TakeLab

Laboratorij za analizu teksta i inženjerstvo znanja
Text Analysis and Knowledge Engineering Lab
Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
Unska 3, 10000 Zagreb, Hrvatska

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 5330

Predviđanje vijestodostojnosti novinskih članaka pomoću strojnog učenja

Antonio Šajatović

Zagreb, lipanj 2017.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA
ODBOR ZA ZAVRŠNI RAD MODULA

Zagreb, 3. ožujka 2017.

ZAVRŠNI ZADATAK br. 5330

Pristupnik: **Antonio Šajatović (0036478985)**
Studij: Računarstvo
Modul: Računarska znanost

Zadatak: **Predviđanje vijestodostojnosti novinskih članaka pomoću strojnog učenja**

Opis zadatka:

Računalna analiza događaja opisanih u novinskim tekstovima aktivno je područje istraživanja u okviru obrade prirodnog jezika. Pored informacijske vrijednosti, važan aspekt svakog novinskog članka jest njegova vijestodostojnost, odnosno kriteriji koji određuju istaknutost i percipiranu vrijednost novinskog članka kod čitateljstva. Računalno predviđanje vijestodostojnosti nov je smjer istraživanja s primjenama u praćenju događaja te sažimanju dokumenata, a potencijalnu primjenu ima i u društvenim znanostima.
Tema završnoga rada jest predviđanje vijestodostojnosti novinskih članaka na hrvatskome i engleskome jeziku temeljeno na strojnom učenju. Proučiti postupke za klasifikaciju teksta temeljene na strojnom učenju te postojeće pristupe za automatsku analizu vijestodostojnosti novinskih članaka. Razviti model za klasifikaciju vijestodostojnosti na temelju naslova članaka ili punog teksta članka u kategorije vijestodostojnosti koje su predložili Harcup i O'Neill (2011). Izraditi prikladnu zbirku novinskih članaka ručno označenih kategorijama vijestodostojnosti. Razmotriti diskriminativne modele temeljene na reprezentacijama riječi te generativne modele koji modeliraju interakciju između značajki. Provesti vrednovanje modela, usporedbe s referentnim modelom, statističku obradu rezultata te analizu pogrešaka. Radu priložiti izvorni i izvršni kod razvijenog sustava, označene skupove podataka i potrebnu dokumentaciju te citirati korištenu literaturu.

Zadatak uručen pristupniku: 10. ožujka 2017.
Rok za predaju rada: 9. lipnja 2017.

Mentor:

Izv. prof. dr. sc. Jan Šnajder

Djelovođa:

Doc. dr. sc. Tomislav Hrkać

Predsjednik odbora za završni rad modula:

Prof. dr. sc. Siniša Srbljić

Zahvaljujem svojoj obitelji i prijateljima na podršci, razumijevanju i pomoći tijekom cijelog studija.
Zahvaljujem mentoru Janu Šnajderu i suradnici Mariji Piji Di Buono na motivaciji, savjetima, strpljenju i prilici da budem dio TakeLaba.

SADRŽAJ

1. Uvod

U današnje vrijeme svakodnevno se generira ogromna količina informacija, a sve brža globalizacija i digitalizacija diljem svijeta omogućuju pohranjivanje sve većih količina tih informacija. Većina njih pohranjena je u nestrukturiranom tekstnom obliku. Istovremeno, ljudski kapaciteti pamćenja i pažnje ostaju gotovo nepromijenjeni, zbog čega sustavi za analizu i pretraživanje teksta postaju vrlo značajni u obradi i analizi informacija. Umjetna inteligencija postaje neizostavan dio rješenja navedenog problema. Najistaknutija grana umjetne inteligencije u navedenoj domeni danas jest strojno učenje (engl. *machine learning*), koje je proizašlo iz raspoznavanja uzoraka, a bavi se oblikovanjem algoritama koji mogu učiti i raditi predviđanja na temelju podataka. Srodna grana računarske znanosti koja se bavi interakcijom između računala i prirodnog jezika naziva se obrada prirodnog jezika (engl. *natural language processing*). Zbog obećavajućih dosadašnjih rezultata i mogućnosti rješavanja dosad nerješivih problema u navedenim domenama, strojno učenje i obrada prirodnog jezika danas su jedna od najistaknutijih područja računarske znanosti.

Prilikom odabira događaja koje će popratiti, novinari žele utvrditi koju će razinu istaknutosti dati određenom događaju i koliku će količinu pažnje publika posvetiti toj vijesti. Za navedene potrebe služi im vijestodostojnost (engl. *newsworthiness*) – generalno definiran skup vrijednosti u novinarstvu, koje se često nazivaju vrijednosti vijesti (engl. *news values*) (Boyd, 2001).

Tema ovog rada jest klasifikacija i analiza interakcije emocija i vrijednosti vijesti naslova novinskih članaka na engleskom jeziku korištenjem metoda strojnog učenja. Za potrebe predviđanja emocija i vrijednosti vijesti razvijena su dva generativna modela, Naivan Bayesov klasifikator (engl. *Naive Bayes classifier*) i Bayesova mreža (engl. *Bayesian network*), i diskriminativni model Stroj potpornih vektora (engl. *Support Vector Machine*, SVM). Cilj je bio istražiti međusobni utjecaj vrijednosti vijesti i emocija, a kako bi to bilo ostvareno, kombinirani su generativni i diskriminativni modeli. Prikupljen je i označen prikladan skup podataka naslova novinskih članaka, nad kojima su provedeni učenje i evaluacija modela.

U poglavljima koja slijede najprije je dan pregled srodnih radova. U idućem poglavlju opisan je skup podataka i metodologija označavanja podataka. Zatim su opisani modeli korišteni u rješavanju problema, a postavke eksperimenata i ostvareni rezultati predviđanja prodiskutirani su u poglavlju pet. Naposljetku, zaključak se nalazi u poglavlju šest.

2. Srodni radovi

Znanstveni rad koji je inicijalno motivirao ovaj rad jest (Harcup i O'Neill, 2016). Autori u njemu revidiraju prethodno ustanovljene vrijednosti vijesti (Harcup i O'neill, 2001), potaknuti sve većom zastupljenošću digitalnih medija u objavi vijesti i ulozi društvenih mreža u dijeljenju objavljenih vijesti. Prikupili su skup od ukupno 711 novinskih članaka vodećih zavisnih i nezavisnih novinskih agencija u Ujedinjenom kraljevstvu, istovremeno objavljenih u rasponu od mjesec dana. Označivači su pročitali svaki navedeni članak i označili ga prikladnom vrijednošću vijesti. Vrijednost vijesti kojoj je posvećena posebna pažnja jest *shareability*, koja kategorizira priče za koje se smatra da će najvjerojatnije biti dijeljene i komentirane na društvenim mrežama. U zaključku svog rada iznose 15 revidiranih vrijednosti vijesti: *ekskluzivnost, loša vijest, sukob, iznenađenje, audio-vizuali, shareability, zabava, drama, popratna vijest, moć i elita, relevantnost, važnost, celebrity, dobra vijest, agenda izdavača* (engl. *Exclusivity, Bad news, Conflict, Surprise, Audio-visuals, Shareability, Entertainment, Drama, Follow-up, The power elite, Relevance, Magnitude, Celebrity, Good news, News organisation's agenda*).

Sljedeći relevantan rad je (Upadhyay et al.) u kojem autori predlažu novi pristup za identifikaciju događaja koji je uzrok objave novinskog članka, polazeći od pretpostavke da su događaji opisani u novinskim člancima većinom pozadinski, tj. da ih većina njih nije vrijedna pažnje. Smatraju da identifikacija događaja uzročnika objave vijesti može biti korisna u detekciji prve objave (engl. *first story detection*), sažimanju teksta i u analizama događaja. Formaliziraju definiciju uzročnog događaja i daju mu naziv *news peg*, smatrajući da je on bolja mjera količine pažnje posvećene članku nego sažetak članka. Skup podataka bio je označeni korpus New York Times članaka. Vrijedi istaknuti da su označivači podataka postigli među-ocjenjivačku suglasnost (engl. *inter-rater agreement*) od čak 80%. Demonstriraju izvodljivost svega korištenjem klasifikatorom temeljenog na pravilima. Pravila na temelju kojih algoritam radi predviđanja su sintaktička pravila aktiv (glagolsko radno stanje), glavna klauza (zavisna rečenica) i prvi predikat u rečenici.

U radu (Nye i Nenkova, 2015) cilj je identifikacija bitnih informacija o vijesti temeljena na podacima. Autori se fokusiraju na analizu glagola u sažetku vijesti iz svijeta te su za potrebe toga izdvojili preko tisuću glagola i analizirali njihovu važnost na razini domene vijesti. Korišten je označen skup podataka izdavača New York Times. Rezultati analize obuhvaćenih glagola pokazuju da su glagoli koji se češće pojavljuju u sažetku pretežito negativni i označavaju akciju i neprijateljstvo, dok oni koji su skloniji pojavljivati se u samom članku većinom opisuju osobne postupke. U zaključku navode činjenicu da dobiveni leksikon može biti korišten za predviđanje vijestodostojnosti novinskih članaka, pretpostavljajući da će u sažetku češće biti korišteni glagoli koji su bitniji za određivanje vrijednosti vijesti.

Tema ovog rada je, osim predviđanja vijestodostojnosti članka na temelju samog naslova, i predviđanje oznaka emocija te analiza interakcije između vrijednosti vijesti i emocija. Za razliku od pristupa u navedenim radovima, za predviđanje oznaka korišteni su samo naslovi članaka, što je otežalo predviđanje zbog kratke duljine naslova u odnosu na sažetak i sadržaj članka. Zbog toga je odabran i drugačiji pristup izvlačenja značajki (engl. *feature extraction*) pomoću kojih model uči predviđati, opisan u poglavlju 4. U poglavlju 2 opisan je korišten skup podataka, u kojem je korišten podskup oznaka vrijednosti vijesti preuzet od (Harcup i O'Neill, 2016).

3. Skup podataka

Skup podataka korišten za eksperimente u radu preuzet je od (Strapparava i Mihalcea, 2007). Cilj njihovog zadatka bio je za naslove novinskih članaka na engleskom jeziku predvidjeti intenzitet svake od šest emocija: *bijes, gađenje, strah, radost, tuga i iznenađenje* (engl. *anger, disgust, fear, joy, sadness, surprise*). Korpus se sastojao od naslova novinskih članaka preuzetih s internetskih stranica ili iz novinskih članaka. Za rješavanje zadatka bila su dostupna dva skupa podataka: razvojni skup s 250 označenih primjera i testni skup s 1000 označenih primjera. Autori su odlučili koristiti novinske članke zbog njihovog potencijala da izazovu emocije u čitatelju i kratke duljine od jedne rečenice. Korištene oznake za emocije su istinite oznake objavljene uz skup podataka. Vrijednost oznake pojedine emocije je iz intervala $[0 - 100]$, a za potrebe rada te su vrijednosti binarizirane, koristeći prag binarizacije iznosa 50. Navedeni skup podataka je trebalo označiti oznakama vrijednosti vijesti (jednom oznakom ili više njih) kako bi se mogla predviđati vijestodostojnost. Za potrebe označavanja unajmljena su četiri označivača i provedene su ukupno četiri kalibracijske runde označavanja kako bi se postigla što viša među-ocjenjivačku suglasnost (engl. *inter-rater agreement*). Statistička mjera za među-ocjenjivačku suglasnost između dvoje označivača jest Cohenova kappa (eng. *Cohen's kappa score, kappa*). Definirana je formulom

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

u kojoj je p_o empirijska vjerojatnost suglasnosti za oznaku dodijeljenu bilo kojem primjeru, a p_e hipotetska vjerojatnost suglasnosti pri nasumičnom označavanju primjera. Raspon vrijednosti *kappa*-statistike je iz intervala $[-1, 1]$. Vrijednosti manje od 0 označavaju nesuglasnost, vrijednost 0 nezavisnost, a vrijednosti veće od 0 suglasnost.

Tablica 3.1: Cohenovi kappa rezultati

oznaka	runda 1	runda 2	runda 3	runda 4	prosjek
ekskluzivnost	0.00	0.00	0.00	0.00	0.00
loša vijest	0.44	0.42	0.53	0.51	0.47
sukob	0.32	0.19	0.16	0.10	0.19
iznenađenje	0.06	0.01	0.12	0.05	0.06
audio-vizuali	0.00	/	/	/	0.00
shareability	0.09	0.09	0.01	0.00	0.05
zabava	0.61	0.50	0.50	0.50	0.53
drama	0.34	0.24	0.19	0.24	0.25
popratna vijest	0.17	0.11	0.10	0.03	0.10
moć i elita	0.35	0.36	0.29	0.44	0.36
relevantnost	0.00	0.04	/	0.00	0.01
važnost	0.10	0.20	0.04	-0.01	0.08
celebrity	0.57	0.31	0.52	0.64	0.51
dobra vijest	0.27	0.21	0.32	0.13	0.23
agenda izdavača	0.00	0.0	0.00	/	0.00
prosjek	0.22	0.19	0.21	0.20	0.21

U tablici 3.1 prikazane su prosječne vrijednosti *kappa*-statistike po rundi za svaku oznaku vrijednosti vijesti, zajedno s ukupnim prosječnim rezultatom svake runde u posljednjem retku, i prosječnim rezultatom za svaku od vrijednosti vijesti u posljednjem stupcu. Simbol / u tablici predstavlja odluku o izbacivanju određene vrijednosti vijesti zbog loše vrijednosti *kappa*-statistike iz jedne ili više prethodnih rundi. U posljednjoj rundi označavanja konačne oznake su donesene većinskim glasanjem troje označivača za svaki naslov članka. Konačna verzija skupa podataka sadrži ukupno 840 naslova novinskih članaka na engleskom jeziku, od kojih su svakom naslovu pridijeljene oznake u obliku binarnog indikatorskog vektora emocija i vrijednosti vijesti. Na temelju prosječnih rezultata svake vrijednosti vijesti iz posljednjeg stupca tablice 3.1 odlučeno je da se odbace četiri oznake vrijednosti vijesti te su preostale vrijednosti vijesti koje su korištene u ostatku rada sljedeće: *loša vijest, celebrity, sukob, drama, zabava, popratna vijest, dobra vijest, važnost, shareability, iznenađenje* i *moć i elita*.

4. Pristup zadatku i modeli

Predviđanje vijestodostojnosti i emocija naslova novinskih članaka modelirano je kao višeoznačni (engl. *multi-label*) klasifikacijski zadatak. Cilj treniranja (učenja) modela u nadziranom strojnom učenju jest na temelju skupa primjera i odgovarajućeg skupa oznaka pronaći najbolju hipotezu (Russell i Norvig, 2009). Najbolja hipoteza h je funkcija koja najbolje aproksimira funkciju $f(x) = y$, koja primjerima, tj. vektorima značajki $x = (x_1, x_2, ..., x_n)^T$ pridjeljuje istinite oznake y. Klasifikacija je kategorija strojnog učenja u kojoj oznake y mogu poprimiti samo diskretne vrijednosti. U višeoznačnoj klasifikaciji, svakom pojedinom primjeru dodijeljen je skup oznaka, što omogućava predviđanje izlaznih oznaka koje nisu međusobno isključive. U nastavku poglavlja pregled modela, osim modela Word2vec, načinjen je prema (Russell i Norvig, 2009).

4.1. Word2vec

Word2vec (*w2v*) je skupina modela koji se koriste za stvaranje vektorskih reprezentacija riječi razvijena u (Mikolov et al., 2013). Modeli *w2v* su plitke neuronske mreže s dva sloja, koje se treniraju za rekonstrukciju lingvističkog konteksta riječi. Ulazni podaci su korpus teksta, a izlazni vektorski prostor veličine nekoliko stotina dimenzija. Svakoj jedinstvenoj riječi u korpusu je dodijeljen odgovarajući vektor iz vektorskog prostora. Riječi sličnog konteksta u vektorskom prostoru su blisko pozicionirane, upravo kako se bi se sačuvale informacije o kontekstu. Demonstrativni primjer efikasnosti modela *w2v* je sljedeća vektorska operacija:

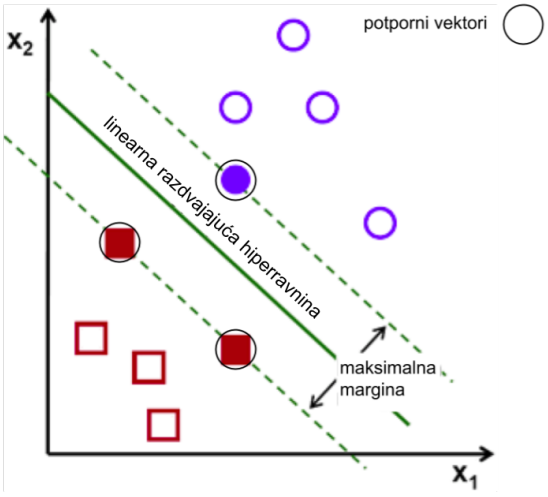
vektor('kralj') - vektor('muškarac') + vektor('žena') $\approx$ *vektor('kraljica')*

Korištenje modela *w2v* prikladno je za primjere s malim brojem riječi, kao što su naslovi članaka, zbog očuvanja informacija o kontekstu i dimenzionalnosti vektora značajki.

Pomoću programske knjižice *gensim* od (Řehůřek i Sojka, 2010) upotrijebljen je model *w2v* predtreniran na Google News skupu podataka, koji broji 100 milijardi riječi. Model sadrži 300-dimenzionalne vektore za 3 milijuna riječi i izraza. Svaki naslov novinskog članka razdvojen je na sastavne riječi, te su potom upotrebom regularnih izraza normalizirane brojke i uklonjeni interpunkcijski znakovi. Svaka riječ iz naslova potom je tokenizirana pomoću programske knjižice *nltk* od (Bird et al., 2009). Naposljetku je stvoren reprezentativni vektor cijelog naslova, kao srednja vrijednost odgovarajućeg *w2v* vektora svake tokenizirane riječi novinskog članka.

4.2. Stroj potpornih vektora

Stroj potpornih vektora je linearni diskriminativni model za nadzirano strojno učenje. SVM tijekom učenja stvara linearnu razdvajajuću hiperravninu (engl. *linear separating hyperplane*). Za donošenje odluke o pripadnosti kategoriji u klasifikaciji koristi marginu – udaljenost između hiperravnine i najbližeg primjera. Tijekom učenja pronalazi maksimalnu marginu, na temelju koje provodi binarnu klasifikaciju primjera. Vektori značajki na rubu margine nazivaju se potporni vektori i od njih potječe ime samog modela.



Slika 4.1: Hiperravnina i potporni vektori

Na slici 4.1 prikazan je jednostavan primjer linearne razdvajajuće hiperravnine, maksimalne margine i potpornih vektora u slučaju dvije značajke i dvije klase. [1]

[1]Preuzeto sa `http://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html`

Višeklasna (engl. *multiclass*) klasifikacija provodi se reduciranjem višeklasne klasifikacije u odgovarajući broj binarnih klasifikacija. Nerijetko se dogodi da podaci nisu zadovoljavajuće linearno razdvojivi u originalnom prostoru, no taj problem rješava se preslikavanjem podataka iz originalnog u višedimenzionalni prostor. Preslikavanjem u višedimenzionalni prostor u većini slučajeva je moguće podatke linearno razdvojiti hiperravninom u novom prostoru, što je poduprto Coverovim teoremom. Preslikavanje se obavlja korištenjem transformacijskih funkcija koje se još zovu i jezgrene funkcije (engl. *kernel functions*). Jezgrene funkcije definirane su kao mjere sličnosti između dva primjera. Vrlo je izgledno da u slučaju preslikavanja granica u originalnom prostoru neće biti linearna, i zbog toga se ova metoda svrstava u nelinearne metode klasifikacije. Jezgrena funkcija korištena u ovom radu jest (Gaussova) radijalna bazna funkcija (engl. *(Gaussian) Radial Basis Function, RBF*), definirana formulom

$$K(x_i, x_j) = exp(-\gamma \|x_i - x_j\|^2) \tag{4.1}$$

Tijekom učenja SVM-a s jezgrom RBF ključan je odabir hiperparametara C i γ, koji su u tom slučaju međusobno zavisni. Hiperparametar C koriste sve jezgrene funkcije i on odlučuje hoće li margina biti meka (engl. *soft margin*), odnosno hoće li model biti jednostavniji i dopuštati više pogrešnih klasifikacija, ili tvrda (engl. *hard margin*), odnosno hoće li model biti složen i pokušati ispravno klasificirati sve primjere. Hiperparametar γ predstavlja inverz utjecaja primjera koji su potporni vektori margine.

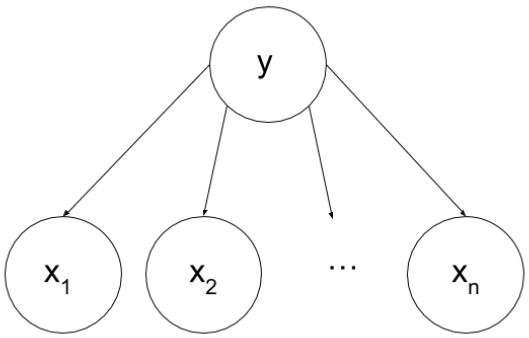
Za implementaciju SVM-a upotrijebljena je široko korištena knjižica za strojno učenje, *scikit-learn*. Korišten je `SVC` (*C-Support Vector Classification*) s jezgrom RBF. Višeklasna i višeoznačna klasifikacija implementirane su pomoću *wrapper* razreda `OneVsRestClassifier`, u kojeg je `SVC` ugniježđen kao estimator. Za svaku pojedinu oznaku stvoren je zaseban klasifikator, koji je treniran protiv klasifikatora za sve ostale oznake. Skup primjera činile su vektorske reprezentacije naslova, a skup oznaka istinite oznake vrijednosti vijesti i emocija. Na podskupu za učenje, kojeg čini nasumično odabranih 80% primjera iz cijelog skupa podataka, pomoću pretraživanja po rešetci (engl. *gridsearch*) odabrana je optimalna kombinacija hiperparametara C i γ, implementirana razredom `GridSearchCV`. Raspon pretrage optimalnih vrijednosti oba hiperparametra je od 2^{-15} do 2^{15}, u koraku po potencijama broja 2. Najbolji model naposljetku je treniran na cijelom podskupu za učenje i korišten za stvaranje predviđanja oznaka vrijednosti vijesti na testnom skupu.

4.3. Naivan Bayesov klasifikator

Kao referentni model korištena je verzija Naivnog Bayesovog klasifikatora. Naivni Bayesovi klasifikatori su skupina modela za nadzirano strojno učenje zasnovani na Bayesovom teoremu. Probabilistički model Naivnog Bayesovog klasifikatora izračunava uvjetne vjerojatnosti klase y uz uvjet vektora značajki $(x_1, ..., x_n)$ pomoću Bayesovog teorema

$$P(y \mid x_1, ..., x_n) = \frac{P(y)\, P(x_1, ..., x_n \mid y)}{P(x_1, ..., x_n)} \tag{4.2}$$

Naziv modela potječe od činjenice da pretpostavlja da su unutar svake klase sve značajke međusobno nezavisne.



Slika 4.2: Usmjereni aciklički graf Naivnog Bayesovog klasifikatora

Naivan Bayesov klasifikator je najjednostavniji primjer Bayesove mreže, koja pripada skupini probabilističkih grafičkih modela (engl. *probabilistic graphical models, PGMs*), a detaljnije je opisana u potpoglavlju 4.4. Probabilistički grafički modeli korištenjem grafova kompaktno mogu modelirati potpune združene distribucije vjerojatnosti (engl. *full joint probability distribution*) svih varijabli u modelu. Na slici 4.2 je pomoću usmjerenog acikličkog grafa prikazan model Naivnog Bayesovog klasifikatora. Bridovi između čvorova u grafu predstavljaju zavisnost varijabli u čvorovima, a varijable čiji čvorovi nemaju zajedničke bridove su međusobno nezavisne. Smjer brida između čvorova određuje uzajamnu zavisnost varijabli, npr. na slici 4.2 čvor y zavisan je o svakoj varijabli x_i iz vektora značajki $(x_1, ..., x_n)$, dok za $\forall x_i, x_j, i \neq j$ vrijedi da su međusobno nezavisni unutar klase y. Uz pretpostavku da su sve značajke međusobno nezavisne, formula za izračun uvjetne vjerojatnosti pojedine značajke uz uvjet klase i ostalih značajki je

$$P(x_i \mid y, x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) = P(x_i \mid y) \tag{4.3}$$

Korištenjem izraza 4.3, formula 4.2 se pojednostavljuje na

$$P(y \mid x_1, ..., x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, ..., x_n)} \tag{4.4}$$

Za učenje parametara modela koristi se procjena najveće izglednosti (engl. *Maximum Likelihood Estimation, MLE*). Cilj postupka *MLE* je pronaći parametar y koji maksimizira izglednost, korištenjem formule

$$y_{MLE} = \arg\max_{y} \prod_{i=1}^{n} P(x_i \mid y) \tag{4.5}$$

Naivan Bayesov klasifikator odluku o pripadnosti pojedinoj klasi y donosi na temelju pravila odlučivanja koje odabire najvjerojatniju hipotezu, tzv. *Maximum A Posteriori (MAP)* hipotezu. Formula za izračun *MAP* hipoteze je

$$y_{MAP} = \arg\max_{y} \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, ..., x_n)} \tag{4.6}$$

Uzevši u obzir da je nazivnik s desne strane u izrazu 4.4 konstantan za svaki vektor značajki, možemo ga zanemariti tijekom računanja *MAP* hipoteze budući da vrijedi proporcionalnost

$$P(y \mid x_1, ..., x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y) \tag{4.7}$$

Upotrebom izraza 4.7, formulu 4.6 možemo zamijeniti konačnom formulom za izračun *MAP* hipoteze:

$$y_{MAP} = \arg\max_{y} P(y) \prod_{i=1}^{n} P(x_i \mid y) \tag{4.8}$$

Razred `MultinomialNB` iz knjižice *scikit-learn* (Pedregosa et al., 2011), koji implementira Naivni Bayesov klasifikator za multinomnu distribuciju, odabran je za implementaciju Naivnog Bayesovog klasifikatora. Vektori značajki su vektori u kojima se za svaki naslov novinskog članka pamti broj pojavljivanja pojedinih riječi iz rječnika od tisuću najbrojnijih riječi. Takva shema izdvajanja značajki poznata je pod imenom “vreća riječi" (engl. *bag of words*). Za dobivanje “vreće riječi" korišten je razred `CountVectorizer`. Svaki naslov novinskog članka razdvojen je na sastavne riječi te su potom upotrebom regularnih izraza sve brojke zamijenjene oznakom *NUM* i uklonjeni su interpunkcijski znakovi. Svaka riječ iz naslova tokenizirana je pomoću programske knjižice *nltk*. Podskup za učenje sadrži 80% nasumično odabranih primjera iz skupa podataka i sasvim je istovjetan skupu za učenje iz potpoglavlja 4.2.

4.4. Bayesova mreža

Bayesova mreža je usmjereni aciklični graf koji modelira odnose čvorova, definiran sljedećim pravilima:

- Svaki čvor predstavlja slučajnu varijablu, čije vrijednosti mogu imati diskretnu ili kontinuiranu razdiobu.
- Između čvorova postoje usmjerene veze, no usmjereni ciklusi nisu dozvoljeni. Ako je veza usmjerena od čvora A prema čvoru B, kažemo da je čvor A roditelj čvora B.
- Svaki čvor X_i sadrži pripadajuću uvjetnu distribuciju vjerojatnosti: $P(X_i \mid roditelji(X_i))$.

Za razliku od stroja potpornih vektora, Bayesova mreža je generativni model za nadzirano strojno učenje. Razlika u odnosu na diskriminativne modele je upravo da generativni modeli reprezentiraju potpunu združenu vjerojatnost svih varijabli. Ta činjenica u praksi znači da se sve varijable tretiraju jednako, tj. da se značajke mogu tretirati kao oznake i oznake kao značajke. Struktura mreže kompaktno modelira uvjetne ovisnosti između čvorova grafa. Uvjetna distribucija za varijable koje poprimaju diskretne vrijednosti, kakve su prisutne u ovom radu, reprezentirana je tablicom uvjetne vjerojatnosti (engl. *conditional probability table*). Najjednostavniji primjer Bayesove mreže je Naivan Bayesov klasifikator iz potpoglavlja 4.3., koji pretpostavlja da su sve varijable međusobno nezavisne. Za razliku od izraza 4.3, izraz za uvjetnu vjerojatnost vektora značajki, tj. konjunkcije pridodijeljenih specifičnih vrijednosti svake od n varijabli u Bayesovoj mreži je

$$P(X_1 = x_1, ..., X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i \mid roditelji(X_i)) \tag{4.9}$$

Korištenjem pravila produkta, potpunu združenu vjerojatnost za pojedini primjer možemo izračunati formulom

$$P(x_1, ..., x_n) = P(x_n \mid x_{n-1}, ..., x_1) P(x_{n-1}, ..., x_1) \tag{4.10}$$

Uzastopnim korištenjem pravila ulančavanja, formulu 4.10 možemo svesti na oblik

$$\begin{aligned} P(x_1, ..., x_n) &= P(x_n \mid x_{n-1}, ..., x_1) P(x_{n-1}, ..., x_1) ... P(x_2 \mid x_1) P(x_1) \\ &= \prod_{i=1}^{n} P(x_i \mid x_{i-1}, ..., x_1) \end{aligned} \tag{4.11}$$

Formulu 4.10 možemo dodatno pojednostaviti, koristeći izraze za uvjetnu nezavisnost više varijabli. Odaberemo li redoslijed varijabli takav da zadovoljava uvjet $roditelji(X_i) \subseteq \{X_{i-1}, ..., X_1\}$ za svaku varijablu X_i, jednadžbu 4.11 možemo poopćiti u

$$P(X_i \mid X_{i-1}, ..., X_1) = \prod_{i=1}^{n} P(X_i \mid roditelji(X_i)) \tag{4.12}$$

Učenje parametara, tj. vrijednosti tablica uvjetnih vjerojatnosti Bayesove mreže ostvaruje se korištenjem postupka procjene najveće izglednosti (engl. *Maximum Likelihood Estimation, MLE*), dok se predviđanje oznake za pojedini vektor značajki donosi upotrebom *MAP-upita* (engl. *MAP query*). Formule za *MLE* i *MAP* u Bayesovim mrežama analogne su formulama 4.5 i 4.6, uz korištenje izraza za uvjetnu zavisnost više varijabli, no njihov izvod je izvan okvira ovoga rada.

Bayesova mreža implementirana je pomoću programske knjižice *pomegranate* (Schreiber, 2016). Prvotno je korištena implementacija u programskoj knjižici *pgmpy*, međutim budući da je ta knjižica u ranoj verziji, nije optimizirana za brzo izvođenje niti je omogućena paralelizacija prilikom učenja.[2] Navedeni nedostaci ograničavaju upotrebu spomenute knjižice na Bayesove mreže s manje od sedam čvorova, i rezultirali su odlukom korištenja knjižice *pomegranate*. U trenutku pisanja ovog rada podrška za paralelizaciju je nažalost ograničena te je zbog toga implementirana reducirana Bayesova mreža. Pomoću knjižice *pomegranate*, moguće je stvoriti Bayesovu mrežu na dva načina: stvaranjem čvorova i inicijalizacijom njihove distribucije te uvjetnih vjerojatnosti i modeliranjem topologije, ili učenjem i topologije i razdioba vjerojatnosti iz samih podataka. Za potrebe ovog rada odabran je potonji pristup, jer bi u suprotnom bilo potrebno prethodno poznavanje odnosa varijabli i vrijednosti uvjetnih vjerojatnosti. Za učenje strukture korištena je pohlepna heuristička funkcija iz knjižice *pomegranate*, koja iterativno dodaje jednu po jednu varijablu u strukturu mreže. Problem pronalaženja strukture Bayesove mreže transformiran je u problem traženja najkraćeg puta u grafu koristeći pristup razvijen u (Yuan et al., 2011). Cilj funkcije je pronaći Bayesovu mrežu s najmanjom vrijednošću statističkog kriterija *MDL/BIC (Minimum description length/Bayesian Information Criterion)*, koji je definiran formulom

$$BIC = -2 \cdot \ln(L) + k \cdot \ln(n) \tag{4.13}$$

U formuli 4.13 varijabla L je najveća vrijednost funkcije izglednosti modela na uzorku, varijabla n veličina promatranog uzorka (broj primjera u uzorku), a k broj slobodnih parametara za procjenu (Wit et al., 2012).

[2] `http://pgmpy.org/`

Prednost korištenja heuristike u pronalasku strukture mreže je ravnoteža između često optimalne strukture grafa i relativno malo potrebnih računalnih resursa. Bitno je napomenuti da takav pristup ne garantira globalno optimalnu strukturu. Skup primjera za učenje je podskup 80% ukupnih istinitih oznaka i emocija i vrijednosti vijesti, koje odgovaraju primjerima za učenje iz potpoglavlja 4.2. i 4.3.

5. Eksperimenti i rezultati

Testni skup podataka na kojem je provedeno vrednovanje svih verzija svih modela čini 20% nasumično odabranih primjera iz ukupnog skupa podataka. Za vrednovanje modela na testnom skupu korištena je F_1-mjera (engl. F_1 *score*), pojedinačno za svaku oznaku, čime je u biti ostvareno 15 binarnih klasifikacija (za šest oznaka emocija i 11 oznaka vrijednosti vijesti). U kontekstu binarne klasifikacije s klasama *pozitivno* i *negativno*, F_1-mjeru moguće je definirati pomoću matrice zabune (engl. *confusion matrix*):

$$M = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix}$$

Elementi matrice zabune imaju sljedeća značenja:

- TP je broj ispravno klasificiranih pozitivnih primjera (engl. *true positive*),
- FP je broj pogrešno klasificiranih pozitivnih primjera (engl. *false positive*),
- FN je broj pogrešno klasificiranih negativnih primjera (engl. *false negative*),
- TN je broj ispravno klasificiranih negativnih primjera (engl. *true negative*).

F_1-mjera je harmonička sredina mjera *preciznost* (engl. *precision*, P) i *odziv* (engl. *recall*, R) definirana formulom

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R}$$

$Preciznost$ je definirana kao udio točno klasificiranih primjera u skupu pozitivno klasificiranih primjera:

$$P = \frac{TP}{TP + FP}$$

$Odziv$ je definiran kao udio točno klasificiranih primjera u skupu svih pozitivnih primjera

$$R = \frac{TP}{TP + FN}$$

5.1. Postavke eksperimenta

Eksperimenti su provođeni zasebno za predviđanje emocija i vrijednosti vijesti. Provedeno ih je ukupno deset: pet eksperimenata za predviđanje emocija i pet eksperimenata za predviđanje vijestodostojnosti vijesti. Cijeli skup podataka slučajnim odabirom podijeljen je na skup za učenje (treniranje), kojeg čini 80% primjera i pripadajućih istinitih oznaka emocija i vrijednosti vijesti, i skup za testiranje, kojeg čini 20% preostalih primjera i pripadajućih istinitih oznaka. U svakom eksperimentu korišten je drugačiji model, a u dva od pet eksperimenata za svaki skup oznaka kombinirani su diskriminativni i generativni modeli. Dva modela stroja potpornih vektora i Naivan Bayesov klasifikator uče na temelju vektora značajki i istinitih oznaka iz skupa za učenje, dok predviđanja donose samo na temelju vektora značajki iz skupa za testiranje. Tri modela Bayesovih mreža uče na temelju svih istinitih oznaka iz skupa za učenje, a predviđanja rade zaključivanjem na temelju opaženih varijabli (engl. *observed variables*) iz skupa za testiranje: ako predviđaju emocije za određeni skup naslova novinskih članaka, za predviđanje su im pružene oznake vrijednosti vijesti tih naslova, i obrnuto. Korišteni modeli su:

- NB: Naivan Bayesov klasifikator čiji su vektori značajki “vreća riječi" naslova novinskih članaka,
- SVM1: stroj potpornih vektora čiji su vektori značajki *w2v* reprezentacije naslova novinskih članaka,
- SVM2: stroj potpornih vektora čiji su vektori značajki *w2v* reprezentacije naslova novinskih članaka, i istinite oznake emocija u slučaju predviđanja vrijednosti vijesti ili istinite oznake vrijednosti vijesti u slučaju predviđanja emocija,
- gold-BN: Bayesova mreža koja predviđanja radi na temelju istinitih oznaka vrijednosti vijesti ili emocija,
- SVM1+BN: Bayesova mreža koja predviđanja radi na temelju predviđanja (emocija ili vrijednosti vijesti) SVM1 modela,
- SVM2+BN: Bayesova mreža koja predviđanja radi na temelju predviđanja (emocija ili vrijednosti vijesti) SVM2 modela.

5.2. Rezultati predviđanja emocija

Tablica 5.1: F_1-mjere predviđanja emocija

emocija	NB	SVM1	SVM2	SVM1+BN	SVM2+BN	gold-BN
ljutnja	0.00	0.00	0.00	0.00	0.00	0.00
gađenje	0.00	0.00	0.00	0.00	0.00	0.00
strah	0.18	**0.42**	0.33	0.00	0.00	0.00
radost	0.28	0.41	**0.48**	0.18	0.29	0.45
tuga	0.35	0.48	**0.67**	0.29	0.48	0.69
iznenađenje	0.00	0.00	0.00	0.00	0.00	0.00
makro-prosjek	0.14	0.22	**0.25**	0.08	0.13	0.19

Tablica 5.1 prikazuje F_1 mjere predviđanja emocija u šest različitih slučajeva. Najbolje rezultate u predviđanju emocija postiže model SVM2 čiji su vektori značajki *w2v* reprezentacije naslova novinskih članaka i istinite oznake vrijednosti vijesti. Drugi najbolji rezultat postigao je modeli SVM1, u kojem je korišten SVM čiji su vektori značajki bile samo *w2v* reprezentacije. Rezultati naivnog Bayesovog klasifikatora (model NB) su osjetno lošiji i usporedivi s Bayesovim mrežama. U ostala tri slučaja također su ostvareni lošiji rezultati u odnosu na SVM, a u svima je predviđanja stvarala određena Bayesova mreža na temelju sljedećih opaženih varijabli:

- SVM1+BN: opažene varijable su predviđanja vrijednosti vijesti modela SVM1,
- SVM2+BN: opažene varijable su predviđanja vrijednosti vijesti modela SVM2,
- gold-BN: opažene varijable su istinite oznake vrijednosti vijesti.

Bitno je spomenuti da se makro rezultati po emociji modela SVM1 poboljšaju samo 13,6% dodavanjem istinitih oznaka vrijednosti vijesti u skup ulaznih značajki, tj. korištenjem modela SVM2. Na prvi pogled, zbog lošijeg prosječnog rezultata Bayesovih mreža i prethodno navedene činjenice proizlazi zaključak da emocije vrijednosti vijesti slabo utječu na predviđanje emocija. Međutim, detaljnijom analizom rezultata modela SVM1 i SVM2 vidljivo je da uključivanjem istinitih oznaka vrijednosti vijesti u verziju SVM2 rezultat značajno raste za vrijednosti vijesti *radost* i *tuga*, dok se suprotno dogodi za *strah*, i nema promjene za ostale tri emocije. Zanimljive varijacije u rezultatima za pojedinu emociju, za razliku od prosječnih rezultata, podupiru hipotezu da postoji određena razina interakcije između emocija i vrijednosti vijesti.

5.3. Rezultati predviđanja vrijednosti vijesti

Tablica 5.2: F_1-mjere predviđanja vrijednosti vijesti

vrijednost vijesti	NB	SVM1	SVM2	SVM1+BN	SVM2+BN	gold-BN
loša vijest	0.38	0.48	**0.73**	0.00	0.00	0.00
celebrity	0.29	**0.60**	0.53	0.00	0.00	0.00
sukob	0.16	0.34	**0.36**	0.00	0.00	0.00
drama	0.55	**0.67**	**0.67**	0.00	0.00	0.00
zabava	0.60	0.75	**0.77**	0.53	0.51	0.55
popratna vijest	0.25	0.27	**0.29**	0.00	0.00	0.00
dobra vijest	0.34	0.46	**0.47**	0.00	0.00	0.00
važnost	**0.60**	0.18	0.20	0.00	0.00	0.00
shareability	0.27	0.30	**0.38**	0.00	0.00	0.00
iznenađenje	0.00	**0.20**	0.00	0.00	0.00	0.00
moć i elita	0.42	0.55	**0.58**	0.00	0.00	0.00
makro-prosjek	0.35	0.44	**0.45**	0.05	0.05	0.05

F_1 mjere predviđanja vrijednosti vijesti za šest različitih slučajeva prikazane su u tablici 5.2. U predviđanju vrijednosti vijesti naslova novinskih članaka najbolje prosječne rezultate po pojedinoj kategoriji postižu dva model SVM-a. Vektori značajki modela SVM1 su *w2v* reprezentacije naslova, dok su za model SVM2 osim njih dodane istinite oznake emocija. Iako su prosječni makro rezultati u obje verzije isti, postoje razlike u rezultatima za pojedine vrijednosti vijesti. Primjerice, rezultat za oznake *loša vijest* i *shareability* znatno se poveća kad se u skup vektora značajki uključe istinite oznake emocija, dok se suprotno dogodi za vrijednosti vijesti *iznenađenje* i *celebrity*. Naivan Bayesov klasifikator, čiji su rezultati prikazani u stupcu NB, postiže tek treći najbolji prosječni rezultat, no iznenađujuće je što ostvaruje najbolji rezultat u predviđanju *važnosti*. U preostala tri preostala slučaja predviđanja je stvarala određena Bayesova mreža na temelju sljedećih opaženih varijabli:

- SVM1+BN: opažene varijable su predviđanja emocija modela SVM1,
- SVM2+BN: opažene varijable su predviđanja emocija modela SVM2,
- gold-BN: opažene varijable su istinite oznake emocija.

Bayesova mreža je u sva tri slučaja ostvarila iznimno loše rezultate za sve vrijednosti vijesti, osim za vrijednost vijesti *zabava*, što je razumljivo s obzirom na to da sadrži samo 15 varijabli. Rezultati dobiveni predviđanjem SVM-a i u ovom slučaju indiciraju

da postoji interakcija između vrijednosti vijesti i emocija. Preliminarni zaključak na temelju rezultata u tablicama 5.1 i 5.2 jest da vrijednosti vijesti više ovise o emocijama nego što to vrijedi obrnuto. Uočljivo je da su najbolji rezultati postignuti za vrijednosti vijesti koje imaju najbolju *kappa*-statistiku u tablici 3.1. Hipoteza o nezanemarivoj interakciji između emocija i vrijednosti vijesti i u ovom je skupu eksperimenata djelomično poduprta dobivenim rezultatima.

6. Zaključak

Cilj ovog rada bio je predviđanje vijestodostojnosti i kategorija emocija novinskih članaka na engleskom jeziku pomoću strojnog učenja na temelju naslova novinskih članaka. Za potrebe rada ručno je označen prikupljeni skup podataka s kategorijama vrijednosti vijesti, uz prethodno dostupne oznake emocija. Korištena su tri modela, od kojih je Naivni Bayesov klasifikator služio kao referentni model. Model SVM je rezultatima nadmašio referentni model u predviđanju emocija i gotovo svih vrijednosti vijesti. Bayesova mreža je ostvarena u reduciranom obliku te nije ostvarila očekivane rezultate.

Uzevši u obzir povoljne dobivene rezultate, cilj rada je ostvaren, no ostaje dovoljno prostora za nadogradnju. Pretpostavka je da bi Bayesova mreža s većim brojem varijabli mogla bolje modelirati odnose između riječi, vrijednosti vijesti i emocija te nadmašiti najbolji trenutno razvijeni model. Zbog neadekvatne programske podrške takav pristup problemu u trenutku izrade rada nije bio ostvariv. Trenutno razvijen sustav potencijalnu primjenu ima u domeni društvenih znanosti, u sažimanju dokumenata, praćenju događaja i sl. Mogući pravci za budući rad su razvoj i implementacija Bayesove mreže koja bi modelirala odnose riječi, vrijednosti vijesti i emocija, rješavanje problema korištenjem metoda dubokog učenja i korištenje sažetka članka umjesto naslova za predviđanje vijestodostojnosti.

LITERATURA

Steven Bird, Ewan Klein, i Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

Andrew Boyd. *Broadcast journalism: techniques of radio and television news*. Taylor & Francis, 2001.

Tony Harcup i Deirdre O'neill. What is news? galtung and ruge revisited. *Journalism studies*, 2(2):261–280, 2001.

Tony Harcup i Deirdre O'Neill. What is news? news values revisited (again). *Journalism Studies*, stranice 1–19, 2016.

Tomas Mikolov, Kai Chen, Greg Corrado, i Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013. URL `http://arxiv.org/abs/1301.3781`.

Benjamin Nye i Ani Nenkova. Identification and characterization of newsworthy verbs in world news. U *HLT-NAACL*, stranice 1440–1445, 2015.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, i E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Radim Řehůřek i Petr Sojka. Software Framework for Topic Modelling with Large Corpora. U *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, stranice 45–50, Valletta, Malta, Svibanj 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

Stuart Jonathan Russell i Peter Norvig. Artificial intelligence: a modern approach (3rd edition), 2009.

Jacob Schreiber. pomegranate. `https://github.com/jmschrei/pomegranate`, 2016.

Carlo Strapparava i Rada Mihalcea. Semeval-2007 task 14: Affective text. U *Proceedings of the 4th International Workshop on Semantic Evaluations*, stranice 70–74. Association for Computational Linguistics, 2007.

Shyam Upadhyay, Christos Christodoulopoulos, i Dan Roth. "making the news": Identifying noteworthy events in news articles.

Ernst Wit, Edwin van den Heuvel, i Jan-Willem Romeijn. 'all models are wrong...': an introduction to model uncertainty. *Statistica Neerlandica*, 66(3):217–236, 2012.

Changhe Yuan, Brandon Malone, i Xiaojian Wu. Learning optimal bayesian networks using a* search. U *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

Predviđanje vijestodostojnosti novinskih članaka pomoću strojnog učenja

Sažetak

Računalno predviđanje vijestodostojnosti nov je smjer istraživanja s primjenama u praćenju događaja te sažimanju dokumenata, a potencijalnu primjenu ima u društvenim znanostima. Tema rada jest predviđanje vijestodostojnosti i emocija novinskih članaka na engleskome jeziku temeljeno na strojnom učenju. Razvijena su dva generativna i jedan diskriminativni model za klasifikaciju vijestodostojnosti i emocija na temelju naslova članaka u kategorije vijestodostojnosti koje su predložili Harcup i O'Neill (2011) i šest osnovnih emocija. Na prikupljenoj zbirci naslova novinskih članaka ručno označenih kategorijama vijestodostojnosti je provedeno učenje i vrednovanje modela te usporedbe s referentnim modelom.

Ključne riječi: obrada prirodnog jezika, strojno učenje, vijestodostojnost, emocije, stroj potpornih vektora, Bayesova mreža

Predicting Newsworthiness of News Stories Using Machine Learning

Abstract

The automatic prediction of newsworthiness is a new research direction with applications in event tracking and document summarization, and has potential applications in social sciences. The topic of this thesis is the prediction of newsworthiness of news stories and emotions in English using machine learning. Two generative and one discriminative model were developed for classifying newsworthiness and emotions into newsworthiness categories proposed by Harcup and O'Neill (2011) and six basic emotions, based on the article headline. An experimental evaluation of the models and comparisons against a baseline model were carried out on a compiled collection of articles, manually labeled for newsworthiness.

Keywords: natural language processing, machine learning, newsworthiness, emotions, support vector machine, Bayesian network