Andmeteadus ja digihumanitaaria (HVEE.00.016)

Klõpsulõkslike pealkirjade tuvastamine masinõppega

Lisanna Lehes, Karel Paan, Gregor Eesmaa

Sissejuhatus

Pealkiri on üks uudisloo olulisemaid komponente, sest see on esimene asi, mida inimesed näevad. Hästi kirjutatud pealkiri peaks olema informatiivne ja köitma tähelepanu, andes lugejatele selge ettekujutuse, millest lugu räägib. Kuna paljudel juhtudel sõltub väljaande käekäik just klikkide arvust, siis jätavad uudislugude autorid erinevatel viisidel pealkirjadesse lünki, mida lugeja saab enda jaoks täita just klõpsulõkslikul pealkirjal klõpsates. Uurisime klõpsulõkslike pealkirjade (ingl *clickbait titles*) omadusi ja tegime kindlaks, kas neid saab masinõpet kasutades tuvastada. Kogusime ja analüüsisime pealkirju klõpsulõksudel ja mitte-klõpsulõksudel, arvestades erinevaid arvutatud tunnuseid (ingl *feature*), mis võivad nende klassifitseerimist mõjutada.

Meie töö klõpsulõkslike pealkirjade mõistmisel ja täpsel klassifitseerimisel on oluline, sest klõpsulõks võib avaldada negatiivset mõju nii kasutajatele kui ka veebisaitidele. Eksitav või madala kvaliteediga klõpsulõks võib kasutajaid tüüdata, kahjustada selle avaldava veebisaidi usaldusväärsust ja põhjustada ärilist või poliitilist kaost. Teisest küljest aitab klõpsulõksu täpne tuvastamine kasutajatel vältida aja raiskamist väheväärtuslikule sisule ning võimaldab veebisaitidel parandada oma pealkirjade ja sisu kvaliteeti. Meie leiud aitavad teadvustada ja luua klõpsulõksude avastamise ja käsitlemise strateegiaid, mis on kasulikud nii lugejatele kui ka veebisaitidele.

Valdkond

Kaalusime erinevaid viise klõpsulõksude määramiseks. Et erinevas kontekstis määratakse klõpsulõkse erinevalt, otsustasime ise kasutada kahest küsimusest koosnevat definitsiooni. Kui

vastus mõlemale küsimusele on jaatav, ei loeta artiklit klõpsulõksuks, kuid kui vähemalt üks neist on ei, siis on tegu klõpsulõksuga.

1. Kas artikkel on pärit uudisteallikast, mis on tuntud kvaliteetse sisu loojana?
2. Kas artikli sisu on põhjalik ja informatiivne?

Kuna meie töö ei keskendu artiklite sisule, vaid pealkirjale, siis aitab selline määratlus meil leida ja valideerida oma lähtematerjali ja treeningandmeid, seejuures olles sõltumatu pealkirjast endast.

Üritasime ka leida töid, mis juba on püüdnud määratleda ja klassifitseerida klõpsulõkse. Näiteks on varasemalt klõpsulõkse määratletud tehisnärvivõrkude abil [1], kus lahendus keskendus POSAM (ingl *Part of Speech Analysis Module*) võimaluste rakendamisele - teksti-/kõneosade arvutuslikule tuvastamisele. Samuti on kasutatud tavalisemaid masinõppe algoritme, nagu logistiline regressioon ja Naiivne Bayesi klassifitseerija (ingl *Naïve Bayes classifier*) [2] või SVM (ingl *support vector machine*), otsustuspuu (ingl *decision tree*) [3]. Kuigi eeltöö käigus ei suutnud me tuvastada, et masinõppega oleks sarnast asja ka varasemalt eestikeelsete pealkirjade peal tehtud, on klõpsulõksude esinemisest pikemalt kirjutatud Kairi Jansoni magistritöös [4].

Andmed ja meetodid

Projekti eesmärgiks oli leida tunnused, mis eristavad klõpsulõksude pealkirju mitte-klõpsulõksude pealkirjadest. Selleks kogusime andmestiku nii klõpsulõksude- kui ka mitte-klõpsulõksudega. Klõpsulõksud kogusime "Klikisäästja" Facebooki lehelt (facebook.com/klikimasin) (vt joonis 1). Andmed kraapisime kokku, kasutades JS funktsionaalsust brauseri konsoolist. Mitte-klõpsulõksud kogusime Eesti Rahvusringhäälingust (ERR, err.ee), millel on API viimaste uudiste pärimiseks. Kood andmete kogumiseks ja analüüsimiseks ning puhas andmestik asuvad GitHubis [5].

Loodud tunnuste aluseks oli eelnevalt tutvumine varasemalt tehtud sarnaste uurimustega kui ka “Klikisäästja” Facebooki lehe põhjal nähtud pealkirjade kvalitatiivne hindamine.

Tabel 1. Pealkirjade teksti põhjal genereeritud andmestik

Koodis kasutatud väärtus	Seletus
sentence	Pealkirja tekst.
article_url	Artikli aadress.
has_adjective	Näitab, kas pealkiri sisaldab omadussõnu (väärtus 1) või mitte (väärtus 0).
no_adjectives	Pealkirjas esinevate omadussõnade arv.
has_verb	Näitab, kas pealkiri sisaldab tegusõnu (väärtus 1) või mitte (väärtus 0).
no_verbs	Pealkirjas esinevate tegusõnade arv.
has_name	Näitab, kas pealkiri sisaldab nimesi (väärtus 1) või mitte (väärtus 0).
no_names	Pealkirjas esinevate nimede arv.
no_words_starting_with_capslock	Pealkirjas esinevate sõnade arv, mis algavad suure tähega.
no_words_with_only_capslock	Pealkirjas esinevate sõnade arv, mis sisaldavad ainult suuri tähti.
has_exclamation_mark	Näitab, kas pealkiri sisaldab hüüumärki (väärtus 1) või mitte (väärtus 0).
has_question_mark	Näitab, kas pealkiri sisaldab küsimärki (väärtus 1) või mitte (väärtus 0).
has_numbers	Näitab, kas pealkiri sisaldab numbreid (väärtus 1) või mitte (väärtus 0).
starts_with_number **(NB! Lõpplahenduses ei kasutatud, kuna ei leidunud ühtegi esinemisjuhtu!)**	Näitab, kas pealkiri algab numbriga (väärtus 1) või mitte (väärtus 0).
no_words	Pealkirjas esinevate sõnade arv.
no_characters	Pealkirjas esinevate tähemärkide arv.
contains_gtp_reccomended_word	Indikeerib, kas pealkiri sisaldab Chat-GPT poolt välja toodud tüüpilisi klõpsulõksu sõnu või mitte.
label	Näitab, kas tegu on klõpsulõksuga (väärtus 1) või mitte (väärtus 0).

Joonis 1. Näide Klikisäästja defineeritud klõpsulõksust ja ERR-st võetud pealkirjast.

Pealkirjade põhjal uute tunnuste loomiseks kasutasime TartuNLP tööriista [7]. Eelistasime seda, kuna see pakub spetsiifilisemaid mudeleid erinevate ülesannete jaoks. Eestikeelsed stoppsõnad [6] võtsime välja vaid wordcloudi, teiste visuaalide jaoks, sest mudeldamiseks sõnu ei kaasanud.

Klõpsulõksu tuvastamise ülesannet saab masinõppes defineerida binaarse klassifikatsiooni probleemina. Iga pealkirja saame klassifitseerida ühte kategooriasse kahest – tegemist on kas klõpsulõksuga või mitte-klõpsulõksuga. Eesmärgiks seame välja selgitada, mis tegurid kallutavad ühe või teise valiku kasuks. Märgendatud pealkirju ja loodud tunnuseid kasutasime logistilise regressiooni treenimiseks.

Lisades 1 ja 2 on sõnapilvega vastavalt välja toodud klõpsulõksude ning mitte-klõpsulõksude artiklites enamlevinud sõnad (kust eemaldatud on stoppsõnad).

Murekohad

Esialgu kartsime, et ei suuda 600 erinevast allikast pärit klõpsulõksust koosnevat andmekogu ära tasakaalustada mitte-klõpsulõkslike pealkirjadega ega valida treenimiseks õigeid tunnuseid. Esimese mure lahenduseks kogusime andmeid tuntud uudisteallika ERR kõigist rubriikidest, et suurendada enda mitte-klõpsulikslike pealkirjade hulka ja vähendada riski õpetada välja mudel, mis oskab eristada vaid nn “kollaseid” ja “valgeid” artikleid. Jõudsime ka järeldusele, et teised uudisteallikad, olles keskendunud eelkõige tulu teenimisele, võivadki kasutada suurema tõenäosusega klõpsulõkslikke pealkirju enda kasumi suurendamiseks. Seetõttu ongi kõige puhtamaks eestikeelseks andmeallikaks loodetavasti just maksumaksjate poolt rahastatud ERR.

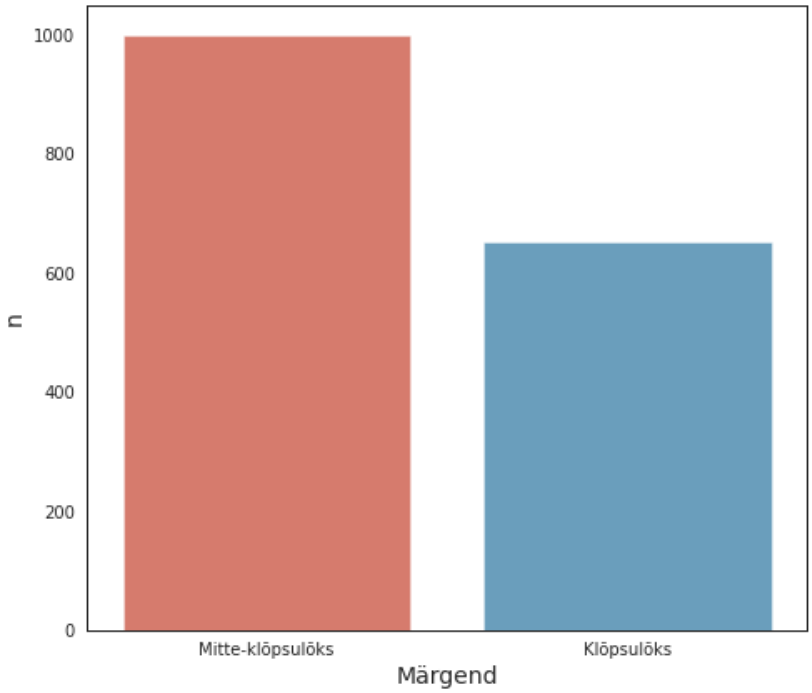
Teise mure lahenduseks kaalusime mitmeid tunnuseid, nagu suurtähtede, hüüumärkide ja sõnade arvu pealkirjas, jättes samas välja sõnad ise, et vältida mudelile hoopis nn “kollaste” ja “valgete” teemade eristamise õpetamist. See sunniks meie mudelit mõistma üldisemaid omadusi, mis määravad klõpsulõksliku pealkirja, mitte seostama just kindlaid teemasid või sõnu klõpsulõksudega.

Tulemused

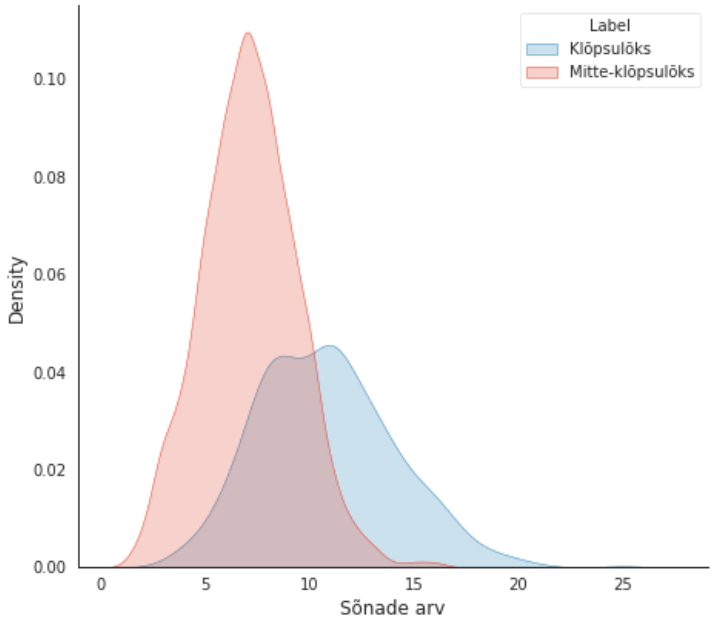
Andmestiku tutvustus

Valimis oli meil kokku 652 klõpsulõksu artiklit ja 1000 mitte-klõpsulõksu artiklit (vt joonis 2).

Vaadates sõnade arvu jaotust märgendite lõikes (joonis 3), on näha, et klõpsulõksude ja mitte-klõpsulõksude jaotus erineb. Klõpsulõksude puhul on keskmiseks pealkirja pikkuseks 10,8 sõna, 95% CI[10,56; 11,05] ja mitte-klõpsulõksude puhul 7,2 sõna, 95% CI[7,02; 7,29]. Seega saab erinevuse klõpsulõksude ja mitte-klõpsulõksude keskmiste vahel lugeda statistiliselt oluliseks ($t=26,98$, $p<0,001$).

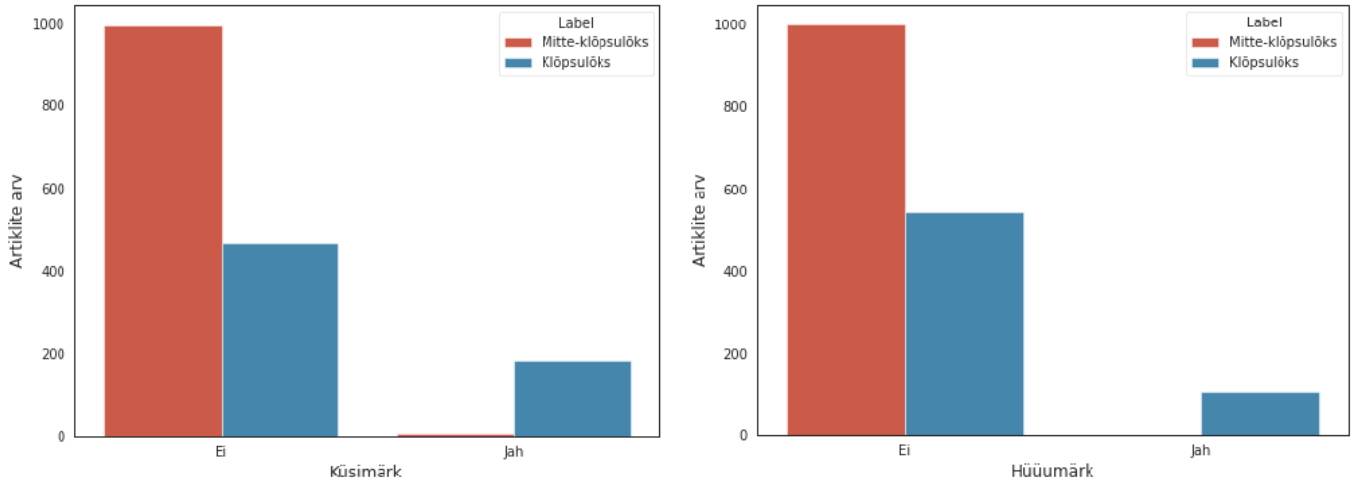


Joonis 2. Klasside osakaal valimis



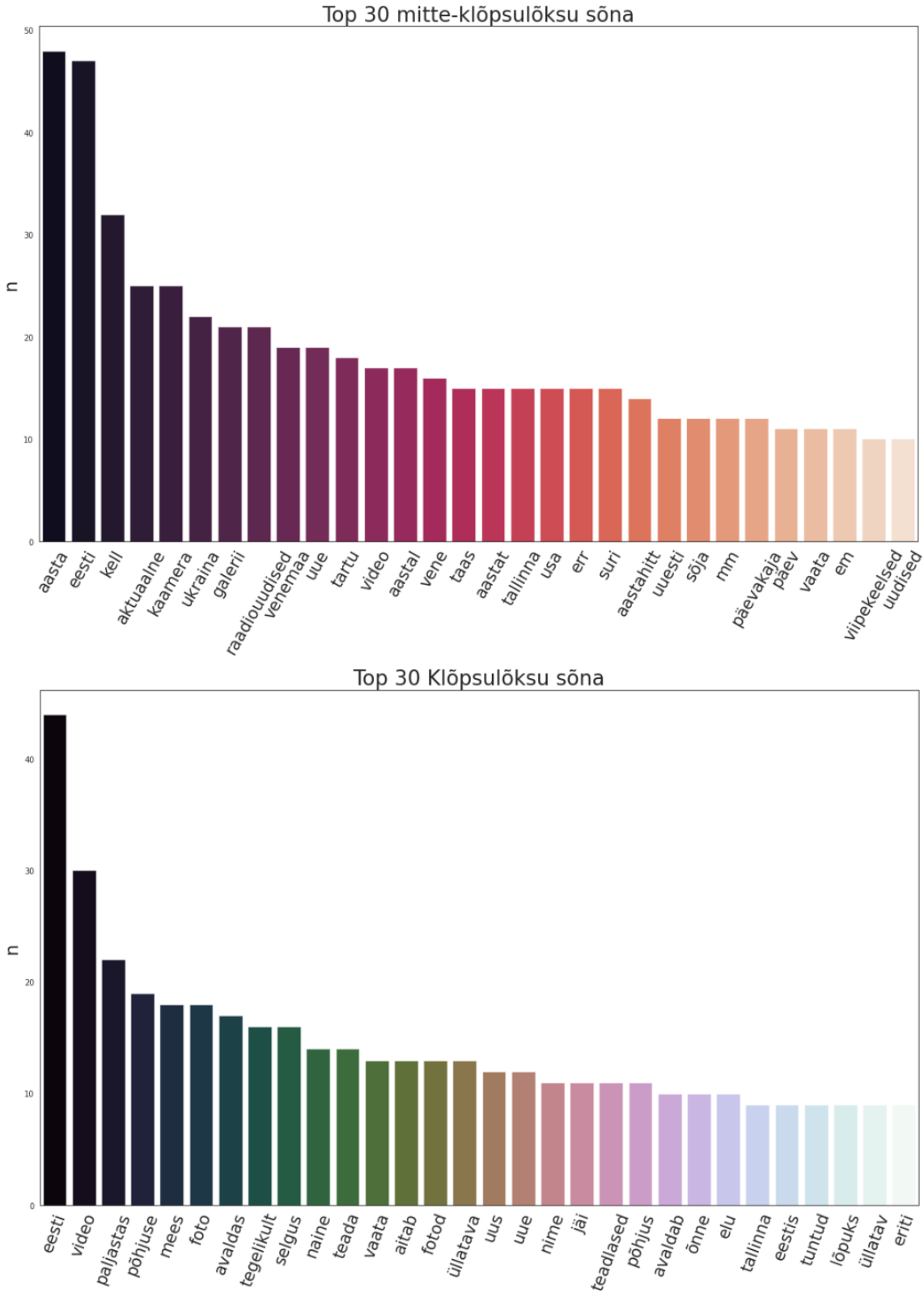
Joonis 3. Pealkirjades olevate sõnade arvude jaotus

Mitte-klõpsulõksudel üldiselt ei ole pealkirjades hüüd- või küsilauset, kuid umbes pooled klõpsulõksu pealkirjad sisaldavad, kas küsi- või hüüdlauset (joonis 4).



Joonis 4. Küsi- ja hüüumärkide esinemine märgendite lõikes

Joonisel 5 on näha kõige rohkem nii klõpsulõksu kui ka mitte-klõpsulõksu pealkirjades esinevad sõnad. Antud juhul on eemaldatud ka eestikeelsed stoppsõnad. Kahe kategooria vahel võib märgata mitmeid erinevusi. Klõpsulõksu pealkirjad sisaldavad mingil määral omadussõnu, nagu *paljastas, ülatava, üllatav*. Lisaks ka sõnu, nagu *video* ja *foto/fotod*, mille peale tõesti klõpsulõksu artiklid tihti oma sisu on üles ehitanud. Mitte-klõpsulõksu sõnade valikus esineb palju ERR-i sisule iseloomulikke sõnu, nagu *aktuaalne* ja *kaamera, päevakaja, viipekeelsed* ja *uudised.* Samuti on nende puhul näha, et sisu on päevakajaline - juttu on Ukrainast ja Venemaast.



Joonis 5. Top 30 sõna mitte-klõpsulõksu ja klõpsulõksu pealkirjade lõikes

Mudeldamine

Loodud andmestiku treenisime ja testisime kasutades logistilist regressiooni. Treenimiseks jäi 70% ja testimiseks 30% andmestikust. Mudel klõpsulõksu ja mitte-klõpsulõksu pealkirjadest peegeldab, milliseks osutuvad klõpsulõksuks kategoriseerimise šansid sõltuvalt loodud tunnustest, nagu sõnade arv, küsi-/hüüumärgi olemasolu, tegusõnade arv jne (vt tabel 1).

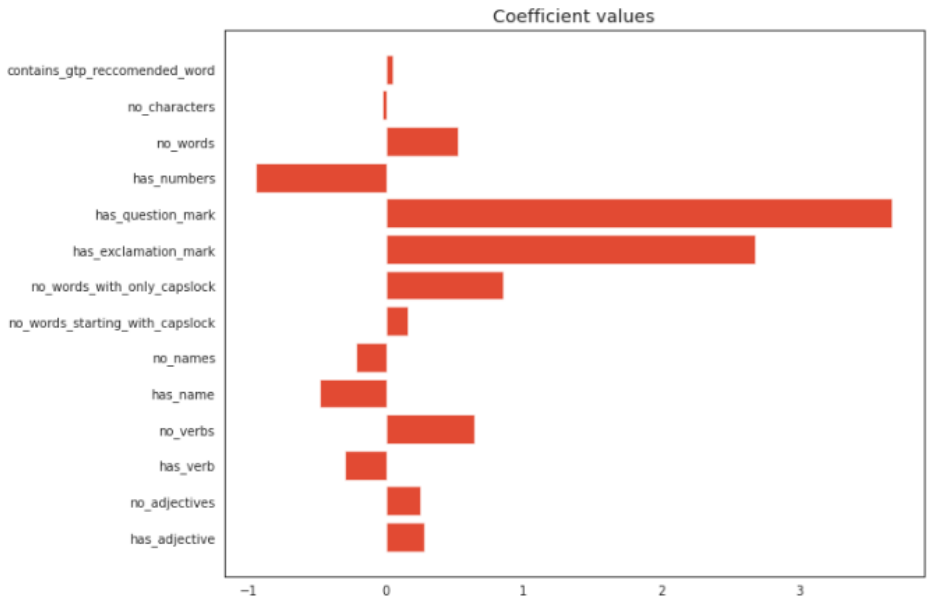
```
intercept  -4.46169285990515
classes [0 1]
----------------------
F1 score of the model
0.8215962441314555
----------------------
Accuracy of the model
0.8608058608058609
----------------------
Confusion Matrix
[[295  23]
 [ 53 175]]
----------------------
Classification Report
              precision    recall  f1-score   support

           0       0.85      0.93      0.89       318
           1       0.88      0.77      0.82       228

    accuracy                           0.86       546
   macro avg       0.87      0.85      0.85       546
weighted avg       0.86      0.86      0.86       546

----------------------
auc/roc
0.8476084078119829
```

	coeff	exp(coeff)	p
has_question_mark	3.6750	39.4475	0.0000
has_exclamation_mark	2.6762	14.5304	0.0000
no_words_with_only_capslock	0.8488	2.3369	0.0000
no_verbs	0.6439	1.9038	0.0000
no_words	0.5251	1.6906	0.0000
has_adjective	0.2717	1.3122	0.0000
no_adjectives	0.2490	1.2827	0.0000
no_words_starting_with_capslock	0.1591	1.1724	0.0000
contains_gtp_reccomended_word	0.0490	1.0503	0.0041
no_characters	-0.0293	0.9711	0.0000
no_names	-0.2135	0.8078	0.0171
has_verb	-0.3016	0.7396	0.0000
has_name	-0.4836	0.6166	0.0006
has_numbers	-0.9453	0.3886	0.0000



Joonis 6. Mudeli tulemused

Joonisel 6 on näha peamised olulised mõõdikud. Mudeli täpsus oli 86,1%. F1-skoor (0,82) on mõõdik mudelite kvaliteedi hindamiseks, kombineerides täpsuse (ingl *precision*) ja saagise (ingl *recall*). Kuna andmestikus ei ole otseselt probleemi taskaalustamata andmetega, siis need kaks mõõdikut ka väga sarnaste väärtustega on. Küll aga on näha, et väärtuse 1 (klõpsulõks) puhul on saagis (ingl *recall*) madalam, mis viitab sellele, et mudel võib ennustada mõned valenegatiivsed tulemused (klõpsulõks määratakse mitte-klõpsulõksuks).

Joonisel 6 on toodud ka koefitsientide väärtused (coeff, exp(coeff)) ja nende olulisuse tõenäosused (p). Küsi- ja hüüumärgi olemasolu ning omadussõnade olemasolu viitavad klõpsulõkslikule pealkirjale. Sõnade arvu, ainult suurte tähtedega sõnade arvu (*caps lock*), tegusõnade arvu ning omadussõnade arvu suurenemine põhjustab tõenäolisemalt klõpsulõksu pealkirja. Näiteks kõige suurema mõjuga klõpsulõksu pealkirjale on küsimärgi olemasolu ehk pealkirjas küsimärgi olemasolu korral on klõpsulõksu pealkirja tõenäosus küsimärgi mitte esinemisega võrreldes 39 korda kõrgem.

Numbrite olemasolu, tähemärkide arv, nimede esinemine ja arv ning tegusõnade olemasolu viitavad aga rohkem mitte-klõpsulõksu pealkirjale (vt joonis 6). Näiteks numbri olemasolul on 2,6 (1/0,3886) korda tõenäolisem, et tegu on mitte-klõpsukõksuga.

Valenegatiivseid pealkirju ehk klõpsulõksud, mis klassifitseeriti mitte-klõpsulõksudeks, oli 53. Näiteks:

a. *Tassitäis seda armastatud jooki aitab põletada rasva*
b. *Teivashüppe maailmameister ootab surma: mul ei ole enam paranemisvõimalust*
c. *Suri rockiajaloo ühe suurima bändi laulja*

Valepositiivseid pealkirju ehk mitte-klõpsulõksud, mis klassifitseeriti klõpsulõksudeks, oli 23. Näiteks:

a. *Sõja 302. päev: Venemaale võib olla küüditatud sadu tuhandeid Ukraina lapsi*
b. *Spekulatsioon: EKRE võimalik ministrikandidaat on Tea Varrak*
c. *Kannapööre: tuntud Eesti muusik liitus kinnisvaramaaklerite ridadega*

Arutelu

Projekti põhjal saame öelda, et klõpsulõksude tuvastamist on võimalik automatiseerida. Välja tulid ka mitmed olulised näitajad klõpsulõksude tuvastamisel, nagu küsi- ja hüüumärkide olemasolu, sõnade arv, ainult suurte tähtedega sõnade arv, omadus- ning tegusõnade arv jne. Tulevikus võiks olla mitmeid suundi, kuhu siit liikuda. Üks võimalus on lisada täiendavaid funktsioone, nagu ikkagi ka täpsete sõnade kaasamine tunnustesse, mis eeldaks suuremat ja mitmekesisemat hulka treeningandmeid. Sõnade kaasamisel mudeli treenimisse oleks võimalik esmalt määrata nende sentiment - selgitada, kas tegu on positiivse või negatiivse sõnaga. Lisaks on välja mõelda veel mitmeid näitajaid, kuidas klõpsulõksu määrata. Näiteks võiks kaaluda ka asesõnade olemasolu. Masinõppe algoritmide valik on lai - edasistes töödes võiks kaaluda mitme olemuselt kergete aga ka keerulisemate *black-box* mudelite võrdlemist kui ka tehisnärvivõrke. Samuti tuleks edasistes mudelites prioritiseerida saagist (ingl *recall*), et minimeerida valenegatiivseid tulemusi.

Teine suund võiks olla kaaluda ka artikli konteksti: veebilehte/plavormi või ka rubriiki; et paremini mõista, kuidas klõpsulõkse erinevates kontekstides kasutatakse. Eesti keele tekstist rubriigi tuvastamiseks võib kasutada näiteks mudelit “tartuNLP/mtee-domain-detection”[1]. Samuti oleks huvitav uurida klõpsulõkslike pealkirjade mõju kasutaja käitumisele, näiteks seda, kui tihti neil klõpsatakse ja kui palju neisse süvenetakse. Analüüsida võib ka autorite kaupa - millised autorid kirjutavad enim klõpsulõkse, millist väljaannet nad esindavad, ning kui sagedasti nad kirjutavad klõpsulõksu pealkirjaga artikli. Veel saaks analüüsida, kuidas erinevad sama autori poolt kirjutatud tavalised artiklid (mitte-klõpsulõksud) ning klõpsulõksu artiklid - selgitada, kas autor teadlikult muudab oma kirjutamise stiili klõpsulõksu kirjutades.

[1] https://huggingface.co/tartuNLP/mtee-domain-detection

Viited

[1] Naeem, B., Khan, A., Beg, M.O. et al. A deep learning framework for clickbait detection on social area network using natural language cues. *J Comput Soc Sc* 3, 231–243 (2020). https://doi.org/10.1007/s42001-020-00063-y

[2] J. Huette, M. Al-Khassaweneh and J. Oakley, "Using Machine Learning Techniques for Clickbait Classification," *2022 IEEE International Conference on Electro Information Technology (eIT),* Mankato, MN, USA, 2022, pp. 091-095, doi: 10.1109/eIT53891.2022.9813776

[3] Pujahari, A., & Sisodia, D. S. (2021). Clickbait detection using multiple categorisation techniques. *Journal of Information Science, 47*(1), 118–128. https://doi.org/10.1177/0165551519871822

[4] Janson, K. (2019). *Vaata ja imesta: lünkpealkirjad veebiuudises portaali Elu24 näitel.* Magistritöö. Tartu Ülikool.

[5] Projektiga seotud koodi ja andmete repositoorium GitHub-is https://github.com/gregoreesmaa/digihum-clickbait.

[6] Uiboaed, K. Eesti keele stoppsõnad / Estonian stop words. http://datadoi.ee/handle/33/78

[7] TartuNLP. https://github.com/tartunlp

Lisad

Lisa 1: Sõnapilv klõpsulõksu artiklitest (eemaldatud on stoppsõnad).

Lisa 2: Sõnapilv mitte klõpsulõksu artiklitest (eemaldatu on stoppsõnad).