaciju važnosti poruka elektroničke pošte na hrvatskome jeziku koji će kombinirati klasifikaciju temeljenu na govornim činovima, sadržajnu klasifikaciju i eventualno uporabu dodatnih značajki (npr. vremenskih izraza) dobivenih jednostavnim postupcima ekstrakcije informacija. Ispitati različite algoritme nadziranog strojnog učenja, uključivo generativne i diskriminativne. Provesti postupak odabira značajki. Provesti eksperimentalno vrednovanje točnosti označavanja govornih činova i klasifikacije važnosti te detaljnu analizu značajki i pogrešaka.

Zadatak uručen pristupniku: 12. listopada 2012.
Rok za predaju rada: 8. veljače 2013.

Mentor: Doc.dr.sc. Jan Šnajder

Predsjednik odbora za diplomski rad profila: Prof.dr.sc. Siniša Srbljić

*Veliko hvala mom mentoru doc. dr. sc. Janu Šnajderu na svom prenesenom znanju i pomoći kad god je ona bila potrebna. Hvala Igoru Čanadiju, Alenu Rakipoviću, Matiji Hanževačkom i doc. dr. sc. Stjepanu Grošu što su ustupili svoje poruke elektroničke pošte koje su bile nužne za izradu ovog rada. Hvala Tatjani Perčinlić i Matiji Hanževačkom na pomoći prilikom označavanja skupa podataka.*

*Hvala mojim roditeljima, Ljiljani i Ivi-Bruni, na apsolutnoj i bezuvjetnoj podršci za vrijeme mog studija, bez njih ništa od ovog ne bi bilo moguće. I na kraju, hvala Tatjani na svim prelijepim trenucima zbog kojih će mi ovo razdoblje života ostati u divnom sjećanju.*

# SADRŽAJ

# 1. Uvod

U današnje vrijeme elektronička pošta nametnula se kao jedan od najkorištenijih komunikacijskih medija, kako u privatnom životu tako i u akademskom i poslovnom svijetu. Posljednja istraživanja ukazuju na to da većina poslovnih korisnika elektroničke pošte provede i do dva sata dnevno čitajući, pišući i odgovarajući na poruke. Takav pritisak sve se više dovodi u vezu sa stresom na radnom mjestu (Dabbish i Kraut, 2006) te može drastično umanjiti produktivnost korisnika. Stoga, očita je potreba za sustavima za automatsku klasifikaciju poruka elektroničke pošte koji bi mogli skratiti vrijeme koje korisnik troši na čitanje i organiziranje pretinca poruka. Klasifikacija dolazećih poruka pruža korisniku informaciju o procijenjenoj važnosti ili sadržaju poruke prije nego što korisnik otvori i pročita poruku. To korisniku omogućuje da se posveti samo onim porukama koje su istaknute kao važne ili one čiji sadržaj smatra zanimljivim. Klasifikacija poruka elektroničke pošte prvotno je postala popularna u obliku filtriranja neželjene (engl. *spam*) pošte, gdje se iz korisnikovog ulaznog pretinca uklanjaju poruke neželjenog sadržaja te ih se premješta u poseban pretinac. Alternativa takvom označavanju i izdvajanju poruka je označavanje poruka na temelju važnosti kao što je učinjeno u servisu Google Mail gdje se poruke koje su smatrane važnima premještaju u poseban pretinac pod nazivom *Priority Inbox*. To je mnogo teži problem, s obzirom na to da je važnost često vrlo subjektivna te ovisi o percepciji korisnika i kontekstu. U sklopu ovog rada pokazat će se kako postoji način za objektivnu i automatsku klasifikaciju važnosti poruka elektroničke pošte, neovisan o kontekstu ili stavu korisnika.

Obje metode koje su prethodno navedene organiziraju poruke na osnovu procijenjene važnosti. S druge strane, moguće je poruke označavati i na osnovu njihovog sadržaja kako bi se korisniku pružio skraćeni pregled sadržaja poruke gdje ima mnogo manje prostora za subjektivnost. Jedan od načina označavanja poruka na temelju sadržaja temelji se na govornim činovima (engl. *speech acts*). Govorni činovi (Searle, 1965) u kontekstu poruka elektroničke pošte pružaju učinkovit način za sažimanje sadržaja i namjene poruke. Korisniku se tada daje mogućnost da na osnovu sadržaja procijeni kojim porukama će pridati veću važnosti. U ovom se radu označavanju po-

ruka na temelju govornih činova pristupa u okviru problema klasifikacije teksta s više oznaka (engl. *multilabel text classification*) i prelaže se rješenje u obliku nadziranog strojnog učenja.

Rad se temelji na pokušaju objedinjavanja klasifikacije na osnovu važnosti i na osnovu sadržaja na tako da se oznake dobivene klasifikacijom na osnovu sadržaja iskoriste kao značajke prilikom klasifikacije na temelju važnosti. Koliko je autoru rada poznato, takav način klasifikacije na temelju važnosti dosad nije proveden niti za jedan svjetski jezik te stoga ovaj rad pruža uvid u jedno potpuno neistraženo područje automatske klasifikacije poruka. Pretpostavlja se da će oznake govornih činova, kao sažeti prikaz sadržaja poruke, doprinijeti klasifikaciji poruka na temelju važnosti koja se temelji na "vreći riječi" (engl. *bag of words*) kao skupu značajki. Zbog toga je prvo razvijen sustav za označavanje poruka na temelju govornih činova te je onda dobiveni skup klasifikatora primijenjen za crpljenje informacija kod klasifikacije prema važnosti.

Rad sadrži detaljne pokuse u kojima je vrednovano šest tipova klasifikatora za označavanje govornih činova te pet tipova klasifikatora za klasifikaciju važnosti poruke. Za klasifikaciju govornih činova koristile su se tri vrste značajki i tri razine (poruka, odlomak i rečenica), dok je za klasifikaciju na temelju važnosti korištena samo razina poruke uz dvije vrste značajki. Provedeno je i usporedno vrednovanje klasifikatora koji koriste govorne činove sa klasifikatorima koji se temelje samo na osnovnom skupu značajki kako bi se procijenio utjecaj oznaka govornih činova na klasifikaciju prema važnosti. Svi pokusi provedeni su nad skupom poruka na hrvatskom jeziku, prikupljenim i označenim za potrebe ovog rada.

Rad je strukturiran kako slijedi. U poglavlju 2 opisana su trenutna dostignuća i suvremeni trendovi u klasifikaciji i označavanju elektroničke pošte na temelju govornih činova i na temelju važnosti. Poglavlje 3 uvodi definiciju pojma govornih činova te taksonomiju činova koji će biti korišteni u radu. U poglavlju 4 opisuje se postupak prikupljanja i označavanja skupa podataka. Poglavlje 5 opisuje postupak učenja klasifikatora, dok poglavlje 6 navodi rezultate vrednovanja oba tipa klasifikatora. Naposljetku, poglavlje 7 donosi zaključak nakon kojeg slijedi pregled korištene literature.

# 2. Klasifikacija poruka elektroničke pošte

Kao što je navedeno u uvodu, broj dnevno poslanih i primljenih poruka elektroničke pošte rapidno raste posljednjih godina te je stoga potreba za sustavima koji bi korisnicima olakšavali čitanje poruka sve izraženija. No, područje koje se bavi klasifikacijom elektroničke pošte i nije toliko novo.

Prvim pokušajima razvoja sustava koji klasificiraju poruke elektroničke pošte smatraju se sustavi poput *The Coordinator* (Winograd, 1987). Glavna manjkavost takvih sustava bila je potreba da se poruke označavaju ručno, odnosno da korisnik ručno doda određene anotacije koje se koriste pri klasifikaciji. To je nepraktično jer korisnici uglavnom ne šalju poruke koje su snažno strukturno određene i prikladne za takvo označavanje. Slični sustavi postoje i danas, poput primjerice sustava *Semanta* (Scerri et al., 2009), koji korisnicima omogućuje označavanje dijelova poruka iz grafičkog sučelja.

Radovi na kojima se temelji ovaj rad više su usredotočeni na automatiziranje klasificiranja elektroničke pošte te najčešće u tu svrhu koriste tehnike strojnog učenja. S obzirom na to da se radi o području vrlo široke primjene, takvi automatizirani sustavi počinju se sve više koristiti. Primjerice, Google je početkom rujna 2010. godine u svoju uslugu Google Mail uveo dodatak pod nazivom *Priority Inbox* koji se služi tehnikama strojnog učenja kako bi poruke klasificirao prema važnosti. Dalianis et al. (2011) za potrebe švedske vlade rade na sustavu koji klasificira poruke elektroničke pošte prema često postavljanim pitanjima te korisnicima šalje odgovarajući automatizirani odgovor.

## 2.1. Klasifikacija na temelju govornih činova

Klasifikacija na temelju govornih činova je jedan od zanimljivih zadataka i izazova u području obrade prirodnog jezika (engl. *natural language processing, NLP*). Iz te

perspektive, klasifikacija govornih činova je zanimljiva posebice u području interakcije čovjeka i računala na temelju dijaloga. Uspješni sustavi za analizu dijaloga sposobni su razumjeti namjere pošiljatelja te sadržaj poruke koju on nastoji prenijeti. Takva analiza pošiljateljevih namjera uglavnom se temelji na proučavanju i detekciji govornih činova.

Projekt *Clarity* (Finke et al., 1998) je jedan od prvih radova u kojima se govorni činovi koriste u pokušaju razumijevanja dijaloga. Cilj projekta bio je otkriti tri strukturalne razine u telefonskim razgovorima na španjolskom jeziku: govorne činove, dijaloške igre te segmente razgovora. Sustav AutoTutor (Marineau et al., 2000) je računalni tutor na engleskom jeziku koji je osjetljiv na govorne činove izražene u prethodnom koraku dijaloga što mu omogućuje izbor sljedeće radnje na temelju namjere govornika. Keizer (2001) je oblikovao konverzacijskog agenta za nizozemski jezik koji govorne činove interpretira na probabilistički način. Serafin et al. (2003) koristi latentnu semantičku analizu (engl. *Latent Semantic Analysis, LSA*) kako bi klasificirao govorne činove iz korpusa dijaloga na španjolskom jeziku. Louwerse i Crossley (2006) koriste algoritme koji rade s n-torkama za klasifikaciju govornih činova u dijalozima na engleskom jeziku koji se bave temom rekonstrukcije lokacijskih mapa. U svojem radu, Kim et al. (2006) provode detaljnu analizu studentskih diskusija na internetskim forumima s ciljem da pomoću govornih činova identificiraju uloge sudionika u razgovoru te pronađu poruke i teme u kojima postoje nedoumice ili pitanja. U radu se također pokazuje kako proučavanje govornih činova može pomoći prilikom automatskog odgovaranja na pitanja. Na sličan način Ravi i Kim (2007) nastoje u komunikaciji temeljenoj na porukama na internetskim forumima uz pomoć govornih činova i značajki temeljenih na n-torkama pronaći pitanja koja zahtijevaju posebnu pažnju nastavnika ili poruke koje sadrže odgovore na navedena pitanja. Namjeru odnosno svrhu poruka elektroničke pošte sa gledišta govornih činova razmatraju i Dabbish et al. (2005) pokušavajući definirati žanrove poruka koristeći podskup govornih činova. Konačan cilj im je identifikacija glavnog komunikacijskog cilja poruke kao i žanra kojem ta poruka pripada.

Kao motivacija za ovaj rad poslužili su radovi koji koriste govorne činove kao pomoć u klasifikaciji poruka elektroničke pošte na temelju sadržaja. Primjerice, Cohen et al. (2004) razvili su sustav za klasifikaciju poruka elektroničke pošte temeljen na nadziranim metodama strojnog učenja i vlastitoj taksonomiji govornih činova. U kasnijem radu, Carvalho i Cohen (2006) koriste lingvističke aspekte problema sadržajne klasifikacije poruka elektroničke pošte tako što kombiniraju predobradu poruka i ekstrakciju značajki na temelju n-torki kako bi poboljšali klasifikaciju.

Ovaj rad u pogledu klasifikacije govornih činova slijedi većinom ideje koje su iz-

nijeli Cohen et al. (2004) te ih prilagođava korištenju pri klasifikaciji poruka na hrvatskom jeziku. Neke od ideja koje su uzete u obzir kasnije su odbačene zbog nedostatka potrebnih resursa i alata prilagođenih hrvatskom jeziku.

## 2.2. Klasifikacija na temelju važnosti

Što se tiče klasifikacije poruka elektroničke pošte prema važnosti, ona se do sad uglavnom isključivo svodila na tehnike i metode izdvajanja neželjenih (engl. *spam*) poruka od ostatka poruka u ulaznom pretincu korisnika. Budući da je razvojem elektroničkog oglašavanja i programskih paketa koji omogućavaju lako slanje velikog broja poruka neželjena pošta postala velik problem pogotovo za poslovne korisnike, razvijen je velik broj različitih rješenja koja pokušavaju automatski izdvojiti neželjene poruke koristeći metode strojnog učenja (Zhang et al., 2004; Sakkis et al., 2001, 2003; Yu i Xu, 2008).

S druge strane, problem procjene važnosti poruka elektroničke pošte i njihovo isticanje u ulaznom pretincu nije toliko raširen te je stoga broj relevantnih znanstvenih radova na tu temu relativno malen. Najpoznatiji primjer automatskog izdvajanja poruka važnih za korisnika jest već spomenuti servis *Priority Inbox* integriran u Googleov klijent elektroničke pošte (Aberdeen et al., 2010). Sustav koristi logističku regresiju zajedno sa značajkama o načinu na koji korisnik reagira na određenu vrstu poruka kako bi procijenio vjerojatnost da će korisnik učiniti neku radnju nad trenutno promatranom porukom. Te značajke se mogu odnositi na pošiljatelja poruke, broj otvaranja (čitanja) sličnih poruka ili odgovaranja na iste, ključne riječi koje se nalaze u poruci i slično. Sustav podržava i *on–line* učenje, tako da se naučena pravila mijenjaju zajedno s korisnikovim radnjama, pa je model koji određuje važnost poruke strogo personaliziran za svakog korisnika. Yoo et al. (2009) temelje svoje istraživanje važnosti poruka na socijalnim mrežama svakog korisnika kako bi stvorili bogat skup značajki temeljen na društvenim ulogama koje pojedinci imaju iz perspektive korisnika. Također, razvili su i polu-nadzirani algoritam strojnog učenja koji propagira oznake važnosti sa poruka u skupu za učenje na ispitni skup. Hart (2008) je patentirao sustav za procjenu važnosti poruka elektroničke pošte koji prilikom zaprimanja poruke uzima u obzir učestalost interakcije između pošiljatelja i primatelja te na osnovu toga pridjeljuje poruci oznaku važnosti. Ta oznaka se kasnije koristi kako bi sustav automatski odabrao najpogodniju radnju vezanu uz tu poruku. Ranije su Scannell et al. (1994) patentirali sličan sustav za automatsko raspoređivanje poruka prema prioritetu, no navedeni sustav nije imao mogućnost učenja već se u potpunosti temeljio na pravilima koje bi korisnici morali eksplicitno zadati. Pravila su se odnosila na ključne riječi u tekstu koje poruku čine

više ili manje važnom i koje je korisnik morao navesti. Sustav tad automatski svakoj pristigloj poruci pridjeljuje procjenu važnosti i odlučuje u koji pretinac ju smjestiti.

# 3. Govorni činovi

> *Rečeno na banalan način, govorni činovi su upravo ono što nam sam termin govori – radi se o činovima, radnjama koje činimo najčešće pomoću govora. (Blečić, 2010)*

Kao što uvodni citat kaže, govorni činovi označavaju radnje koje činimo pomoću govora, bilo da se radi o razgovoru ili o pisanoj komunikaciji, kao što je slučaj s porukama elektroničke pošte. Kako navodi Searle (1965), govorni se činovi karakteristično izvode izgovaranjem glasova ili stvaranjem oznaka. Za razliku od običnog izgovaranja glasova ili stvaranja oznaka, govorni činovi su ilokucijski činovi koji imaju određeno značenje koje govornik želi prenijeti na slušatelja. To značenje može biti bilo što, primjerice pozdrav, zamolba, zahvala ili zahtjev.

Govorne činove moguće je pronaći u svakoj poruci elektroničke pošte, jer bez njih sama poruka ne bi imala smisla, stoga zbog svoje učestalosti, raznovrsnosti i informativnosti čine dobru osnovu za klasifikaciju poruka elektroničke pošte.

U nastavku je dana taksonomija govornih činova korištena u ovom radu.

## 3.1. Taksonomija govornih činova

Govorne činove možemo prema Searleu na osnovu tri temeljne dimenzije (ilokucijska svrha, smjer prilagodbe i psihološka karakteristika) i devet kriterija podijeliti u pet kategorija:

- asertivi
- direktivi
- ekspresivi
- komisivi
- deklaracije

*Asertivi* su govorni činovi kojima se govornik obvezuje na istinitost tvrdnje te su podložni procjeni prema odnosu istinito – neistinito. U ovom radu koriste se sljedeći asertivi:

- izmijeniti
- predvidjeti
- zaključiti

*Direktivi* se koriste kako bi govornik potaknuo slušatelja da nešto učini. Pokazalo se kako bi označavanje direktiva bilo vrlo korisno s obzirom da je iz njih vidljivo koje poruke donose novi zadatak za primatelja. U ovom radu koriste se sljedeći direktivi:

- zahtijevati
- podsjetiti
- predložiti

*Ekspresivi* su najčešće korišteni za izražavanje nekog psihološkog stanja, primjerice zahvale, čestitke, dobrodošlice. Oni su vrlo česti u porukama elektroničke pošte, no nisu od prevelike informativne važnosti. Korišteni su sljedeći ekspresivi:

- ispričati se
- pozdraviti
- zahvaliti

*Komisivi* su govorni činovi kojima se govornik obvezuje na neko buduće djelovanje. Komisivi korišteni u ovom radu su:

- obvezati se
- odbiti
- upozoriti

*Deklaracije* su obilježene time da njihovo uspješno izvođenje dovodi do toga da sadržaj odgovara stvarnosti. U nekim slučajevima dolazi do preklapanja područja deklarativa i asertiva, pogotovo kad je potrebno da neki autoritet deklarativom utvrdi neku činjenicu. U ovom radu korišten je samo jedan deklarativ, a to je *isporučiti*.

# 4. Prikupljanje i označavanje skupova podataka

U sklopu ovog rada bilo je potrebno je naučiti i evaluirati klasifikatore koji će se koristiti za označavanje poruka elektroničke pošte. Za učenje klasifikatora i njihovo vrednovanje potreban je označen skup podataka (engl. *dataset*) iz kojeg se odabiru pozitivni i negativni primjeri za učenje te primjeri za testiranje. Mnogi takvi skupovi podataka dostupni su na Internetu, poput primjerice skupa poruka izmjenjivanih u tvrtki Enron (Klimt i Yang, 2004), no s obzirom da se u ovom radu obrađuje označavanje poruka elektroničke pošte na hrvatskom jeziku, navedeni skupovi podataka nisu iskoristivi. Budući da u trenutku izrade ovog rada prikladnih skupova podataka nije bilo, pristupilo se prikupljanju i obradi vlastitog skupa podataka. Navedeni postupak možemo podijeliti na dva osnovna dijela: prikupljanje podataka i označavanje. Budući da se u ovom radu razmatraju dva tipa klasifikatora (temeljen na govornim činovima i temeljen na važnosti poruke), označavanje poruka bilo je potrebno provesti dvaput, dok se u oba slučaja koristio isti skup poruka.

## 4.1. Prikupljanje poruka elektroničke pošte

Prikupljanje podataka sastoji se od pribavljanja potrebnih podataka iz različitih izvora te njihove obrade kako bi se pretvorili u oblik pogodan za označavanje i daljnje korištenje. Samo prikupljanje može se ostvariti na jedan od tri načina: korištenje već dostupne arhive objavljene na Internetu (Klimt i Yang, 2004), praćenje komunikacije na nekom projektu kako bi se prikupili podaci (Cohen et al., 2004) te korištenje osobne arhive elektroničke pošte određenog broja volontera. U ovom slučaju korišten je potonji pristup.

Kako se elektronička pošta uglavnom koristi za osobnu korespondenciju te su mnogi korisnici vrlo osjetljivi na privatnost vlastite elektroničke pošte, u ovakvim je slučajevima nužno pribaviti suglasnost osoba koje su uključene u komunikaciju kako

bi se izbjegla bilo kakva povreda tajnosti. Stoga je odlučeno da se u obzir uzme samo poslana pošta volontera, s obzirom na to da je u tom slučaju autor poruke već pristao na njezino korištenje. Loša strana tog pristupa jest u konačnici relativno malen uzorak osoba uključenih u komunikaciju, što zbog razlika u stilu pisanja i korištenom rječniku može utjecati na sposobnost generaliziranja klasifikatora.

Kako bi se podaci što lakše prikupili te kako bi se olakšala pretvorba u oblik pogodan za označavanje uzeti su u obzir volonteri sa korisničkim računom Google Mail te je svatko od njih koristio sljedeći postupak kako bi pripremio podatke:

1. omogućiti IMAP pristup[1] svom korisničkom računu,
2. označiti poruke koje se žele koristiti,
3. odabrati opciju "izvezi u tekstni format" (engl. *export to plain text*) u kojoj se svaka poruka posebno sprema,
4. dobiven direktorij arhivirati u format .zip.

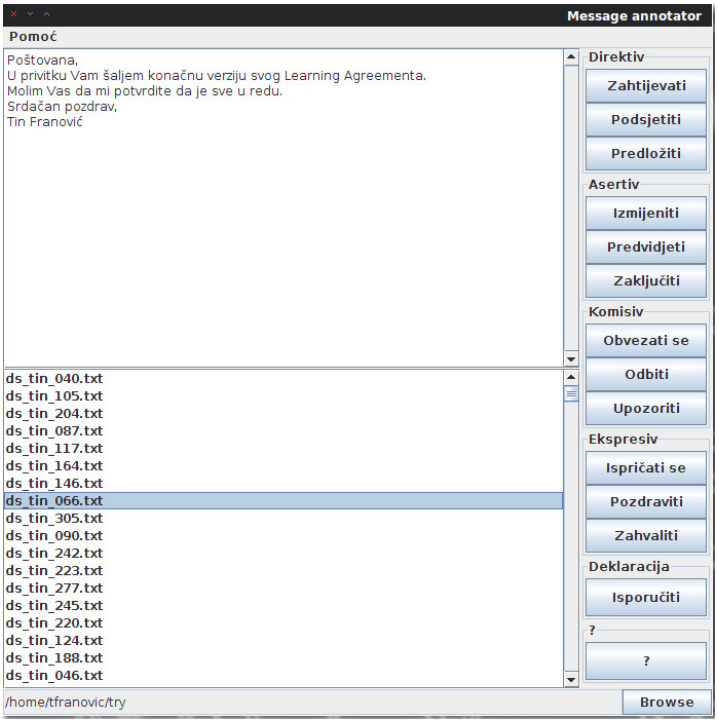
Tako je dobivena po jedna tekstna datoteka za svaku poruku elektroničke pošte koju su volonteri stavili na raspolaganje. Osnovni skup podataka čine poruke ustupljene od strane četiri volontera. Uz poruke prikupljene od korisnika servisa Google Mail, manji dio skupa podataka čine i poruke razmijenjene u sklopu projekta IKEV2 na Fakultetu elektrotehnike i računarstva Sveučilišta u Zagrebu. U slučaju tog podskupa poruka, nije bilo moguće automatski odvojiti poruke u zasebne tekstne datoteke, već je to učinjeno ručno.

Nakon odvajanja poruka, pokrenuta je skripta napisana u programskom jeziku Python koja iz svake poruke uklanja zaglavlje te citirane dijelove prošlih poruka (ukoliko postoje). Nakon toga, poruke su spremne za označavanje. Ukupno je prikupljeno 3059 poruka elektroničke pošte.

## 4.2. Označavanje govornih činova

Cilj označavanja skupa podataka na temelju govornih činova jest pridjeljivanje oznaka govornih činova pojedinim dijelovima poruka elektroničke pošte kako bi se ti dijelovi kasnije mogli koristiti za stvaranje pozitivnih i negativnih primjera za učenje klasifikatora. Kako bi se taj postupak olakšao i ubrzao, razvijena je jednostavna aplikacija

[1] IMAP – *Internet message access protocol*, protokol za pristup elektroničkoj pošti koji ne sprema čitav sadržaj pretinca na disk.




**Slika 4.1:** Aplikacija za označavanje skupa podataka

u programskom jeziku Java (slika 4.1) koja korisniku omogućava označavanje dijela poruke te klikom na odgovarajući gumb stvara novu oznaku (anotaciju).

Anotacija je posebno definirana datoteka koja sadrži meta informacije i tekst koji je prepoznat kao govorni čin. Meta informacije nalaze se na početku datoteke i označene su znakom "@". Podaci koje sadrže su:

- `@parent`: naziv datoteke koja sadrži poruku iz koje je nastala anotacija,
- `@from` i `@to`: oznaka od kojeg do kojeg znaka originalne datoteke se proteže anotacija,
- `@annot`: prepoznat govorni čin.

Nakon meta informacija slijedi tekst koji je prepoznat kao govorni čin. Primjer dobivenih anotacija vidljiv je na slici 4.2.


**(a)** Isporučiti **(b)** Pozdraviti **(c)** Zahtijevati

**Slika 4.2:** Primjeri anotacija

### 4.2.1. Uzajamno slaganje označivača

U svrhu smanjenja utjecaja subjektivnosti označivača na oznake govornih činova (ali i ubrzanja procesa označavanja), posao je podijeljen između dvije osobe. Oba označivača dobila su jednak broj poruka za označavanje te je svatko dodatno označio $15\%$ poruka koje je označavao drugi označivač (15-postotno preklapanje). Te su se poruke koristile kako bi se izračunala mjera podudarnosti između dva označivača (engl. *Inter-Annotator Agreement – IAA*).

Nakon što su oba označivača završila svoj posao, načinjena je usporedba oznaka u preklapajućim dijelovima kako bi se izračunala brojčana mjera podudarnosti među označivačima – $\kappa$-statistika (Cohen, 1960; Carletta, 1996) za svaki govorni čin posebno.

$\kappa$-statistika se računa prema formuli:

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)},$$

gdje $P(a)$ označava izmjerenu podudarnost između dva označivača (vjerojatnost da su se označivači složili oko odluke), dok $P(e)$ označava očekivanu (engl. *expected*) vjerojatnost da su se označivači slučajno složili oko odluke. Ukoliko je $\kappa = 1$, tada postoji potpuno slaganje među označivačima, dok se skup podataka za većinu primjena smatra pogodnim za upotrebu ukoliko je $\kappa \geq 66{,}67\%$. Tablica 4.2 prikazuje dobivene rezultate zasebno za svaki govorni čin.

### 4.2.2. Daljnja obrada

Nakon označavanja skupa podataka i izračuna mjere podudarnosti između označivača, potrebno je za svaki govorni čin odrediti pozitivne i negativne primjere. Budući da se klasifikacija na temelju govornih činova radila na tri razine (poruka, odlomak, rečenica), bilo je potrebno stvoriti zasebne skupove pozitivnih i negativnih primjera za svaku razinu. U tu je svrhu razvijena skripta u programskom jeziku Python koja dijeli svaku poruku na dijelove ovisno o razini koja se koristi te provjerava nalazi li se u tom dijelu neka od anotacija. Pomoću meta informacija za svaku anotaciju možemo odrediti nalazi li se anotacija za neki govorni čin u dijelu koji promatramo, te tada taj dio smatramo pozitivnim primjerom za učenje klasifikatora, dok u suprotnom taj dio svrstavamo u negativne primjere. Ovakvim pristupom pronađeno je mnogo više negativnih primjera nego pozitivnih za svaki govorni čin, pa je bilo potrebno ograničiti broj negativnih primjera tako da u konačnici imamo podjednak broj pozitivnih i negativnih primjera za svaki govorni čin na svakoj razini.

Nakon što su pozitivni i negativni primjeri obrađeni, bilo je potrebno odrediti koji će govorni činovi biti odbačeni, a koji će se koristiti u konačnoj inačici označivača. Kriteriji zbog kojih su govorni činovi mogli biti odbačeni su:

- neinformativnost – govorni čin nije previše koristan za konačnog korisnika (primjer: POZDRAVITI, ZAHVALITI),
- slaba zastupljenost – broj primjera za učenje je premalen (primjer: odbiti, upozoriti),
- neslaganje označivača – niska vrijednost $\kappa$-statistike za govorni čin (primjer: zaključiti, predvidjeti).

Na osnovu navedenih kriterija početni popis od 13 govornih činova skraćen je na skup od šest činova. Izbačeni govorni činovi su: ISPRIČATI_SE, ODBITI, POZDRAVITI, PREDVIDJETI, UPOZORITI, ZAHVALITI i ZAKLJUČITI. Konačan skup korištenih govornih činova čine: ISPORUČITI, IZMIJENITI, OBVEZATI_SE, PODSJETITI, PREDLOŽITI i ZAHTIJEVATI.

### 4.2.3. Statistike skupa podataka

U nastavku su navedeni neki osnovni statistički podaci o skupu podataka korištenom za učenje klasifikatora govornih činova.

#### Osnovne statistike

Skup podataka sadrži ukupno $1337$ poruka elektroničke pošte u kojima je označeno 4498 govornih činova. Ukupno se u označenim porukama nalazi $76760$ riječi u $4468$ odlomaka.

U Tablici 4.1 nalaze se podaci o broju označenih primjera za učenje za svaki govorni čin koji je korišten prilikom učenja klasifikatora, podijeljeni po razinama.

| | Poruka | | Odlomak | | Rečenica | |
|---|---|---|---|---|---|---|
| | + | − | + | − | + | − |
| ISPORUČITI | 293 | 263 | 350 | 331 | 376 | 384 |
| IZMIJENITI | 104 | 102 | 131 | 131 | 178 | 176 |
| OBVEZATI_SE | 288 | 259 | 362 | 347 | 469 | 452 |
| PODSJETITI | 29 | 30 | 38 | 39 | 42 | 43 |
| PREDLOŽITI | 240 | 219 | 346 | 334 | 456 | 442 |
| ZAHTIJEVATI | 679 | 421 | 989 | 865 | 1384 | 1228 |

**Tablica 4.1:** Broj pozitivnih i negativnih primjera za svaki govorni čin, po razinama

**Inter-Annotator Agreement - $\kappa$–statistika**

**Tablica 4.2:** $\kappa$-statistika za sve govorne činove

| Govorni čin | $\kappa$ –statistika (%) |
|---|---|
| ISPORUČITI | 79,17 |
| ISPRIČATI_SE | 85,55 |
| IZMIJENITI | 71,44 |
| OBVEZATI_SE | 85,13 |
| ODBITI | 0,00 |
| PODSJETITI | 74,67 |
| POZDRAVITI | 77,92 |
| PREDLOŽITI | 54,41 |
| PREDVIDJETI | 26,67 |
| UPOZORITI | 17,44 |
| ZAHTIJEVATI | 58,85 |
| ZAHVALITI | 94,88 |
| ZAKLJUČITI | 0,53 |

## 4.3. Označavanje na temelju važnosti

Označavanje skupa podataka na temelju važnosti ima kao svrhu pridjeljivanje oznake važnosti prikupljenim porukama. Oznaka važnosti može biti binarna (važno/nevažno)

ili se sastojati od više kategorija, odnosno razina važnosti. Budući da je konačan cilj klasifikatora isticanje važnih poruka, odnosno podjela skupa poruka na važne i nevažne, označavanje je u ovom radu ograničeno na slučaj binarnih oznaka, odnosno oznaka važno ili nevažno.

Važnost poruke elektroničke pošte može biti veoma subjektivna te vrlo ovisiti o vremenskom kontekstu kao i kontekstu razgovora u sklopu kojeg je poruka poslana. Također, pošiljatelj poruke odnosno odnos primatelja prema pošiljatelju može utjecati na primateljevu percepciju važnosti pojedine poruke. Primjer kontekstno zavisne poruke bila bi poruka sadržaja "Slažem se s napisanim." bez konkretnog osvrtanja na sadržaj prethodne poruke. Na taj način kontekst odnosno sadržaj prethodne poruke, a ovisno o situaciji i sam pošiljatelj, određuju važnost promatrane poruke. U sklopu ovog rada, pošiljatelj i kontekst poruke su zanemareni te je cilj bio postići što objektivniju podjelu na važne i nevažne poruke. Poruke koje nije bilo moguće objektivno označiti (primjerice zbog kontekstne zavisnosti), izbačene su iz konačnog skupa podataka.

Kako bi se postigla objektivnost prilikom označavanja poruka prema važnosti, bilo je potrebno definirati skup pravila koja određuju na koji će se način pojedina poruka klasificirati. U tu svrhu je, uz manje izmjene, iskorišten skup pravila i kriterija korišten u (Dabbish et al., 2005). Konačan skup korištenih kriterija sastoji se od:

1. važnost sadržaja poruke (bez uloge konteksta) za pošiljatelja,
2. važnost sadržaja poruke za primatelja,
3. postojanje istaknutih zadataka ili rokova u poruci,
4. količina posla koji poruka zahtijeva,
5. mijenja li poruka na neki način trenutnu situaciju u pogledu komunikacije između korisnika,
6. ukoliko primatelj propusti pročitati poruku ili ju izbriše, hoće li ostati bez važne informacije.

Odlučeno je da će se označavanje važnosti raditi isključivo na razini poruke, s obzirom na to da je konačan cilj podjela poruka prema važnosti, bez podjele segmenata unutar poruke. Također, koristeći navedene upute, mnogo je lakše objektivno procijeniti važnost pojedine poruke nego što bi to bio slučaj za odlomke ili rečenice. Za potrebe označavanja prilagođena je aplikacija opisana u odjeljku 4.2 tako da omogućuje označavanje cijelih poruka i razvrstavanje u jednu od tri moguće kategorije: *važno*,

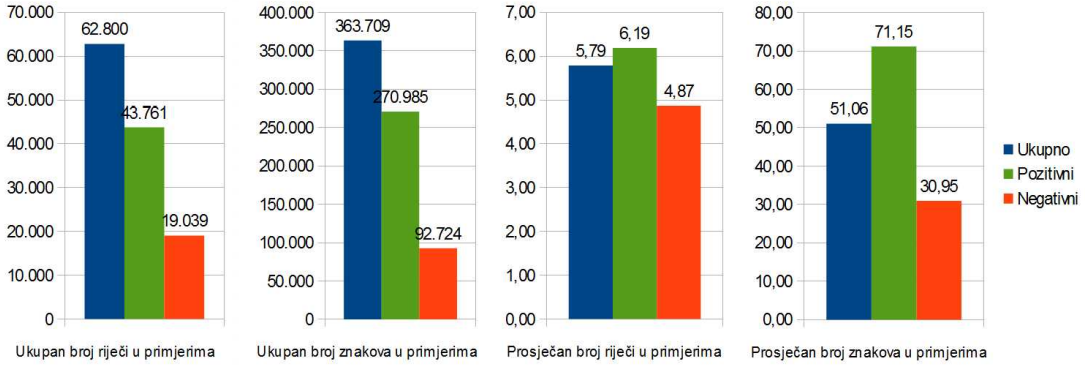
*nevažno* ili *nije moguće objektivno procijeniti*. Treća kategorija služi za poruke kojima označivač nije mogao procijeniti važnost na temelju danih kriterija već je važnost uvjetovana primjerice pošiljateljem ili kontekstom razgovora i tako označene poruke su isključene iz daljnjeg razmatranja. Označavanje cijelog skupa podataka proveo je jedan označivač.

### 4.3.1. Uzajamno slaganje označivača

Kao što je već navedeno, označavanje cijelog skupa poruka proveo je jedan označivač. S obzirom na to da je procjena važnosti poruke često pod utjecajem subjektivnosti označivača, pogotovo ako mu je poznat kontekst, bilo je potrebno procijeniti objektivnost označivača. Procjena objektivnosti je važna i zato što daje odgovor na pitanje je li važnost poruke uopće objektivna kategorija, odnosno je li moguće u potpunosti isključiti subjektivnost i kontekst iz označavanja, a pritom dobiti kvalitetnu podjelu poruka prema važnosti.

Kako bi se procijenila objektivnost označivača, iskorištena je metoda opisana u odjeljku 4.2.1, odnosno računanje $\kappa$-statistike koja izražava mjeru slaganja dva označivača. Za drugog označivača uzet je volonter kojemu je pripremljen podskup originalnog skupa podataka koji se sastojao od $25\%$ označenih pozitivnih primjera i $25\%$ negativnih. Budući da je broj pozitivnih i negativnih primjera jednak, takav odnos zadržan je i u podskupu koji je ponovno označavan. Volonteru su dane upute opisane u prethodnom odjeljku i zadatak da podijeli poruke u jednu od dvije kategorije. Nakon označavanja, primijećen je veći broj pozitivno označenih primjera u odnosu na negativne te su volonteru dati primjeri označavanja nekih poruka koje se ne nalaze u skupu za ponovno označavanje s objašnjenjem zašto je prvotni označivač pojedinu poruku razvrstao kao važnu ili nevažnu prema prethodno navedenih šest kriterija. To je učinjeno kako bi se označivača upozorilo na mogućnost da je neke kontekstno zavisne poruke označio kao važne, iako one možda nisu zadovoljavale navedene kriterije. Nakon toga, volonter je ponovno označio isti skup poruka. Primijećeno je da je nakon ponovnog označavanja smanjio broj označenih važnih poruka, što može biti znak da je volonter doista neke kontekstno zavisne poruke svrstao u važne.

Za oba slučaja označavanja izračunata je $\kappa$–statistika odnosno mjera slaganja označivača. U prvom slučaju, samo uz korištenje popisa pravila, $\kappa = 0.818$, dok je u drugom slučaju gdje je bilo više utjecaja na volontera, $\kappa = 0.816$. Vidljivo je da je uzajamno slaganje označivača vrlo visoko u oba slučaja (mnogo više od graničnih $0.67$), što potvrđuje da se označeni podaci mogu koristiti u daljnjem radu. Također,




**Slika 4.3:** Statistike skupa podataka za klasifikaciju važnosti

moguće je zaključiti i kako dodatne upute volonteru nisu poboljšale uzajamno slaganje, odnosno oznake koje je pridijelio volonter su doista objektivne. Time je potvrđeno da je moguća objektivna procjena važnosti poruka elektroničke pošte.

### 4.3.2. Statistike skupa podataka

Iz originalnog skupa poruka $1096$ ih je označeno kao važno, a $615$ kao nevažno. Kako bi se ujednačio broj pozitivnih i negativnih primjera i time uklonila pristranost klasifikatora, iz skupa pozitivnih primjera slučajnim je odabirom izdvojeno $615$ poruka. Time je veličina skupa primjera smanjena na ukupno $1230$ poruka, uz jednak broj pozitivnih i negativnih primjera.

Prikupljeni skup podataka, nakon ujednačavanja broja pozitivnih i negativnih primjera, sastoji se od $62.800$ riječi, odnosno $363.709$ znakova. Prosječna duljina riječi je $5{,}79$ znakova, a prosječan broj riječi po primjeru je $51{,}06$.

Pozitivni primjeri sastoje se ukupno od $43.761$ riječi, odnosno $270.985$ znakova. Time sadrže gotovo $70\%$ riječi koje se nalaze u skupu podataka. Prosječna duljina riječi je $6{,}19$ znakova, dok u svakom primjeru ima u prosjeku $71{,}15$ riječi. Negativni primjeri sačinjeni su od $19.039$ riječi sa $92.724$ znakova. Prosječna duljina riječi je $4{,}87$ znakova, a u svakom negativnom primjeru nalazi se prosječno $30{,}95$ riječi. Slika 4.3 grafički prikazuje dobivene rezultate.

Iz navedenih podataka vidljivo je da su važne poruke u pravilu bogatije riječima i sadrže dulje riječi, iako postoje suprotni primjeri poput poruka u kojima se primatelju isporučuje neki privitak (engl. *attachment*) i takve se poruke najčešće sastoje od jedne do dvije rečenice.

# 5. Učenje klasifikatora

Nakon prikupljanja i obrade skupa podataka, slijedi najvažniji dio rada koji se sastoji od učenja klasifikatora. Proces učenja klasifikatora podijeljen je na dva dijela, gdje se u prvom dijelu uče i vrednuju klasifikatori koji označavaju govorne činove, dok se u drugom dijelu ti klasifikatori koriste kako bi se proširio skup značajki za klasifikaciju na temelju važnosti. Na taj je način moguće dobiti uvid u to kako klasifikacija na temelju govornih činova utječe na konačnu klasifikaciju poruka elektroničke pošte prema važnosti.

U okviru postupka učenja klasifikatora bilo je potrebno odabrati značajke koje će se koristiti pri klasifikaciji i na temelju odabranih značajki stvoriti primjere za učenje. Ti primjeri za učenje koriste se kasnije za učenje klasifikatora i optimizaciju parametara. Nakon što su parametri klasifikatora optimirani, koristi se ispitni skup podataka kako bi se rezultati mogli vrednovati. U slučaju klasifikacije govornih činova, rezultati vrednovanja nam omogućuju odabir najpogodnijeg klasifikatora za svaki govorni čin.

## 5.1. Programska potpora

### 5.1.1. Programski paket RapidMiner

Programski paket RapidMiner (Mierswa et al., 2006) jest programski paket otvorenog koda (engl. *open-source*) stvoren kao dio projekta koji su 2001. godine započeli Ingo Mierswa, Ralf Klinkenberg i Simon Fischer na Sveučilištu u Dortmundu. Mierswa i Klinkenberg kasnije su osnovali tvrtku Rapid-I, koja je danas glavni distributer tog paketa. Najnovija verzija RapidMinera u vrijeme pisanja ovog rada bila je $5.2.8$ te je ta verzija korištena za učenje klasifikatora.

Programski paket RapidMiner omogućuje jednostavan i učinkovit razvoj procesa koji izvršavaju najrazličitije zadaće strojnog učenja poput primjerice crpljenja informacija (engl. *information extraction*), dubinske analize podataka (engl. *data mining*) te strojne analize i obrade teksta (engl. *text mining*). U paket su uključeni brojni pri-

mjeri za učenje te je razvijen velik broj dodataka koji proširuju osnovne mogućnosti paketa. Primjerice, u osnovnoj verziji RapidMiner nudi brojne operatore koji omogućuju predobradu i vizualizaciju podataka te modeliranje i vrednovanje sustava strojnog učenja.

Najpoznatiji dodatci za RapidMiner navedeni su prilikom prvog pokretanja sustava te korisnik može odabrati koje želi instalirati i koristiti dok sustav automatski provjerava postoje li novije verzije tih dodataka od trenutno korištenih. Na primjer, u osnovnoj verziji RapidMinera ponuđena je instalacija sljedećih dodataka:

- dodatak za obradu teksta,
- dodatak koji omogućava korištenje operatora iz programskog paketa Weka,[1]
- dodatak koji omogućuje paralelno izvršavanje zadataka,
- dodatak koji pomaže pri odabiru metoda za učenje klasifikatora i slično.

Zajedno s operatorima već dostupnima u osnovnoj verziji, to RapidMiner čini vrlo moćnim programskim paketom koji uvelike olakšava razvoj sustava za strojno učenje.

Rad RapidMinera temelji se na procesima sačinjenima od stabla operatora između kojih cjevovodima putuju podaci kako bi ih svaki operator mogao obraditi. Svaki operator ima definirane ulaze i izlaze i vlastite parametre te ponekad može sadržavati i neke druge ugniježđene operatore. Ulaz u operator mogu biti lokalne datoteke, baze podataka, izvori na Internetu ili podaci spremljeni u repozitorij RapidMinera koji se definira prilikom prvog pokretanja programa. Izlaz iz zadnjeg operatora može se ili spojiti na izlaz cijelog procesa ili na ulaz nekog od odgovarajućih operatora za pohranu podataka. Ukoliko je izlaz spojen na izlaz iz procesa, nakon završetka obrade otvara se prozor u kojem je moguće pregledati rezultate dobivene izvođenjem procesa. Kako je izvorni Java-kôd RapidMinera javno dostupan, ukoliko se za tim ukaže potreba vrlo je lako napraviti vlastite operatore koji pružaju dodatne funkcionalnosti programskom paketu.

Uz rad u grafičkom sučelju, RapidMiner je moguće uključiti kao zasebnu biblioteku u vlastitu Java aplikaciju.

## 5.2. Korišteni klasifikatori

Klasifikatori korišteni u ovom radu su:

- Naivan Bayesov klasifikator

[1] Weka – programski paket za strojno učenje, razvijen na University of Waikato, Novi Zeland.

- SVM
- k-NN
- AdaBoost
- DecisionStump
- Ripple Down Rule

### Naivan Bayesov klasifikator

Naivan Bayesov klasifikator je klasifikator koji je temeljen na Bayesovom teoremu o uvjetnoj vjerojatnosti. Naziv "naivan" ima zbog vrlo strogih pretpostavki oko međusobne zavisnosti značajki, odnosno pretpostavlja da su sve značajke međusobno uvjetno nezavisne za zadanu klasu. Navedeni klasifikator u RapidMiner okruženju nema parametara koji se mogu optimirati.

### SVM

SVM ili Support Vector Machine je klasifikator koji su opisali Cortes i Vapnik (1995), a temeljen je na ideji maksimizacije udaljenosti između separacijske hiperravnine i primjera za klasifikaciju (engl. *max-margin classification*). Konačan cilj jest iz skupa primjera za učenje odabrati potporne vektore koji će definirati hiperravninu na način da je njezina udaljenost od najbližih pozitivnih i negativnih primjera maksimalna (čime se poboljšava generalizacija). U slučaju skupa podataka za koji ne postoji savršena separacija u danom prostoru, SVM podržava korištenje meke (engl. *soft*) margine gdje se pogrešna klasifikacija primjera kažnjava u skladu s parametrom C. Jezgrene (engl. *kernel*) funkcije koriste se iz razloga što većina primjera nije linearno odvojiva,[2] pa se nastoji povećati dimenzionalnost problema kako bi se u višoj dimenziji uspjela konstruirati hiperravnina koja uspješno odvaja pozitivne primjere od negativnih. Te funkcije su najčešće linearne, polinomijalne, radijalne i sigmoidalne. U ovom radu koristi se linearna funkcija, budući da je u radovima poput (Cohen et al., 2004; Carvalho i Cohen, 2006) pokazala najbolje performanse kod klasifikacije teksta. U RapidMiner okruženju korištena je LibSVM-implementacija koju su razvili Chang i Lin (2001). Optimiran je parametar C koji određuje složenost modela mijenjajući cijenu pogrešne klasifikacije.

[2]Linearna odvojivost – pozitivni i negativni primjeri mogu se odvojiti povlačenjem hiperravnine između njih.

### $k$-NN

Klasifikator $k$-NN (engl. *$k$-nearest neighbor*) klasificira primjer na temelju $k$ najbližih primjera u skupu za učenje. Dakle, klasa primjera određena je većinskim glasovanjem između $k$ već prethodno klasificiranih primjera koji su najbliži primjeru koji promatramo na osnovu neke od mjera kojima izražavamo sličnost između dva primjera. U ovom radu korištena je kosinusna sličnost (engl. *cosine similarity*), s obzirom da se ista koristi u brojnim pretraživačima kako bi se usporedila sličnost dokumenta s traženim izrazom. Parametar koji je moguće optimirati u RapidMineru odnosi se na broj susjeda koji će se promatrati. Veći broj susjeda neutralizira šum u skupu podataka, no ujedno uzrokuje i veću pristranost modela.

### AdaBoost

AdaBoost je meta-algoritam učenja opisan u (Freund i Schapire, 1997). Karakteristika tog algoritma je da u određenom broju iteracija poziva jednostavnije algoritme za učenje i na temelju njihovih glasova određuje težinu koju će pridijeliti svakom primjeru za učenje u skupu podataka. Primjerice, svaki pogrešno klasificirani primjer će dobiti veću težinu, kako bi se povećala vjerojatnost da u idućem koraku bude ispravno klasificiran. U sklopu ovog parametra u RapidMineru je moguće optimirati parametar koji određuje broj iteracija. Kao slabiji klasifikator korišten je *Decision Stump*, opisan u idućem odjeljku.

### Decision Stump

*Decision Stump* je vrlo jednostavan klasifikator koji se sastoji od stabla s jednom jedinom razinom ispod korijena. Odluka se donosi isključivo na temelju jedne značajke. S obzirom da se radi o vrlo jednostavnom klasifikatoru, RapidMiner ne pruža nikakve mogućnosti optimiranja parametara za njega.

### Ripple-down pravilo

Ripple-down pravilo (engl. *Ripple-down rule*) (RDR) je klasifikator koji se temelji na stablu odluke (engl. *decision tree*) i koji koristi binarne značajke. Klasifikator prvo definira osnovno pravilo, odnosno pretpostavi da su svi primjeri pozitivni ili negativni, pa onda na temelju riječi koje se pojavljuju stvara iznimke za osnovno pravilo. Primjerice, za govorni čin *pozdraviti* bi početno pravilo moglo biti da su svi primjeri negativni, a ukoliko u razmatranom primjeru postoji riječ *pozdrav*, taj se primjer sma-

tra iznimkom i klasificira kao pozitivan. Parametri koje je moguće optimirati u sklopu RapidMiner-implementacije su minimalna težina pojavnica i način na koji se odabire osnovno pravilo.

### 5.2.1. Odabir značajki

Odabir značajki koje će biti korištene za učenje klasifikatora temeljio se na radu koji su objavili Cohen et al. (2004), no za razliku od postupka u navedenom radu, zbog nedostatka resursa ovdje nisu korištene značajke poput oznaka vrsta riječi ili vremenskih oznaka. Pokušano je učenje korištenjem bigrama kao značajki, no to se nije pokazalo mnogo uspješnijim od osnovne inačice pa je zbog vremenske zahtjevnosti izbačeno.

Osnovne korištene značajke su riječi u dokumentu, takozvana "vreća riječi" (engl. *bag of words*), gdje je svakoj riječi pridijeljen težinski faktor. Kako bi se pronašao način izračuna težinskog faktora koji daje najbolje rezultate, isprobani su sljedeći:

- TF (Term Frequency) – učestalost pojavljivanja pojedine riječi u dokumentu;
- TF-IDF (Term Frequency - Inverted Document Frequency) – važnost riječi određena je učestalošću riječi u dokumentu, a umanjuje se ako je riječ prisutna u više različitih dokumenata;
- binarna – oznaka nalazi li se riječ u dokumentu ili ne.

#### Korjenovanje riječi

Kako klasifikator riječi koje imaju jednak korijen ne bi smatrao različitima i time nepotrebno povećao skup značajki, primijenjen je jednostavan postupak korjenovanja pod nazivom S-1 korjenovanje (Šnajder, 2010, str. 114.). Nakon takve obrade, broj različitih riječi u skupu značajki smanjio sa $15.100$ na $11.856$ kod klasifikacije govornih činova, dok je korjenovanje skupa za klasifikaciju prema važnosti smanjilo broj značajki sa $8.092$ na $5.703$.

#### Filtriranje stop-riječi

U svrhu dodatnog smanjivanja skupa značajki uklanjanjem riječi koje nemaju visoku informativnu vrijednost (stop–riječi), proveden je postupak filtriranja tih riječi pomoću liste koja se sastoji od 2024 hrvatskih stop-riječi. Vrednovanje klasifikatora izvršeno je za slučaj kad nema filtriranja i za slučaj kad je filtriranje provedeno kako bi se dobio uvid u utjecaj filtriranja na performanse klasifikatora.

## 5.3. Učenje klasifikatora govornih činova

Nakon što je popis govornih činova skraćen na šest te su obavljene sve pripremne radnje u vezi skupa podataka, slijedi učenje klasifikatora. Cilj učenja klasifikatora govornih činova jest razviti po jedan binaran klasifikator za svaki govorni čin, koji će biti u stanju za pojedini segment teksta odlučiti nalazi li se u tom segmentu određeni govorni čin ili ne. Učenje i optimiranje parametara te vrednovanje provodi se za svaki tip klasifikatora, vrstu značajki (mjera) i razinu označavanja zasebno te se u konačnici dobivaju podaci o najprikladnijem klasifikatoru, vrsti značajki te razini označavanja za svaki govorni čin. Za svaku razinu postupak učenja je istovjetan, pa će u nastavku biti prikazan postupak učenja za samo jednu razinu, dok se u odjeljku 6.1 navode dobivene mjere preciznosti za sve razine i sve klasifikatore.

### 5.3.1. Stvaranje primjera za učenje

Primjeri za učenje stvarani su pomoću RapidMiner operatora "Process Documents from Files" koji kao parametar prima listu oznaka (u ovom slučaju samo dvije jer se radi o binarnoj klasifikaciji) i direktorija u kojima se nalaze datoteke koje pripadaju toj oznaci. U ovom radu su za svaki govorni čin na svakoj razini postojala dva direktorija, jedan s pozitivnim primjerima i jedan s negativnim. Također, drugi parametar operatora pružao je mogućnost odabira koja se metoda izračuna težinskih faktora koristi (binarna, TF ili TF-IDF).

Operator "Process Documents from Files" je operator koji u sebi sadrži ugniježđene operatore koji služe za obradu svakog od dokumenata koji prolazi kroz operator. U ovom slučaju je željeno ponašanje bilo da se sva velika slova u dokumentu pretvore u mala (kako klasifikator ne bi riječi na početku rečenice gledao drugačije), što se postiglo operatorom "Transform Case" s parametrom koji označava mala slova. Nakon toga se svaki dokument razlaže na pojavnice (engl. *token*) koji označavaju riječi (što je postignuto operatorom "Tokenize") i u konačnici se izbacuju dijelovi čija je duljina kraća od dva znaka ili dulja od 35 znakova, za što je korišten operator "Filter Tokens (by length)". Ulančavanjem navedenih operatora unutar operatora "Process Documents from Files" dobiveni su željeni primjeri za učenje u obliku RapidMiner "Example Seta".

Još jedna karakteristika operatora "Process Documents from Files" jest da kao ulaz može primiti listu riječi koje da promatra prilikom računanja težinskih faktora, odnosno da sve ostale ignorira. Također, jedan od izlaza iz operatora jest upravo lista riječi koje je operator koristio prilikom računanja težinksih faktora. Ovo je vrlo važno

kako bi se pravilno računali faktori kod vrednovanja, o čemu će više riječi biti u odjeljku 6.1.

### 5.3.2. Optimizacija hiperparametara

Optimizacija hiperparametara je korak u kojem se za svaki klasifikator pokušavaju pronaći najbolji parametri kako bi mu se maksimizirale performanse. U ovom radu, optimizacija hiperparametara provodi se tako da se iterira kroz moguće vrijednosti parametara za svaki klasifikator i unakrsnom provjerom (engl. *cross-validation*) vrednuje prikladnost skupa hiperparametara za klasifikaciju.

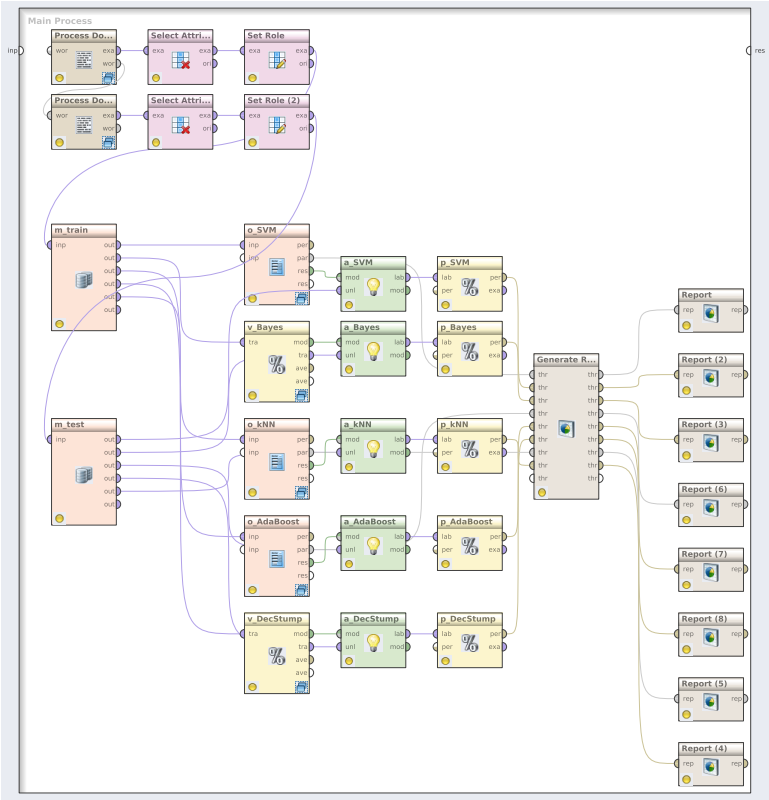
Kako bi se optimirali parametri, za svaki je govorni čin stvoren skup pozitivnih i negativnih primjera operatorom "Process Documents from Files". Ti podaci spojeni su na ulaz operatora koji iterira po parametrima klasifikatora i to za svaki od klasifikatora kojima možemo optimirati parametre (AdaBoost, SVM, k–NN i RDR). Unutar tog operatora nalazi se operator koji uči klasifikator i provodi deseterostruku unakrsnu provjeru. Time se na izlazu operatora koji optimira parametre dobiva naučeni model klasifikatora s optimalnom vrijednosti parametra.

Optimirani su sljedeći parametri:

- SVM: parametar *C* (cijena pogrešne klasifikacije) u logaritamskom rasponu od $0{,}2$ do $1024$, $50$ vrijednosti;
- k–NN: parametar *k* (broj promatranih susjeda) u linearnom rasponu od $1$ do $15$, $15$ vrijednosti;
- AdaBoost: parametar *I* (broj iteracija) u linearnom rasponu od $1$ do $10$, $10$ vrijednosti;
- RDR: parametar *N* (minimalna težina primjera u podjeli) u linearnom rasponu od $1$ do $3$, $3$ vrijednosti te parametri *M* i *A* koji imaju binarne vrijednosti i određuju način na koji se izabire pretpostavljena klasa.

Nakon što je za svaki klasifikator dobiven optimalan model, potrebno je vrednovati taj model na skupu podataka koje klasifikator dosad nije vidio. Zbog toga je početni skup podataka podijeljen u omjeru 70%-30%, od kojih 70% služi za učenje klasifikatora i optimiranje parametara, dok 30% služi za vrednovanje konačnih modela. Time se na izlazu dobiva procjena performansi klasifikatora nad skupom podataka koji dosad nije viđen.

S obzirom na to da metode TF i TF-IDF uzimaju u obzir čitav skup podataka kod računanja, važno je prije izračuna odijeliti skup podataka za učenje i za ispitivanje i




**Slika 5.1:** RapidMiner-proces koji služi za jednu iteraciju učenja klasifikatora

onda na svakom zasebno računati težinske faktore. No, također, potrebno je i uskladiti koje će riječi biti promatrane u oba skupa. To je moguće učiniti tako da se lista riječi koja se dobije kao drugi izlaz iz operatora "Process Documents from Files" (koji računa težinske faktore metodama TF i TF-IDF za prijere za učenje) proslijedi na ulaz operatora koji istu funkciju obavlja nad ispitnim skupom. Tu je listu riječi potrebno pohraniti kako bi mogla biti korištena za kasniju klasifikaciju dodatnih poruka.

## 5.4. Primjena klasifikatora govornih činova

Nakon što je za svaki govorni čin dobiven klasifikator koji ga najuspješnije otkriva i označuje u segmentu teksta, potrebno je pronaći način da se ta informacija iskoristi za klasifikaciju poruka prema važnosti. Kao prvi korak, izlaz klasifikatora govornih činova promijenjen je tako da umjesto binarne oznake koja indicira sadrži li segment teksta govorni čin klasifikator vrati vjerojatnost da je govorni čin sadržan u tom segmentu teksta. Ideja je da za svaki segment teksta dobijemo po šest vrijednosti koje označavaju vjerojatnosti da taj segment sadrži svaki od šest analiziranih govornih činova. Za neke klasifikatore (RDR i naivan Bayes) to nije moguće, budući da je njihov

izlaz u ovom slučaju bio isključivo binaran te je kao takav uzet u razmatranje. Tih šest vrijednosti koristi se kako bi se proširio skup značajki klasifikatora na temelju važnosti, koji je dotad sadržavao samo značajke temeljene na "vreći riječi". Usporedbom rezultata dobivenih uz korištenje originalnog skupa značajki i skupa proširenog oznakama govornih činova moguće je odrediti doprinos govornih činova klasifikaciji na temelju važnosti.

S obzirom na relativno visoku stabilnost rezultata klasifikacije govornih činova na razini poruke, odnosno relativno dobre performanse, odlučeno je da će se za proširivanje skupa značajki koristiti oznake klasifikatora govornih činova samo na toj razini. Time se ujedno olakšava rukovanje podacima, jer je za svaku poruku dovoljno pohraniti vjerojatnost da se u njoj pojavljuje pojedini govorni čin, dok bi kod klasifikacije na razini odlomka ili rečenice imali više oznaka za jedan čin po poruci te bi bilo potrebno razraditi način na koji bi koristili taj skup oznaka.

Tehnička izvedba u programskom paketu RapidMiner sastoji se od potprocesa koji čita podatke s ulaza koji sadrže čitav tekst poruke, nad njima provodi obradu koristeći operator "Process Documents from Files" namješten tako da izbor mjere učestalosti sastavnica najbolje odgovara korištenom klasifikatoru uzimajući u obzir skup značajki nad kojima je učen klasifikator za pojedini govorni čin. Riječi odnosno značajke koje se ne pojavljuju prilikom učenja klasifikatora za govorni čin, a koje postoje u skupu za klasifikaciju prema važnosti, zanemaruju se. Popis riječi i model za klasifikaciju svakog govornog čina učitavaju se iz unaprijed pripremljenih datoteka. Dobiveni se rezultati spajaju u jedan skup primjera te pridružuju skupu značajki vreća riječi i u takvom obliku prosljeđuju na izlaz potprocesa.

Idući korak bio je odabir klasifikatora koji su se za pojedine govorne činove pokazali najboljima na razini dokumenta, uz pripadajuće parametre i vrstu značajki. Korišteni klasifikatori prikazani su u Tablici 5.1, uz postignute performanse nad skupom za ispitivanje pojedinog govornog čina. Nakon odabira primjera, svi su klasifikatori učeni nad cijelim označenim skupom za pojedini govorni čin te spremljeni za kasniju upotrebu kod označavanja poruka.

### Ujednačavanje skupa značajki

Prilikom upotrebe klasifikatora na temelju govornih činova nad skupom poruka koje se klasificiraju prema važnosti potrebno je uzeti u obzir i mogućnost da klasifikator za govorne činove bude previše prilagođen skupu podataka nad kojim je učen i nad kojim su optimirani parametri. Primjerice, moguć je slučaj gdje za klasifikator postoji određeni

**Tablica 5.1:** Odabrane postavke klasifikatora govornih činova

| | Klasifikator | Parametri | Tip značajki | Filtriranje stop-riječi | % F1 |
|---|---|---|---|---|---|
| ISPORUČITI | RDR | N=3,A=0,M=0 | binarno | + | 86,59 |
| IZMIJENITI | NB | – | TF-IDF | – | 79,31 |
| OBVEZATI_SE | RDR | N=3,A=0,F=0 | binarno | – | 83,75 |
| PODSJETITI | AB | I=3 | TF | + | 94,74 |
| PREDLOŽITI | AB | I=1 | TF-IDF | – | 71,88 |
| ZAHTIJEVATI | SVM | C=1,553 | TF | + | 70,09 |

podskup riječi koje su veoma diskriminatorne i pomoću njih klasifikator uspijeva održati relativno dobre rezultate. No, lako je moguće da se niti jedna riječ iz tog podskupa ne pojavi prilikom klasifikacije prema važnosti, pa tada klasifikator gubi na točnosti. Kako bi se ispitala ta mogućnost i pokušalo ponuditi rješenje problema, pristupilo se ujednačavanju skupa značajki između klasifikatora govornih činova i klasifikatora na temelju važnosti.

Ujednačavanje je provedeno tako da je načinjen popis riječi koji sadrži sve riječi koje se pojavljuju u skupu za učenje klasifikatora na temelju važnosti. Nakon toga su pomoću skripte napisane u programskom jeziku Python uklonjene sve riječi koje se ne nalaze na navedenom popisu, a nalazile su se u skupu za učenje nekog od klasifikatora govornih činova. Time se postiglo da su svi klasifikatori govornih činova prisiljeni učiti samo nad skupom riječi koje se mogu kasnije pojaviti prilikom klasifikacije prema važnosti. Pritom i dalje svaki od klasifikatora zadržava vlastitu listu korištenih riječi, no sad te liste više ne mogu sadržavati niti jednu riječ koja nije prisutna u skupu za učenje klasifikatora prema važnosti.

To je dovelo do potrebe za ponovnim optimiranjem parametara svih klasifikatora te odabirom najpogodnijeg klasifikatora i učenjem tog klasifikatora nad cijelim skupom za govorni čin. Odabrani klasifikatori i njihove postavke zajedno s F1-mjerom prikazani su u Tablici 5.2. Usporedbom rezultata u Tablici 5.1 i Tablici 5.2 vidljivo je da su dobiveni rezultati s obzirom na F1-mjeru u pravilu niži nego oni dobiveni bez ujednačavanja, što govori u prilog pretpostavci da su doista postojale značajke koje su bile diskriminativne, a nisu se uopće pojavljivale u skupu za učenje klasifikatora na temelju važnosti.

Za svaki od govornih činova odabran je najbolji klasifikator za slučaj bez ujednačavanja i s ujednačavanjem, tako da je moguće da se tipovi klasifikatora razlikuju,

**Tablica 5.2:** Odabrane postavke klasifikatora govornih činova uz ujednačavanje skupa značajki

| | Klasifikator | Parametri | Tip značajki | Filtriranje stop-riječi | % F1 |
|---|---|---|---|---|---|
| ISPORUČITI | SVM | C=1,3094 | TF | – | 84,47 |
| IZMIJENITI | SVM | C=78,98 | TF-IDF | – | 77,97 |
| OBVEZATI_SE | AB | I=8 | TF-IDF | – | 79,10 |
| PODSJETITI | AB | I=7 | TF | + | 90,00 |
| PREDLOŽITI | SVM | C=3,0761 | TF | – | 75,20 |
| ZAHTIJEVATI | SVM | C=2,1859 | TF | + | 70,94 |

što može objasniti razlike u distribucijama pouzdanosti pripadnosti pozitivnoj klasi za određene govorne činove, što je prvenstveno uočeno kod govornih činova ISPORUČITI, IZMIJENITI te u manjoj mjeri za OBVEZATI_SE. Za navedene govorne činove bez ujednačavanja skupova značajki korišteni su naivan Bayesov klasifikator i RDR te su modeli na izlazu dali binarnu oznaku sadrži li segment teksta govorni čin ili ne. U slučaju kad su skupovi značajki ujednačeni, kao bolji klasifikatori pokazali su se SVM i AdaBoost koji kao izlaz daju pouzdanost pripadnosti pojedinom razredu, što uzrokuje razlike u dobivenim distribucijama.

No, neovisno o tome, primjećeno je kako klasifikatori nakon ujednačavanja skupova značajki imaju mnogo bolju distribuciju, odnosno veći broj vrijednosti koje se koriste kao mjera pripadnosti razredu. To je posebice uočljivo u slučaju govornog čina PREDLOŽITI. I u tom slučaju korištena su dva različita klasifikatora, AdaBoost prije i SVM nakon ujednačavanja, no budući da oba klasifikatora u pravilu vraćaju pouzdanost pripadnosti klasi, rezultati dobiveni prije i nakon ujednačavanja skupa značajki su usporedivi.

Iz navedenih rezultata moguće je pretpostaviti kako bi ujednačavanjem značajki postigli veći raspon vrijednosti za oznake govornih činova koje bi kasnije mogao koristiti klasifikator prema važnosti te bi time klasifikacija trebala biti bolja. No, također, potrebno je imati na umu i činjenicu da ujednačavanjem skupova smanjujemo dimenzionalnost skupa značajki za svaki pojedini klasifikator, što u konačnici može uzrokovati lošije performanse od očekivanih.

Također, učinjena je i usporedba distribucija pouzdanosti za klasifikatore koji su se pokazali najboljima u slučaju bez ujednačavanja. Za te je klasifikatore provedeno optimiranje parametara uz ujednačavanje značajki kako bi se usporedile dobivene distribucije pouzdanosti u slučaju kad se klasifikatori ne mijenjaju. Primjećeno je da nije

bilo većih promjena u izgledu distribucija, što bi moglo značiti kako uklonjene značajke nisu bile previše informativne za te tipove klasifikatora.

## 5.5. Učenje klasifikatora važnosti poruka

Cilj učenja klasifikatora važnosti poruka jest naučiti binaran klasifikator koji uspješno odvaja važne poruke od nevažnih, koristeći tekst poruke i dodatne podatke dobivene označavanjem govornih činova. Klasifikator radi na razini cijelih poruka i ne uzima zasebno u obzir manje dijelove poput odlomaka ili rečenica. Učenje se provodi koristeći pet klasifikatora od kojih se svaki uči za dvije vrste značajki (TF i TF-IDF) te uz ujednačavanje skupa značajki i bez njega. Kako bi bilo moguće ocijeniti utjecaj govornih činova na klasifikaciju prema važnosti, prethodno opisani postupci učenja ponovljeni su i za slučaj kad se koriste samo značajke vreće riječi, bez oznaka govornih činova.

### 5.5.1. Stvaranje primjera za učenje

Postupak stvaranja primjera za učenje s implementacijske je strane gotovo istovjetan procesu opisanom o odjeljku 5.3.1. I u ovom slučaju se pojavnice izdvajaju iz teksta u obliku riječi obrađenih tako da se velika slova pretvore u mala i uklone pojavnice koje su kraće od dva znaka i dulje od 35 znakova. No, za razliku od prethodno opisanog procesa, u ovom slučaju je potrebno tekst poruke predati potprocesu koji će izračunati vjerojatnosti da se pojedini govorni činovi nalaze u tom tekstu te vratiti rezultat u obliku značajki. Te dobivene značajke pridružuju se značajkama vreće riječi u slučaju kad želimo klasificirati uz pomoć oznaka govornih činova.

Prilikom stvaranja primjera za učenje također je važno održavati listu riječi koje se pojavljuju u primjerima za učenje kako bi se samo te riječi kasnije koristile kod ispitivanje i kako bi težinski faktori, posebice u slučaju metode TF-IDF, mogli biti uspješno izračunati.

Budući da je konačan cilj ovog rada usporedba performansi klasifikatora važnosti poruke uz korištenje oznaka govornih činova i bez njih, prilikom izrade primjera za učenje osigurano je da skup za učenje i ispitivanje budu istovjetni za ova slučaja korištenja klasifikatora. To je postignuto tako da su primjeri za učenje i ispitivanje klasifikacije prema važnosti temeljeni na značajkama vreće riječi proslijeđeni na izlaz prije obrade potprocesom koji služi za označavanje govornih činova. Isti skupovi primjera su kasnije obrađeni tako da su im pridodane oznake govornih činova te se time postigla usporedivost dobivenih rezultata za oba slučaja klasifikacije.

### 5.5.2. Optimizacija hiperparametara

Proces optimiranja parametara klasifikatora u suštini je vrlo sličan istovjetnom procesu kod klasifikacije govornih činova. Za svaki se klasifikator pretražuje skup mogućih vrijednosti parametara te se za svaku vrijednost deseterostrukom unakrsnom validacijom procjenjuju performanse klasifikatora. Optimiranje se provodi nad skupom za učenje, koji čini $70\%$ ukupne količine podataka. U ovom slučaju postoje tri klasifikatora čije parametre je moguće optimirati, a to su SVM, k–NN te klasifikator AdaBoost.

Za njih su optimirani sljedeći parametri:

- SVM: parametar *C* (cijena pogrešne klasifikacije) u logaritamskom rasponu od $0{,}0125$ do $1024$, $25$ vrijednosti;
- k–NN: parametar *k* (broj promatranih susjeda) u linearnom rasponu od $1$ do $10$, $10$ vrijednosti;
- AdaBoost: parametar *I* (broj iteracija) u linearnom rasponu od $1$ do $10$, $10$ vrijednosti.

# 6. Eksperimentalno vrednovanje klasifikatora

Prilikom eksperimentalnog vrednovanja klasifikatora u obzir su uzete četiri mjere kojima se iskazuje rezultat klasifikatora nad ispitnim skupom. To su: točnost (engl. *accuracy*), preciznost (engl. *precision*), odziv (engl. *recall*) i F1 mjera (engl. *F1-score*). Od navedenih, kao najvažnija uzeta je F1-mjera prilikom optimiranja parametara i konačnog vrednovanja klasifikatora, no i ostale mjere uzete su u obzir prilikom analize rezultata.

**Točnost** je mjera koja prikazuje odstupanje izmjerene količine od stvarne vrijednosti, odnosno koliko je dobivena klasifikacija udaljena od stvarne raspodjele po klasama. Točnost od $100\%$ predstavlja dobivenu raspodjelu po klasama koja je istovjetna stvarnim klasama primjera koje želimo klasificirati. Definirana je kao omjer zbroja stvarnih pozitivnih (engl. *true positive, TP*) i stvarnih negativnih (engl. *true negative, TN*) rezultata te zbroja stvarnih pozitivnih i negativnih rezultata i lažnih pozitivnih (engl. *false positive, FP*) i lažnih negativnih (engl. *false negative, FN*) rezultata, odnosno $A = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$.

**Preciznost** za razliku od točnosti, nije moguće definirati za jedno mjerenje jer označava mjeru pouzdanosti mjernog uređaja, odnosno ponovljivost dobivenih rezultata. Preciznost i točnost su pojmovi koji se često koriste kao sinonimi, no u kontekstu analize rezultata klasifikacije, njihova značenja su vrlo različita. Preciznost se definira kao omjer stvarnih pozitivnih rezultata i svih pozitivnih rezultata klasifikacije, odnosno kao $P = \frac{\text{TP}}{\text{TP+FP}}$.

**Odziv** ili osjetljivost klasifikatora je mjera koja iskazuje koliko dobro klasifikator izdvaja pozitivne primjere. Definirana je kao udio stvarnih pozitivnih primjera koje je klasifikator označio u cijelom skupu pozitivnih primjera, odnosno kao omjer stvarnih

pozitivnih primjera i stvarnih pozitivnih i lažnih negativnih rezultata, $R = \frac{\text{TP}}{\text{TP+FN}}$.

**F1-mjera** je također mjera točnosti klasifikacije. Ona uzima u obzir i preciznost i odziv klasifikatora te predstavlja njihovu harmonijsku sredinu. Kao i u slučaju računanja točnosti, vrijednost F1-mjere kreće se u rasponu od $0\%$ do $100\%$. Matematički, F1-mjera klasifikatora definirana je na sljedeći način: $F1 = 2 \cdot \frac{P \cdot R}{P+R}$. Prema potrebi, F1-mjera može uključivati i parametar $\beta$ kojim se određuje hoće li se više vrednovati visoka preciznost ili odziv.

## 6.1. Vrednovanje klasifikatora govornih činova

Nakon što se provede optimizacija parametara opisana u odjeljku 5.3.2, u svrhu vrednovanja klasifikatora potrebno je izmjeriti performanse svakog klasifikatora govornih činova nad dotad neviđenim primjerima za ispitivanje izražene u obliku F1-mjere.

Ispitivanje klasifikatora sastojalo se od učenja klasifikatora uz optimalnu vrijednost parametara na cijelom skupu podataka za učenje (70% cijelog označenog skupa za pojedini govorni čin) te primjene na klasifikaciju dotad neviđenog skupa podataka (preostalih 30% označenog skupa).

Prilikom ispitivanja bilo je vrlo bitno uskladiti skupove značajki nad kojima se provodi klasifikacija, i to pogotovo kod metode TF-IDF budući da ona uzima u obzir učestalosti pojedinih značajki u različitim klasama. U tu svrhu korištena je lista riječi dobivena prilikom učenja klasifikatora nad cijelim skupom podataka za učenje, budući da je time skup značajki koje će se koristiti kod testiranja ograničen na značajke koje su već viđene prilikom učenja, dok se ostale značajke ignoriraju. Također, na taj način se prilikom izračunavanja mjere TF-IDF može uzeti u obzir raspodjela značajki po klasama.

### Analiza rezultata

Tablica 6.1 prikazuje performanse šest naučenih klasifikatora za šest govornih činova uz korištenje F1-mjere. Prikazani su samo modeli koji su postigli najbolje rezultate, neovisno o razini označavanja ili korištenim značajkama. Vidljivo je kako klasifikatori SVM i RDR u pravilu postižu bolje rezultate od ostalih klasifikatora, uz F1 mjeru koja prelazi 88%. Klasifikator SVM ne samo da je pokazao najbolje performanse, već je imao i najmanju razliku između najboljeg i najlošijeg rezultata i to u rasponu od 75% (za govorni čin PODSJETITI) do 88.16% (za govorni čin ISPORUČITI). Klasifi-

**Tablica 6.1:** Performanse klasifikatora prema govornim činovima (% F1)

| | NB | k-NN | SVM | DS | AB | RDR |
|---|---|---|---|---|---|---|
| ISPORUČITI | 69,70 | 83,72 | 88,16 | 85,71 | 87,50 | **88,51** |
| IZMIJENITI | **79,31** | 71,43 | 77,97 | 72,29 | 74,63 | 77,27 |
| OBVEZATI_SE | 62,45 | 67,44 | 78,61 | 79,37 | 81,97 | **83,75** |
| PODSJETITI | 60,87 | 63,64 | 75,00 | 76,92 | **94,74** | 76,92 |
| PREDLOŽITI | 67,06 | 70,27 | **76,84** | 76,27 | 75,12 | 71,50 |
| ZAHTIJEVATI | 69,69 | 75,44 | **78,76** | 70,57 | 75,23 | 74,46 |

**Tablica 6.2:** Performanse klasifikatora prema razini označavanja (% F1)

| | Poruka | Odlomak | Rečenica |
|---|---|---|---|
| ISPORUČITI | 86,59 | 83,64 | **88,51** |
| IZMIJENITI | **79,31** | 77,27 | 72,38 |
| OBVEZATI_SE | **83,75** | 81,97 | 78,93 |
| PODSJETITI | **94,74** | 76,92 | 69,57 |
| PREDLOŽITI | 71,88 | **76,84** | 69,74 |
| ZAHTIJEVATI | 70,09 | **78,76** | 72,19 |
| *Ukupno* | 94,74 | 83,64 | 78,93 |

kator AdaBoost također je pokazao prilično dobre rezultate te je najuspješnije klasificirao govorni čin PODSJETITI. Klasifikator *Decision Stump* ostvario je začuđujuće kvalitetne rezultate ako uzmemo u obzir jednostavnost modela. Iz navedenih rezultata vidljivo je i da većina klasifikatora najbolje rezultate ostvaruje na govornom činu ISPORUČITI. S druge strane, govorni čin PODSJETITI pokazao se kao najteži za klasifikaciju, što bi se moglo pripisati i uvjerljivo najnižem broju primjera za učenje od svih govornih činova.

U Tablici 6.2 moguće je vidjeti rezultate koje su ostvarili najbolji klasifikatori (bez obzira na tip i vrstu značajki) za svaki govorni čin prema razini označavanja. U rezultatima nije primjetan nikakav globalni trend koji bi upućivao na veću prikladnost pojedine razine označavanja naspram ostalih razina. Iako bi možda bilo očekivano da rezultati klasifikacije budu bolji na razini rečenice, budući da primjeri sadrže manje riječi koje nemaju veze s navedenim govornim činom nego što je to primjerice slučaj

**Tablica 6.3:** Performanse klasifikatora na prema vrsti značajki (% F1)

| | Uz stop-riječi | | | Bez stop-riječi | | |
|---|---|---|---|---|---|---|
| | Binary | TF | TF-IDF | Binary | TF | TF-IDF |
| ISPORUČITI | **88,51** | 87,50 | 88,00 | **88,51** | 88,16 | 87.96 |
| IZMIJENITI | 70,07 | 77,19 | **79,31** | 77,27 | 75,86 | **77,19** |
| OBVEZATI_SE | **83,75** | 79,37 | 81,63 | 78,82 | 79,76 | **81,97** |
| PODSJETITI | 76,92 | 76,92 | **77,78** | 75,00 | **94,74** | 77,78 |
| PREDLOŽITI | 71,50 | **76,84** | 76,27 | 68,40 | 73,08 | **73,68** |
| ZAHTIJEVATI | 61,90 | **78,76** | 78,10 | 74,46 | **78,08** | 77,53 |

kod klasifikacije na razini dokumenta, to se nije pokazalo kao pravilo. No, ovi rezultati nam mogu biti vrlo korisni i znakoviti s jezičnog aspekta budući da nam pomažu razumjeti na kojoj razini su pojedini govorni činovi najčešće izraženi. Na primjer, podsjetnik nekome je rijetko izražen samo jednom rečenicom te bi stoga bilo očekivano da klasifikatora na razini odlomka pokazuje bolje rezultate nego klasifikator na razini rečenice. Nadalje, govorni čin ISPORUČITI najčešće se izražava malim brojem riječi, pa bi bilo očekivano da klasifikacija na razini rečenice pokaže najbolje rezultate za taj govorni čin. Ukupno gledajući, klasifikacija na razini dokumenta pokazala se najboljom za većinu govornih činova, boljom od klasifikacije na razini odlomka. To je moguće objasniti tim što većina klasifikatora ima visok odziv (engl. *recall*) te je potrebno više okolnog teksta kako bi se uklonili lažni pozitivni rezultati.

Kako bi se procijenio utjecaj različitih tipova značajki na rezultate klasifikacije, u Tablici 6.3 prikazane su performanse najboljih klasifikatora za svaki par govornog čina i tipa značajki. Filtriranje stop-riječi, opisano u odjeljku 5.2.1, u pravilu ne pokazuje veći utjecaj na rezultate. U slučaju kad stop-riječi nisu filtrirane, performanse klasifikatora koji koriste sve tri vrste značajki su usporedive, osim u slučaju govornog čina ZAHTIJEVATI, gdje binarne značajke daju nešto lošiji rezultat. Apsolutna razlika između različitih vrsta značajki uglavnom je zadržana unutar 3% F1-mjere, što pokazuje kako je problem klasifikacije na temelju govornih činova uglavnom otporan na promjene vrste značajki.

U svrhu vrednovanja rezultata pojedinih tipova klasifikatora, u Tablici 6.4 prikazane su performanse klasifikatora na sve tri razine označavanja. Prikazani rezultati odnose se na najbolji rezultat koji je taj klasifikator postigao, neovisno o govornom činu

**Tablica 6.4:** Ukupne performanse klasifikatora (% F1)

| | Poruka | Odlomak | Rečenica |
|---|---|---|---|
| NB | **79,31** | 69,70 | 72,38 |
| k-NN | 72,73 | 75,44 | **83,72** |
| SVM | 83,87 | 81,55 | **88,16** |
| DS | 78,65 | 79,37 | **85,71** |
| AB | **94,74** | 83,54 | 87,50 |
| RDR | 86,59 | 83,64 | **88,51** |

ili vrsti značajki. Većina klasifikatora najbolje rezultate postiže za razinu rečenice, što je iznenađujuće s obzirom na prethodne rezultate gdje se razina rečenice pokazala najtežom za klasifikaciju. No, to je moguće objasniti ako uzmemo u obzir da su prikazani rezultati pod velikim utjecajem visokih performansi klasifikatora nad govornim činom ISPORUČITI. Ukupne performanse klasifikatora za označavanje govornih činova su relativno dobre, pogotovo u usporedbi s rezultatima koje su za engleski jezik ostvarili Cohen et al. (2004): u radu F1-mjere variraju od $79{,}31\%$ do $94{,}74\%$, dok Cohen et al. (2004) prikazuju rezultate u rasponu od $44\%$ do $85\%$.

## 6.2. Vrednovanje klasifikatora važnosti

Nakon što je završeno optimiranje parametara klasifikatora iz odjeljka 5.5.2, učenje se provodi nad cijelim skupom podataka za učenje ($70\%$ ukupnog broja primjera) te se vrednovanje radi tako što se naučeni model primijeni na skupu za ispitivanje koji nije viđen prilikom učenja. Za svaki klasifikator prilikom vrednovanja bilježe se podaci o točnosti, preciznosti, odzivu te F1–mjera. Svi klasifikatori imaju jednak skup za učenje i ispitivanje te se stoga dobiveni rezultati mogu izravno koristiti za usporedbu klasifikatora i njihove prikladnosti za rješavanje ove vrste problema.

Uz usporedbu rezultata klasifikacije, učinjeno je i vrednovanje važnosti pojedinih značajki za uspješnost klasifikacije. Time je moguće rangirati značajke prema važnosti i tako utvrditi jesu li značajke koje se odnose na oznake govornih činova doista važne ili nisu.

**Tablica 6.5:** Ukupne performanse klasifikatora za klasifikaciju važnosti poruka

| Klasifikator | Točnost | Preciznost | Odziv | F1-mjera |
|---|---|---|---|---|
| NB | 65,35 | 61,70 | **80,93** | 70,02 |
| k-NN | 73,72 | 76,56 | 68,37 | 72,24 |
| SVM | **78,14** | 77,88 | 78,60 | **78,24** |
| DS | 61,16 | **96,15** | 23,26 | 37,45 |
| AB | 70,93 | 78,48 | 57,67 | 66,49 |

### Performanse klasifikatora

U Tablici 6.5 navedene su ukupne performanse za svaki klasifikator. Za svaki klasifikator navedena je najviša dobivena F1-mjera neovisno o vrsti značajki te pripadajuće mjere točnosti, preciznosti i odziva. Vidljivo je da su dobiveni relativno visoki rezultati, odnosno za većinu klasifikatora je F1-mjera iznad $70\%$. Najboljim se pokazao klasifikator SVM uz F1-mjeru od $78{,}24\%$. Isti klasifikator bio je među najboljima i prilikom klasifikacije na temelju govornih činova. Za isti je klasifikator primjetna i najmanja razlika između preciznosti i odziva te su mu sve mjere ujednačene. Ti rezultati ističu SVM kao vjerojatno najpogodniji izbor za rješavanje problema klasifikacije poruka prema važnosti.

S druge strane, klasifikator *Decision Stump* pokazao se kao daleko najlošiji gledajući F1-mjeru, uz rezultat od samo $37{,}45\%$, što ne iznenađuje ako se uzme u obzir jednostavnost modela. No, gledajući samo točnost klasifikacije, razlika i nije toliko drastična te se *Decision Stump* približava rezultatu dobivenom od naivnog Bayesovog klasifikatora. Važno je napomenuti da *Decision Stump* pruža daleko najveću preciznost klasifikacije od svih dobivenih rezultata, postižući gotovo $20\%$ veću preciznost od najboljeg idućeg klasifikatora. Razlog loših rezultata u pogledu točnosti leži u slabom odzivu, pa ako se preciznost klasifikacije uzme kao važniji parametar od odziva, *Decision Stump* bi se mogao pokazati kao vrlo dobar izbor. Također, važno je istaknuti da je unatoč relativno lošem pojedinačnom rezultatu klasifikatora *Decision Stump*, klasifikacija korištenjem metode AdaBoost koja u svom radu koristi klasifikator *Decision Stump* pokazuje mnogo bolje rezultate, što ukazuje na mogućnost kombiniranja navedenog klasifikatora s istovjetnim ili nekim drugim klasifikatorima s ciljem poboljšanje rezultata.

U pogledu odziva, najboljim se pokazao naivan Bayesov klasifikator koji pruža

**Tablica 6.6:** Utjecaj govornih činova na performanse klasifikatora (%F1)

| Klasifikator | BoW[1] | GČ[2] | GČ+BoW |
|---|---|---|---|
| NB | 70,02 | 70,02 | 70,02 |
| k-NN | 70,07 | 71,83 | 72,24 |
| SVM | 66,48 | 72,56 | 78,24 |
| DS | 37,45 | 37,45 | 37,45 |
| AB | 60,31 | 62,21 | 66,49 |
| *Ukupno* | 70,07 | 72,56 | 78,24 |

daleko najveći odziv od $80{,}93\%$, no uz cijenu najniže preciznosti. No, s obzirom na jednostavnost modela i nemogućnost prilagodbe parametara, rezultati dobiveni korištenjem naivnog Bayesovog klasifikatora su vrlo visoki.

### Utjecaj govornih činova na klasifikaciju

Tablica 6.6 prikazuje vjerojatno najvažniji rezultat ovoga rada, a to je način na koji oznake govornih činova koje su pridodane skupu značajki vreće riječi utječu na rezultate klasifikacije. Budući da su ispitivanje s govornim činovima i bez njih provedena u potpuno jednakim okolnostima s obzirom na skup primjera za učenje, primjerima za ispitivanje, brojem i tipom parametara koji su optimirani, pa se razlika između navedenih testova svodi isključivo na šest značajki koje označavaju prisutnost govornih činova, dobiveni rezultati izravno svjedoče o utjecaju oznaka govornih činova na uspješnost klasifikacije poruka elektroničke pošte na osnovu važnosti. Za svaki klasifikator naveden je najbolji zabilježen rezultat nad ispitnim skupom za slučaj korištenja samo značajki vreće riječi, korištenja samo govornih činova te uz dodavanje oznaka govornih činova značajkama vreće riječi, neovisno o metodi koja se koristila za izračunavanje težinskih faktora značajki vreće riječi.

Prema dobivenim rezultatima s obzirom na F1-mjeru, za većinu isprobanih klasifikatora primjetno je poboljšanje prilikom dodavanja oznaka govornih činova. Osim u slučaju naivnog Bayesovog klasifikatora i klasifikatora *Decision Stump*, koji ostvaruju jednaku F1-mjeru neovisno o tome koriste li se oznake govornih činova ili ne. To se može pripisati jednostavnosti modela i utjecaju manjeg broja značajki (primjerice po-

[1] BoW – "Bag of words" skup značajki.

[2] GČ – Oznake govornih činova.

jedinih riječi) koje su karakteristične i za određene govorne činove, pa je stoga moguće da se dobivene F1-mjere podudaraju. U ostalim slučajevima poboljšanje varira od $2\%$ za slučaj k-NN klasifikatora, do $12\%$ u slučaju SVM, što je doista primjetna razlika. Razlika između najboljeg ukupnog rezultata kroz sve klasifikatore za ta dva slučaja iznosi $8{,}17\%$.

Analizom F1-mjera ostvarenih korištenjem samo oznaka govornih činova vidljivo je kako se uz samo tih šest značajki postižu bolji rezultati nego korištenjem značajki vreće riječi. To potvrđuje pretpostavke o informativnosti navedenih značajki i njihovoj korisnosti za klasifikaciju važnosti poruka. Ipak, za većinu klasifikatora rezultati dobiveni samo uz oznake govornih činova su lošiji nego rezultati dobiveni udruživanjem značajki vreće riječi i oznake govornih činova. To svjedoči o potrebi da se i značajke vreće riječi koriste prilikom konačne klasifikacije s obzirom da je očito da obuhvaćaju informacije važne za uspješnu ocjenu važnosti.

Poboljšanje dodavanjem oznaka govornih činova bilo je očekivano sa stanovišta strojnog učenja jer se radi o proširivanju skupa značajki što najčešće pozitivno utječe na klasifikatore. No, s obzirom na to da se radi o samo šest dodatnih značajki, što je približno jedna tisućina ukupnog broja značajki, gotovo je nemoguće pripisati ovoliko značajno poboljšanje samo činjenici da je skup značajki proširen. To govori u prilog pretpostavci da govorni činovi doista imaju važnu ulogu u procjeni važnosti poruke, odnosno da su oznake govornih činova zapravo vrlo informativne značajke koje donose veliku količinu novih informacija potrebnih za uspješno klasificiranje važnih poruka. To potvrđuju i dobiveni rezultati koji ukazuju na informativnost oznaka govornih činova. Govorni činovi su, kao što je prije navedeno u ovom radu, dijelovi poruka koji mnogo govore o sadržaju i namjeni iste poruke, pa je realno očekivati da poruke koje korisnici doživljavaju kao važne sadrže neke od govornih činova. Također, u samom postupku označavanja poruka na temelju važnosti, u uputstvima sastavljenima za označivače, kao jedan od kriterija navedena je količina posla koju poruka zahtijeva od primatelja. Govorni činovi, posebice ZAHTIJEVATI i IZMIJENITI od primatelja upravo i traže da obavi određenu količinu posla te je stoga očekivano da se u velikom broju važnih poruka nalaze navedeni govorni činovi. Uz to, kriterij koji govori o gubitku informacija u slučaju brisanja poruke može se povezati s govornim činom ISPORUČITI budući da poruke koje sadrže navedeni govorni čin vrlo često sadrže i neku dodatnu informaciju koja je od važnosti za primatelja, bilo u obliku privitka, bilo u nastavku teksta poruke.

Uzevši u obzir dobivene rezultate te pretpostavke iznesene i objašnjene u ovom odjeljku, moguće je zaključiti da govorni činovi značajno pridonose klasifikaciji po-

**Tablica 6.7:** Utjecaj korištenog načina izračuna težinskih faktora na performanse klasifikatora (%F1)

| Klasifikator | TF | TF-IDF |
|---|---|---|
| NB | 70,02 | 70,02 |
| k-NN | 71,66 | 72,24 |
| SVM | 66,48 | 78,24 |
| DS | 37,45 | 37,45 |
| AB | 66,49 | 66,49 |
| *Ukupno* | 71,66 | 78,24 |

ruka na temelju važnosti te da taj malen skup oznaka uistinu opravdava dodavanje skupu značajki vreće riječi. Ta saznanja mogu biti od velikog značaja za budućnost područja klasifikacije poruka prema važnosti, s obzirom da svjedoče koliko malen i informativan skup značajki od sadržaju poruke može pozitivno utjecati na performanse klasifikacije, čak i ako rezultati klasifikacije govornih činova nisu apsolutno savršeni.

### Utjecaj načina izračuna težinkih faktora na rezultate klasifikacije

Tablica 6.7 prikazuje utjecaj dva korištena načina za izračunavanje težinskih faktora (engl. *weighting scheme*) kod značajki vreće riječi na klasifikaciju poruka prema važnosti. Dvije korištene metode su TF i TF-IDF, a prikazani su najbolji ostvareni rezultati za svaki klasifikator neovisno o korištenom skupu značajki. Osim u slučaju klasifikatora SVM, dobiveni su gotovo jednaki rezultati za obje korištene metode, što ukazuje na neovisnost klasifikatora o načinu izračuna težinskih faktora. Klasifikator SVM pokazao je prilično loše rezultate u svim slučajevima kad je korištena TF metoda te je očito da je u ovom slučaju za taj klasifikator mnogo bolji izbor korištenje TF-IDF metode.

Taj rezultat za SVM jer vrlo zanimljiv posebice u kontekstu rezultata klasifikacije govornih činova, gdje je u većini slučajeva SVM najbolje radio baš uz značajke TF. Od ukupno pet naučenih SVM klasifikatora koji se koriste za označavanje govornih činova (jedan bez ujednačavanja značajki, četiri uz ujednačavanje skupa značajki), samo jednom je izabran SVM klasifikator sa značajkama TF-IDF naspram klasifikatora sa značajkama TF. Budući da se u oba slučaja klasifikacije koriste iste poruke, odnosno značajke su gotovo iste riječi, ovaj rezultat može ukazivati na to koliko su problem

**Tablica 6.8:** Utjecaj ujednačavanja skupa značajki na performanse klasifikatora (%F1)

| Klasifikator | Bez ujednačavanja | Uz ujednačavanje |
|---|---|---|
| NB | 70,02 | 70,02 |
| k-NN | 72,24 | 72,00 |
| SVM | 78,24 | 75,61 |
| DS | 37,45 | 37,45 |
| AB | 66,49 | 62,21 |
| *Ukupno* | 78,24 | 75,61 |

označavanja govornih činova i problem klasifikacije važnosti poruka zapravo različiti zadaci, iako je pokazano koliko govorni činovi mogu pozitivno utjecati na klasifikaciju po važnosti.

### Utjecaj ujednačavanja skupa značajki na rezultate klasifikacije

Kao što je opisano u odjeljku 5.4, prilikom primjene klasifikatora govornih činova na klasifikaciju na temelju važnosti, iskušana je i metoda kojoj je cilj bio ujednačiti promatrane skupove značajki za obje vrste klasifikatora. Svrha tog postupka bio je onemogućiti klasifikatore govornih činova da nauče kako su pojedine riječi koje se kasnije neće pojavljivati prilikom klasifikacije prema važnosti zapravo važne za označavanje. Pretpostavka je bila da će se tom metodom rezultati poboljšati jer će klasifikatori govornih činova bolje reagirati na skup značajki korišten u klasifikaciji prema važnosti. No, prema rezultatima prikazanim u Tablici 6.8, to se nije pokazalo točnim. U slučaju naivnog Bayesovog klasifikatora i klasifikatora *Decision Stump*, ujednačavanje skupa značajki ne mijenja ostvareni rezultat. No, u svim ostalim slučajevima, performanse klasifikatora uz ujednačavanje skupova značajki su se pokazale lošijima od korištenja neujednačenih skupova. Najočitiji primjer razlike je kod klasifikatora AdaBoost koji u slučaju ujednačavanja značajki gubi $4{,}28\%$ F1-mjere. Iako gubici nisu veliki, dobiveni rezultati su vrlo važni jer opovrgavaju pretpostavku da bi ujednačavanje skupa značajki trebalo dovesti do poboljšanja rezultata klasifikacije. Jedno moguće objašnjenje za takav rezultat jest da postoji mogućnost da se ujednačavanjem skupa podataka, odnosno izbacivanjem nekih značajki koje su imale veću diskriminatornu vrijednost za pojedine klasifikatore govornih činova, veća težina dala značajkama koje možda nisu dovoljno diskriminativne, čime se povećala sigurnost klasifikatora u pripadnost jednoj

od klasa za primjere koji su ranije bili možda na granici klasifikacije. Time se na razini označavanja govornih činova postižu slabiji rezultati, što se odražava na značajke koje se pridružuju značajkama vreće riječi kod klasifikacije prema važnosti te to dovodi do lošijih rezultata. Tome u prilog govore i rezultati, koji ne pokazuju nikakvu razliku u slučaju klasifikatora koji značajkama prilikom učenja ne pridružuju težine poput naivnog Bayesovog klasifikatora ili klasifikatora *Decision Stump*. Također, razlika između dva navedena slučaja za klasifikator k-NN jest minorna, što također govori u prilog prethodnoj tezi, budući da k-NN računa udaljenost između primjera, a ne pridjeljuje težine pojedinim značajkama. Uz to, najveća razlika je prisutna upravo kod AdaBoost klasifikatora koji se temelji na pridjeljivanju težina pojedinim značajkama.

Ovi su rezultati vrlo značajni jer ukazuju i na to da su klasifikatori govornih činova vrlo otporni na promjenu skupa značajki, što daje nadu da ih je moguće iskoristiti i nad skupovima poruka koje su potpuno nevezane uz skup nad kojim su učene, dok god se radi o sličnom tipu poruka elektroničke pošte.

## 6.3. Analiza utjecaja značajki klasifikatora važnosti

Nakon što je analiza performansi klasifikatora pokazala važnost dodavanja oznaka govornih činova u skup značajki za klasifikaciju prema važnosti, bilo je potrebno provjeriti na koji točno način pridodane značajke utječu na klasifikaciju. Najprikladniji način za izvedbu takve provjere sastoji se od analize utjecaja pojedinih značajki na ukupnu klasifikaciju. Postoji više metoda za analizu utjecaja, no svima je zajedničko da u konačnici proizvode listu u kojoj su značajke uređene prema utjecaju tako da se značajkama koje više utječu na klasifikaciju pridaje veći koeficijent, dok se manje utjecajnim značajkama pridaje manji. U ovom radu navedeni koeficijenti su normalizirani kako bi se dobila raspodjela koeficijenata u rasponu od 0 do 1 te je zbroj svih koeficijenata 1. Time se omogućuje kasnija usporedba vrijednosti između različitih metoda.

Kako konačan rezultat ne bi previše ovisio o jednoj izabranoj metodi analize, u obzir je uzeto ukupno šest različitih metoda. Budući da svaka od metoda vraća normaliziranu vrijednost koeficijenta utjecaja za svaku značajku u ulaznom skupu, dobiveni rezultati su uprosječeni i taj prosjek je iskorišten kao vrijednost koeficijenta utjecaja za svaku od značajki. Analiza je izvršena nad cijelim skupom podataka za klasifikaciju prema važnosti, uključujući i skup za ispitivanje.

**Korištene metode**

**$\chi^2$–analiza** procjenjuje vrijednost svake značajke na temelju $\chi^2$–statistike Bienayme (1838).

**Analiza informacijskog dobitka** vrijednost značajke računa kao dobitak informacije (smanjenje entropije) s obzirom na klasu uvođenjem navedene značajke. Koristi se formula:

$$\text{InfoG}(K,A) = H(K) - H(K|A),$$

gdje je $H(K)$ entropija klase, a $H(K|A)$ entropija klase uz uvođenje nove značajke.

**Analiza omjera dobitka** računa vrijednost značajke kao omjer dobitka s obzirom na klasu odnosno normalizira se prethodna metoda s obzirom na entropiju značajke. Koristi se sljedeća formula:

$$\text{GainR}(K,A) = \frac{H(K) - H(K|A)}{H(A)},$$

gdje je $H(K)$ entropija klase, $H(K|A)$ entropija klase uz dodanu značajku, a $H(A)$ entropija značajke.

**Analiza glavnih komponenti (PCA)** je metoda koja koristi ortogonalnu transformaciju kako bi skup varijabli koje su možda korelirane pretvorio u skup vrijednosti koje su međusobno linearno nezavisne i nazivaju se glavnim komponentama (Pearson, 1901).

**Analiza "Relief"** Kira i Rendell (1992) koriste višestruko uzorkovanje kako bi se ispitala vrijednost značajke za najbliži primjer iste i različite klase od uzorkovanog primjera.

**Analiza simetrične nesigurnosti** Witten i Frank (2005) računa simetričnu nesigurnost s obzirom na klasu za danu značajku. Koristi se sljedeća formula:

$$\text{SymmU}(K,A) = 2 \times \frac{H(K) - H(K|A)}{H(K) + H(A)},$$

gdje je $H(K)$ entropija klase, $H(K|A)$ entropija klase uz dodanu značajku, a $H(A)$ entropija značajke.

Iz Tablice 6.9 vidljivo je kako se sve oznake govornih činova nalaze unutar $1\%$ najvažnijih značajki ako se kao mjerilo uzima prosječna normalizirana važnost dobivena

**Tablica 6.9:** Važnost oznaka govornih činova kao značajki

| | Srednja važnost | Rang | Percentil |
|---|---|---|---|
| ZAHTIJEVATI | 0,672 | 1 | 0,0175 |
| ISPORUČITI | 0,521 | 3 | 0,0525 |
| PODSJETITI | 0,276 | 16 | 0,2803 |
| IZMIJENITI | 0,174 | 52 | 0,91 |
| OBVEZATI_SE | 0,165 | 59 | 1,03 |
| PREDLOŽITI | 0,163 | 62 | 1,086 |

od korištenih metoda. To izravno govori o informativnosti navedenih značajki kao i o njihovoj važnosti za klasifikaciju. Dobiveni rezultati mogu se vrlo lako objasniti ako se pobliže pogleda raspored i važnost pojedinih govornih činova i uzme u obzir lista kriterija kojima su se vodili označivači prilikom označavanja poruka

Primjerice, oznaka prisutnosti govornog čina ZAHTIJEVATI izdvojena je kao najvažnija značajka u skupu podataka. Poruke elektroničke pošte u kojima se od primatelja nešto zahtijeva ili očekuje neka radnja, prema uvedenim kriterijima redom su se smatrale važnima, tako da je udio zahtjeva i molbi u skupu podataka s važnim porukama bio dosta velik. Stoga i ne čudi da ukoliko određena poruka s velikom vjerojatnošću sadrži govorni čin ZAHTIJEVATI, velika je vjerojatnost i da će ista biti smatrana važnom porukom. Tome u prilog govori i druga značajka po važnosti, a to je riječ *molim*. Ta riječ vrlo je česta u svim oblicima zahtjeva i stoga ne čudi da je i njezina prisutnost ili odsutnost od velikog utjecaja na važnost poruka. No, i dalje je oznaka govornog čina zahtijevati, s prosječnom mjerom važnosti od $0{,}672$, daleko najvažnija značajka za klasifikaciju. Iako je važnost oznake govornog čina ZAHTIJEVATI bilo vrlo jednostavno logički objasniti, važno je i istaknuti činjenicu da je prema podacima iz Tablice 6.2 klasifikator za taj govorni čin pokazao najlošije performanse za razinu dokumenta na kojoj je korišten te uz F1-mjeru od svega $70{,}09\%$ ipak uspijeva biti najvažnija komponenta klasifikacije poruka prema važnosti. To ukazuje na mogućnost da bi se poboljšanjem performansi navedenog klasifikatora, primjerice korištenjem na razini odlomka ili rečenice ili primjenom druge vrste klasifikatora, moglo dobiti i bolje rezultate za klasifikaciju važnosti.

Iduća prema važnosti je oznaka za govorni čin ISPORUČITI, što je također u skladu s kriterijima za označavanje, pogotovo s kriterijem koji se odnosi na brisanje poruke i gubitak važnih informacija. Gotovo svaka poruka koja je sadržavala taj govorni čin

**Tablica 6.10:** Karakteristične riječi za govorne činove

| | Najkarakterističnije riječi |
|---|---|
| ISPORUČITI | prilog, salj*(-em)*, canad*(-i)*, mi, matematik*(-a)*, bud*(-em)*, attach |
| IZMIJENITI | dod*(-am)*, algorit*(-am)*, sam, tako, tek, verzij*(-a,-u)*, tekst |
| OBVEZATI_SE | cu, adres*(-a)*, agreement, bud*(-em)*, jav*(-im)*, cim, automatsk*(-i,-o)* |
| PODSJETITI | podsjet*(-it)*, vas, cemo, ja, dogovor, svim, srijed*(-a,-u)* |
| PREDLOŽITI | bi, da, prilog, predl, moze, ako, mozd*(-a)* |
| ZAHTIJEVATI | mol*(-im)*, prilog, ic, li, da, vas, ajde |

smatrana je važnom, stoga ne čudi da je oznaka za taj govorni čin ukupno treća po važnosti od svih značajki, uz vrlo visoku prosječnu mjeru važnosti od $0{,}521$. Uz to, klasifikator koji je korišten za označavanje navedenog govornog čina jer prema Tablici 6.2 pokazao vrlo dobre performanse, uz F1-mjeru od $86{,}59\%$. Sve to ukazuje na to da je oznaka govornog čina ISPORUČITI od vrlo visokog značaja za uspješnu klasifikaciju poruka prema važnosti, te da uz oznaku govornog čina ZAHTIJEVATI čini par značajki koje su odgovorne za većinu poboljšanja koje oznake govornih činova donose prilikom klasifikacije poruka prema važnosti.

Preostale oznake govornih činova imaju nešto nižu prosječnu mjeru važnosti, no i dalje su vrlo značajne za ukupnu klasifikaciju poruka prema važnosti. S lingvističkog stajališta, navedeni govorni činovi ne bi mogli biti isključivo vezani uz važne poruke, jer mogu biti korišteni i u odlomcima teksta koji nisu od pretjerane važnosti za primatelja. Primjerice, pošiljatelj poruke može primatelja podsjetiti na neku relativno nebitnu ili banalnu stvar te poruka koja sadrži takvu vrstu podsjetnika nije stvarno važna poruka poput onih koje sadrže zahtjeve ili isporučuju nešto primatelju. Također, govorni čin OBVEZATI_SE može biti prisutan u porukama koje nisu važne, ali primatelju pružaju informaciju o tome da se pošiljatelj na nešto obavezao. Te poruke uglavnom ne očekuju od primatelja neku vrstu povratne radnje, već jedino prenose informaciju koja krajnjem čitatelju može ili ne mora biti od važnosti, ovisno o kontekstu razgovora. No, naravno, sama činjenica da su sve oznake govornih činova unutar $1\%$ najvažnijih značajki govori o tome kako je značajan njihov utjecaj na klasifikaciju prema važnosti te kako se poruke koje sadrže neke od navedenih govornih činova uglavnom smatraju važnima.

Koristeći istu metodologiju koja je korištena za procjenu utjecaja govornih činova kao značajki prilikom klasifikacije prema važnosti provedeno je i istraživanje o karak-

terističnim riječima vezanim uz pojedine govorne činove. Time je za svaki čin dobiven karakterističan jezični profil koji sadrži riječi koje su od velikog utjecaja prilikom označavanja govornih činova. Tablica 6.10 prikazuje dobivene rezultate nakon izbora sedam najutjecajnijih riječi za svaki govorni čin. Vidljivo je kako dobivene riječi odgovaraju pojmovima koje inače vezujemo uz navedene govorne činove. Primjerice, za govorni čin ISPORUČITI karakteristična je riječ *prilog*, dok je za govorni čin ZAHTIJEVATI karakteristična riječ *molim*. Povezanost između karakterističnih riječi i oznaka govornih činova može također poslužiti i kao objašnjenje poklapanja F1-mjera za jednostavnije klasifikatore (naivan Bayes i *Decision Stump*) prilikom korištenja različitih vrsta značajki, budući da je moguće da navedeni klasifikatori jednako promatraju pojavu riječi *molim* i pojavu govornog čina ZAHTIJEVATI.

# 7. Zaključak

Govorni činovi su ilokucijski činovi koji imaju određeno značenje koje govornik želi prenijeti na slušatelja (Searle, 1965), odnosno označavaju radnje koje činimo pomoću govora, bilo da se radi o razgovoru ili o pisanoj komunikaciji. U okviru ovog rada pokazano je kako su govorni činovi učinkovit način za sažimanje sadržaja i namjene poruke elektroničke pošte. Problem klasifikacije govornih činova promatran je u okviru problema klasifikacije poruka elektroničke pošte na hrvatskom jeziku. Pristup problemu zasnovao se na klasifikaciji s više oznaka koristeći detaljno vrednovanje šest algoritama strojnog učenja koji su učeni na tri razine (poruka, odlomak i rečenica) te tri vrste značajki. Pokazano je da razina na kojoj su učeni klasifikatori te tip značajki ne utječu previše na sposobnost klasifikacije. No, pokazano je i da pojedini govorni činovi mogu biti uspješnije označeni ukoliko se koristi odgovarajuća razina. Prilikom korištenja značajki na razini poruke, postignuta je najbolja F1-mjera od $94\%$. Rezultati dobiveni u ovom radu su značajno bolji u pogledu F1-mjere od rezultata dobivenih u prethodnim radovima koji koriste engleski jezik.

Također, pokazano je i kako je moguće pristupiti klasifikaciji poruka na temelju važnosti na objektivan način i definirati objektivne kriterije koje moraju zadovoljiti važne poruke. Korištenjem skupa značajki vreće riječi postignuta je F1-mjera od $70{,}07\%$ za klasifikaciju važnosti poruka. Primjenom oznaka govornih činova, F1-mjera je poboljšana za $8{,}17\%$ što je dokaz pozitivnog utjecaja oznaka govornih činova na klasifikaciju prema važnosti. Uz to, sve značajke koje su se temeljile na oznakama govornih činova prema procijeni utjecaja nalaze se unutar $1\%$ najutjecajnijih značajki za klasifikaciju poruka prema važnosti. Sve to svjedoči o iznimnoj korisnosti oznaka govornih činova u sklopu rješavanja problema automatske klasifikacije poruka elektroničke pošte.

Rezultati izneseni u ovom radu, pogotovo u pogledu korisnosti korištenja oznaka govornih činova kod klasifikacije poruka prema važnosti otvaraju jedno potpuno novo područje u polju klasifikacije važnosti poruka budući da je po prvi put ispitano kako se informacije o sadržaju poruke koje se nalaze u oznakama govornih činova mogu

primijeniti za daljnju klasifikaciju poruka. Ovim se otvara prostor korištenju i drugih značajki vezanih uz sadržaj poruke kako bi se poruka mogla uspješnije klasificirati na temelju važnosti.

Ovaj rad nije se bavio problematikom praktične korisnosti dobivenih oznaka za krajnjeg korisnika u pogledu skraćivanja vremena potrebnog za rad s porukama elektroničke pošte. U tom pogledu moguće je razviti klijent elektroničke pošte koji bi korisnicima automatski označavao važnost poruka na temelju govornih činova, kao i same činove u tekstu poruke te ispitati utjecaj takvog klijenta na produktivnost. Budući da su u ovom radu korišteni samo klasifikatori govornih činova na razini poruke, bilo bi zanimljivo ispitati kako bi korištenje klasifikatora na razini odlomaka ili rečenica za govorne činove kod kojih je klasifikacija bolja na razinama nižim od razine poruke utjecala na klasifikaciju prema važnosti.

Uz oznake govornih činova, u budućem radu bi se za procjenu važnosti mogle iskoristiti i druge vrste značajki temeljenih na sadržaju poruke, kao što su vremenske oznake ili imenovani entiteti. Također bi se, umjesto binarne klasifikacije na važne i nevažne poruke, mogla iskoristiti regresija s ciljem procjene koeficijenta važnosti poruke.

# LITERATURA


Douglas Aberdeen, Ondrej Pacovsky, i Andrew Slater. The Learning Behind Gmail Priority Inbox. Technical report, Google, Inc., 2010.

Irénée-Jules Bienayme. Memoire sur la probabilite des resultats moyens des obseravations; demonstration directe de la regle de Laplace. *Memoires de l' Academie de Sciences de l' Institut de France*, 5:513–558, 1838.

Martina Blečić. Govorni činovi. Magistarski rad, Sveučilište u Rijeci, Filozofski fakultet u Rijeci, 2010.

Jean Carletta. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, stranice 249–254, 1996.

Vitor R. Carvalho i William W. Cohen. Improving "Email Speech Acts" Analysis Via N-gram Selection. U *Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech*, stranice 35–41, 2006.

Chih-Chung Chang i Chih-Jen Lin. *LIBSVM: A Library for Support Vector Machines*, 2001. Dostupno na `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 1960.

William W. Cohen, Vitor R. Carvalho, i Tom M. Mitchell. Learning to Classify Email into "Speech Acts". U *Proceedings of EMNLP 2004*, stranice 309–316, 2004.

Corinna Cortes i Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, stranice 273–297, 1995.

Laura A. Dabbish i Robert E. Kraut. Email Overload at Work: An Analysis of Factors Associated With Email Strain. U *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*, 2006.

Laura A. Dabbish, Robert E. Kraut, Susan Fussell, i Sara Kiesler. Understanding Email Use: Predicting Action on a Message. U *CHI 2005*, stranice 691–700, 2005.

Hercules Dalianis, Jonas Sjöbergh, i Eriks Sneiders. Comparing Manual Text Patterns and Machine Learning for Classification of E-Mails for Automatic Answering by a Government Agency. U *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part II*, stranice 234–243, 2011.

Michael Finke, Maria Lapata, Alon Lavie, Lori Levin, Laura Mayfield Tomokiyo, Thomas Polzin, Klaus Ries, Alex Waibel, i Klaus Zechner. CLARITY: Inferring Discourse Structure from Speech. U *Procceedings of the Workshop on Applying Machine Learning to Discourse Processing*, 1998.

Yoav Freund i Robert E. Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

Matt E. Hart. Method and Apparatus for Determining the Importance of Email Messages. Patent, siječanj 2008. US 2008/0005249 A1.

Simon Keizer. A Bayesian Approach to Dialogue Act Classification. U *BI-DIALOG 2001: Proc. of the 5th Workshop on Formal Semantics and Pragmatics of Dialogue*, stranice 210–218, 2001.

Jihie Kim, Grace Chern, Donghui Feng, Erin Shaw, i Eduard Hovy. Mining and Assessing Discussions on the Web Through Speech Act Analysis. U *Proceedings of the ISWC'06 Workshop on Web Content Mining with Human Language Technologies*, 2006.

Kenji Kira i Larry A. Rendell. A Practical Approach to Feature Selection. U *Proceedings of the 9th International Workshop on Machine Learning*, stranice 249–256. Morgan Kaufmann Publishers Inc., 1992.

Bryan Klimt i Yiming Yang. The Enron Corpus: A New Dataset for Email Classification Research. U *Machine Learning: ECML 2004*, stranice 217–226. 2004.

Max M. Louwerse i Scott A. Crossley. Dialog Act Classification Using N-Gram Algorithms. U *FLAIRS Conference*, stranice 758–763, 2006.

Johanna Marineau, Peter Wiemer-Hastings, Derek Harter, Brent Olde, Patrick Chipman, Ashish Karnavat, Victoria Pomeroy, Victoria Graesser, i the Tutoring Research Group. Classification of Speech Acts in Tutorial Dialog. U *Proceedings of the Workshop on Modeling Human Teaching Tactics and Strategies at the Intelligent Tutoring Systems 2000 Conference*, stranice 65–71, 2000.

Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz, i Timm Euler. YALE: Rapid Prototyping for Complex Data Mining Tasks. U *KDD '06: Proceedings of the 12th ACM SIGKDD International Vonference on Knowledge Discovery and Data Mining*, stranice 935–940, 2006.

Karl Pearson. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2:559–572, 1901.

Sujith Ravi i Jihie Kim. Profiling Student Interactions in Threaded Discussions with Speech Act Classifiers, internal project report, 2007.

Georgios Sakkis, Ion Androutsopoulos, Georgios Paliouras, Vangelis Karkaletsis, Constantine D. Spyropoulos, i Panagiotis Stamatopoulos. Stacking Classifiers for Anti-Spam Filtering of E-Mail. *Empirical Methods in Natural Language Processing*, stranice 44–50, 2001.

Georgios Sakkis, Ion Androutsopoulos, i Constantine D. Spyropoulos. A Memory-Based Approach to Anti-Spam Filtering for Mailing Lists. *Information Retrieval*, 6: 49–73, 2003.

Niamh C. Scannell, Stuart D. Dawson, Anthony J. Redmond, Serge Himbaut, Pascale Bares, Villeneuve Loubet, i Alison Clark. Method and System for Sorting and Prioritizing Electronic Mail Messages. Patent, prosinac 1994. US005377354A.

Simon Scerri, Brian Davis, Siegfried Handschuh, i Manfred Hauswirth. Semanta – Semantic Email Made Easy. U *ESWC '09*, stranice 36–50, 2009.

John R. Searle. What is a Speech Act? *The Philosophy of Language, Oxford University Press*, stranice 44–46, 1965.

Riccardo Serafin, Barbara Di Eugenio, i Michael Glass. Latent Semantic Analysis for Dialogue Act Classification. U *Proceedings of HLT-NAACL 2003–short papers*, svezak 2 od *NAACL-Short '03*, stranice 94–96, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

Terry Winograd. A Language/Action Perspective on the Design of Cooperative Work. *Human – Computer Interaction*, stranice 3–30, 1987.

Ian H. Witten i Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Amsterdam, 2005.

Shinjae Yoo, Yiming Yang, Frank Lin, i Il-Chul Moon. Mining Social Networks for Personalized Email Prioritization. U *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, stranice 967–976, New York, NY, USA, 2009.

Bo Yu i Zong-ben Xu. A Comparative Study for Content-Based Dynamic Spam Classification Using Four Machine Learning Algorithms. *Knowledge-Based Systems*, 21(4):355–362, 2008.

Le Zhang, Jingbo Zhu, i Tianshun Yao. An Evaluation of Statistical Spam Filtering Techniques. *ACM transactions on Asian Language Information Processing*, 3(4): 243–269, 2004.

Jan Šnajder. *Morfološka normalizacija tekstova na hrvatskome jeziku za dubinsku analizu i pretraživanje informacija*. Doktorska disertacija, Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva, 2010.

**Klasifikacija važnosti poruka elektroničke pošte temeljem govornih činova**

## Sažetak

Razmjena informacija putem poruka elektroničke pošte zauzima sve veći udio u poslovnoj i osobnoj komunikaciji. Automatska klasifikacija poruka prema važnosti korisniku omogućava kvalitetniju obradu poruka i uštedu vremena. Tipični sustavi za klasifikaciju važnosti poruka temelje se ne modelu tzv. vreće riječi. S komunikacijskog aspekta, veći značaj od samih riječi imaju tzv. govorni činovi, odnosno radnje izražene govorom ili pismom (zahtjev, izmjena, isporuka i sl). Budući da govorni činovi neupitno utječu na važnost poruke, pretpostavlja se da klasifikacija važnosti poruke temeljena na govornim činovima može dati bolje rezultate od uobičajene klasifikacije temeljene na riječima. U okviru ovog rada proučeni su postojeći postupci za određivanje važnosti poruka elektorničke pošte temeljeni na metodama strojnog učenja te teorija govornih činova. Predložen je postupak za označavanje poruka elektroničke pošte na hrvatskom jeziku govornim činovima. Također, predložen je postupak za klasifikaciju važnosti poruka elektroničke pošte na hrvatskome jeziku koji kombinira klasifikaciju temeljenu na govornim činovima i sadržajnu klasifikaciju temeljenu na modelu tzv. vreće riječi. Nad predloženim postupcima ispitano je šest različitih algoritama nadziranog strojnog učenja te je proveden postupak procjene utjecaja značajki na klasifikaciju. Provedeno je eksperimentalno vrednovanje točnosti označavanja govornih činova i klasifikacije važnosti te analiza utjecaja oznaka govornih činova na klasifikaciju.

**Ključne riječi:** strojno učenje, nadzirano učenje, obrada prirodnog jezika, klasifikacija elektroničke pošte, govorni činovi

**Classification of Email Importance Based on Speech Acts**

## Abstract

Information sharing through email is fastly becoming an integral part of everyday business and personal communication. Automatic classification of messages based on their importance provides the user with high-quality message processing while saving time. Typical classifiers of message importance are based on the bag-of-words model. From the communication standpoint, speech acts are more important than single words. Speech acts are actions performed with words, in writing or orally (request, amendment, delivery, etc). Since speech acts undoubtedly affect message importance, it is assumed that speech act-based classification of message impotance could show better results than classification based on words. This work assesses the existing procedures for importance based message classification using machine learning methods as well as the theory of speech acts. A procedure is proposed for labelling email messages in Croatian language using speech acts. Futhermore, the work proposes a method for classifying the importance of email messages which combines speech-act based classification with content-based classification using the bag-of-words model. For the proposed methods, six supervised machine learning algorithms were tested and estimation of the effect of the features was performed. Experimental evaluation of the accuracy in speech act labelling and importance classification was performed along with the analysis of the effect of speech act labels on classification performance.

**Keywords:** machine learning, supervised learning, natural language processing, email message classification, speech acts