Klasifikacija sporočil spletnega pogovora o knjigah

Adam Prestor, Lojze Žust
ap2408@student.uni-lj.si, lojze.zust@student.uni-lj.si

1 Uvod

V slovenskem prostoru pismenost in želja po branju med mlajšimi generacijami počasi upadata, zato so se pojavile razne iniciative, ki želijo ta trend obrniti in ponovno oživeti zanimanje za branje in ohranjanje slovenskega jezika. Med njimi je tudi obširen projekt v katerem sodeluje več ustanov, med drugim tudi nekaj fakultet iz Univerze v Ljubljani. Cilj projekta je združiti sodobno tehnologijo in knjižno gradivo, ter tako mladim generacijam približati branje. Pomemben del projekta predstavlja orodje IMapBook. Gre za spletni prikazovalnik digitalnih knjig in ostalih gradiv, ki pa ima nekaj močnih dodatnih funkcionalnosti. Omogoča namreč izvedbo spletnih diskusij o posameznem gradivu, kjer bralci lahko svoja vprašanja in mnenja delijo z ostalimi.

V tej seminarski nalogi se osredotočamo na podatke, ki so jih v okviru projekta pridobili z raziskavo na več Ljubljanskih osnovnih šolah. Učenci so v IMapBooku prebrali pripravljeno gradivo, nato pa so se udeležili diskusije, ki je za izhodišče postavila odprto vprašanje na podlagi prebranega gradiva. Učenci so se lahko pogovarjali o temi, pri tem pa so jih usmerjali moderatorji (v tem primeru učitelji/ce). Opazimo lahko, da brez posredovanja moderatorjev, pogovori zelo hitro zaidejo izven teme knjige ali postanejo žaljivi. Zato se je pojavila želja, da bi sistem pomagal moderatorjem in zaznal, v kateri točki pogovora je potrebno posredovanje. V tem delu želimo nasloviti to potrebo in razviti metodo, ki bo v pomoč moderatorjem. Osredotočamo se na podproblem in sicer klasifikacijo posameznih sporočil v različne kategorije. Podproblem je bolj fokusiran in jasno zastavljen, ter lažje merljiv in omogoča enostavnejše primerjanje razvitih metod. Poleg tega taka klasifikacija omogoča osnovno informiranje moderatorja o morebitnem potrebnem posredovanju. Če na primer nekaj časa zelo majhen del sporočil predstavlja sporočila o dejanski temi, potem se lahko izda obvestilo moderatorju, da preveri vsebino.

Problem ni enostaven zaradi specifik spletnih klepetalnic. Sporočila so zelo kratka in vsebujejo ogromno količino slovničnih napak in slangovskih besed, kar otežuje tokenizacijo in ostale dele analize besedil. Poleg tega se sporočila zelo pogosto sklicujejo druga na drugo, uporabljajo se uporabniška imena in imena naučena tekom pogovora. Kontekst sporočila je tako zelo širok in ga je težko jasno določiti. Zbrani podatki temelijo na diskusijah o zgolj treh različnih delih, kar lahko povzroči slabo generalizacijo na ostala dela. Metoda se lahko na primer nauči, da so pogovori, kjer se omenja cefizlja vredu, vendar to glede na kontekst ni nujno res.

2 Sorodna dela

Obstaja več različnih pristopov za klasifikacijo besedil. Osnoven model za klasifikacijo je naivni Bayes, ki predpostavlja neodvisnost in nepomembnost vrstnega reda besed in deluje na podlagi pogojnih verjetnosti, da se beseda w pojavi v besedilu tipa c. Za tak pristop ne pričakujemo velikega uspeha, saj je povprečno število besed v besedilih zelo majhno, ter z njim ne zajamemo konteksta pogovora. Bi pa tak model mogoče znal detektirati sporočila, ki niso relevantna za temo (npr. iskanje identitete, kletvice).

Boljši model bi moral biti sposoben upoštevati ostala sporočila v pogovoru. Če obravnavamo pogovor kot sekvenco sporočil, lahko uporabimo metode za označevanje sekvenc, ki poskušajo za klasifikacijo posameznega elementa upoštevati tudi sosednje elemente. Taka metoda so skriti Markovi modeli (Rabiner, 1989), ki pa jih v našem primeru ni smiselno uporabiti, saj je množica opazovanih stanj (vsa možna sporočila) neskončna. Uporabimo pa lahko pogojne Markove modele (MEMM) (McCallum et al., 2000) in metode CRF (Lafferty et al., 2001), ki namesto

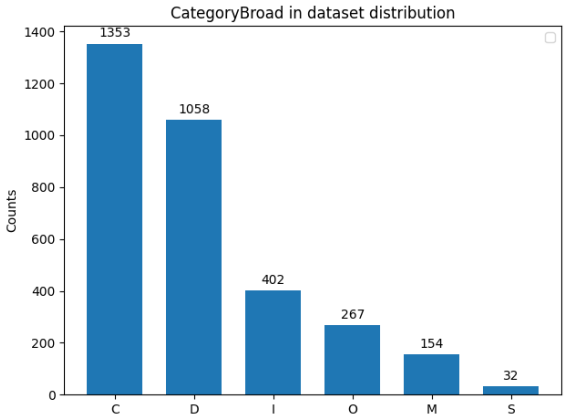
množice opazovanih stanj uvedejo prostor značilk, na podlagi katerih se izračuna prehodna verjetnost. Vsak element sekvence moramo tako predstaviti s točko v tem prostoru značilk. Značilke lahko izluščimo ročno (prisotnost določenih besed, ločil, dolžina, itd.), ali pa z uporabo vložitev. Kot bolj napredne metode je vredno omeniti tudi globoke metode – rekurenčne nevronske mreže (LSTM (Hochreiter and Schmidhuber, 1997)) in transformerje (Vaswani et al., 2017). Trenutno dajejo najboljše rezultate na širokem spektru problemov in so sposobne modeliranja kompleksnih in daljših odvisnosti znotraj sekvenc.

Za značilke lahko vložitve pridobimo na več načinov. Lahko uporabimo redek model bag-of-words, kar pa za metode označevanja sekvenc ni preveč uporabno, saj je s tem načinom število značilk preveliko. Bolj uporabne so goste vložitve, ki bistveno zmanjšajo število dimenzij. Tu pridejo v poštev metode kot so PCA (Jolliffe, 1986), word2vec (Mikolov et al., 2013) in globoke vložitve. V primeru uporabe globokih vložitev, je zanimiva tudi možnost izgradnje end-to-end arhitekture, ki združuje globoke vložitve in klasifikacijo sekvenc v enovit model, kjer lahko obe komponenti optimiziramo hkrati. Tak model zna biti vprašljiv zaradi relativno majhne količine podatkov.

Nekaj del (Weisz; Dong et al., 2006) se ukvarja specifično s problemom klasifikacije teme pogovora v kontekstu spletnih klepetalnic in opisujejo podobne metodologije.

3 Metodologija

Zaradi nezanesljivosti črkovanja napisanih sporočil, je potrebna robustna tokenizacija sporočil. En pristop bi bil, da bi po tokenizaciji obdržali le besede, ki jih najdemo v slovarju. Tako bi se fokusirali na besede, za katere vemo, da nosijo nek pomen, vendar bi morda zavrgli preveč informacije. Drug pristop bi bil, da po tokenizaciji z uporabo Lehvensteinove razdalje (ali kake druge boljše razdalje) poiščemo najbližjo besedo v slovarju. V kolikor je ta manjša od določenega praga, potem sklepamo, da je šlo za napako pri črkovanju in izberemo besedo iz slovarja. Besede, ki so še vedno neznane označimo s posebnim tokenom <neznano>. Poleg tega bi lahko poseben token uvedli tudi za vse pojavitve uporabniških imen (<username>) in tako omogočili lažjo detekcijo pogovorov o identiteti. Pomemben del predprocesiranja bo najbrž predstavljalo tudi odstravnjevanje stop besed, ki predstavljajo velik del sporočil.



Slika 1: Porazdelitev števila sporočil v podatkovni zbirki glede na tip sporočila. (C: Chatting, S: Switching, D: Discusion, M: Moderating, O: Other, I: Identity)

4 Rezultati

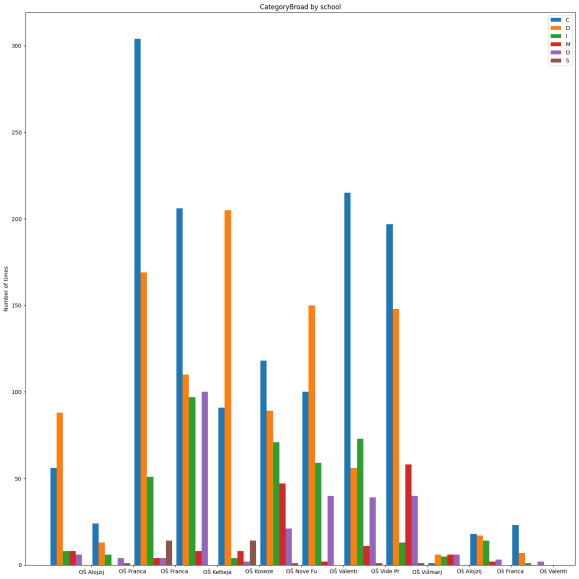
4.1 Analiza podatkovne zbirke

Pred začetkom implementacije smo izvedli preliminarno analizo podatkovne zbirke, da bi pridobili čim več informacij, ki bi nam pomagale pri izbiri ustrezne metodologije.

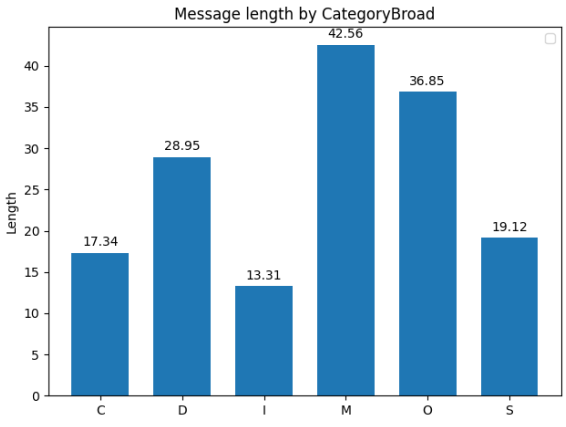
Najprej smo preverili porazdelitev števila sporočil med posameznimi tipi sporočil (glej Sliko 1). Največji del sporočil predstavljata kategoriji Chatting, ki predstavlja neformalen pogovor in Discussion, ki predstavlja pogovor o knjigi. Izmed ostalih tipov je največ Identity, ki vsebuje pogovore o identiteti posameznikov.

Na Sliki 2 je prikaz števila sporočil posameznih tipov po različnih šolah v zbirki. Vidimo da se število sporočil precej razlikuje med šolami. Prav tako je v nekaterih šolah več neformalnega pogovora, drugje pa so se učenci bolje držali teme.

Na Sliki 3 smo analizirali še povprečno dolžino sporočila po posameznih kategorijah. Vidimo da so sporočila moderatorja v povprečju najdaljša. Prav tako so sporočila diskusije povprečno rahlo daljša kot klepetanje. Nekatere razlike so precejšnje, tako da bi dolžino sporočila mogoče lahko uporabili kot uporabno značilko pri napovedih.



Slika 2: Porazdelitev tipa sporočil po različnih šolah. (C: Chatting, S: Switching, D: Discusion, M: Moderating, O: Other, I: Identity)



Slika 3: Povprečna dolžina sporočila (v znakih) glede na tip sporočila. (C: Chatting, S: Switching, D: Discusion, M: Moderating, O: Other, I: Identity)

5 Zaključek

Viri

Haichao Dong, Siu Cheung Hui, and Yulan He. 2006. Structural analysis of chat messages for topic detection. *Online Information Review*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ian T Jolliffe. 1986. Principal components in regression analysis. In *Principal component analysis*, pages 129–155. Springer.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Andrew McCallum, Dayne Freitag, and Fernando CN Pereira. 2000. Maximum entropy markov models for information extraction and segmentation. In *Icml*, volume 17, pages 591–598.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Lawrence R Rabiner. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Justin Weisz. Segmentation and classification of online chats.