SLENGLARNI AVTOMATIK TEGLASHDA LUG‘ATGA VA MASHINALI O‘QITISHGA ASOSLANGAN YONDASHUVLAR

Bekmuradova Iroda Zokir qizi,
Kompyuter lingvistikasi mutaxassisligi magistranti
irodabekmuradova9@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada slenglarni avtomatik teglashning lug‘atga va mashinali o‘qitishga asoslangan usullari xususida so‘z yuritiladi. Bunda lug‘atga asoslangan usulning bosqichlari ketma-ketlikda ko‘rsatib, mazkur usulning afzalliklari va kamchiliklari keltiriladi. Shuningdek, maqolada slenglarni avtomatik teglash uchun Bayes va SVM (support vector machine) modellari ham taklif etilib, bu modellarga xos dastlabki umumiy tavsiflar ham berilib, ilmiy tahlil qilinadi.

Abstract. This article discusses dictionary-based and machine learning-based methods for automatic tagging of slang. The steps of the dictionary-based method are shown sequentially, and the advantages and disadvantages of this method are presented. The article also proposes Bayesian and SVM (support vector machine) models for automatic tagging of slang, provides initial general descriptions of these models, and conducts scientific analysis.

Аннотация. В этой статье рассматриваются методы автоматической разметки сленга на основе словаря и машинного обучения. Здесь будут последовательно показаны этапы метода, основанного на словаре, а также представлены преимущества и недостатки этого метода. В статье также предлагаются байесовские модели и модели SVM (машина опорных векторов) для автоматической разметки сленга, даются начальные общие описания этих моделей и проводится научный анализ.

Kalit so‘zlar: *sleng, teg, teglash, lug‘at, Bayes algoritmi, Support vector machine.*

Ma’lumki, til korpusida sleng birliklarni ajratib tahlil qilish, ularga avtomatik teglash orqali bunday birliklar qatnashgan matnning u yoki bu tahlili amalga oshirilganda slenglarni oddiy leksik birliklar kabi tahlil qilib ketish, avtomatik tarjima jarayonida so‘zni to‘g‘ridan-to‘g‘ri belgilangan tilga tarjima qilib qo‘yish kabi xatoliklarni to‘xtatishga erishish mumkin.

Til korpuslari o‘z maqsadi va ichki tuzilishiga ko‘ra turlicha bo‘lishiga qaramay, ularni umumlashtiruvchi asosiy xususiyat shundan iboratki, ular til birliklarini avtomatik teglash hamda ularni ma’lum bir mezonlarga ko‘ra tahlil qilish

yoki shunga o‘xshash lingvistik tahlillarni amalga oshirish imkonini beradi. Korpusda tahlil uchun kiritilgan matnlardagi birliklarni teglashga oid turli yondashuvlar va bir qator matematik usullar mavjud. Shulardan asosiylari: qoidaga asoslangan yondashuv, lug‘atga asoslangan yondashuv, mashinali o‘qitishga asoslangan yondashuv, statistik va transformatsion yondashuv va hokazolar [2].

Har qanday til korpusi tilning leksik va grammatik xususiyatlarini chuqur o‘rganish imkonini beradi, til tizimini hamda uning o‘zgaruvchanligini kuzatishga xizmat qiladi. Milliy til korpusi esa, tilning barcha xususiyatlarini aks ettiruvchi og‘zaki va yozma matnlarni o‘z ichiga olgani bois, tilning dinamikasini, so‘zlar va grammatik shakllarning kelib chiqishi hamda qo‘llanishini tahlil qilishda, matnlarning lingvistik xususiyatlarini o‘rganishda muhim ahamiyat kasb etadi.

Matndagi so‘z va iboralarni teglash tabiiy tilni qayta ishlash jarayonida muhim vazifalardan biri hisoblanadi. Teglash orqali so‘zlarning tuzilishi, ma’nosi va ularning kontekstdagi o‘rni haqida batafsil ma’lumot beriladi. Bu jarayonda matn tarkibidagi so‘z yoki iboralarga maxsus izohlar qo‘shiladi. Izohlar tadqiqot maqsadiga qarab, grammatik yoki leksik ma’no, shuningdek, boshqa muhim xususiyatlarni ko‘rsatishi mumkin.

Teglar (inglizchasiga “tags”) – bu matn haqidagi qo‘shimcha ma’lumotni o‘zida mujassam etuvchi belgilardir. Matnli ma’lumotlarni (korpuslarni) teglash uchun bir necha universitetlar matnlarning qaysi parametrlarini teglash kerakligini tavsiflovchi tizimni maxsus ishlab chiqdi. Ushbu tizim XMLdan foydalanadi va Text Encoding Initiative Guidelines (TEI Guidelines) deb nomlanadi. Bu kodlash, teglash va indeksirlash mumkin bo‘lgan matnlarning turli xil xususiyatlarining ro‘yxati hisoblanadi. Masalan, tizim matndagi turli xil tuzatishlar, iqtiboslar, qisqrtmalar, atoqli otlar, initsial, akronimlar, chet el so‘zlari va boshqalarni sanab o‘tadi.[1]

So‘zlarni teglashning lug‘atga asoslangan yondashuvi quyidagi bosqichlarni o‘z ichiga oladi:

1-rasm. Slenglarni avtomatik teglashning lug‘atga asoslangan usuli bosqichlari

Lug‘atga asoslangan teglash usuli oldindan tuzilgan lug‘atlardan foydalangan holda matn birliklariga mos teg yoki toifani belgilashga asoslanadi. Masalan, sleng birliklarini avtomatik teglashda dastlab tildagi sleng birliklari aniqlanib, maxsus lug‘at shakllantiriladi. So‘ngra matn kiritilganida, u lug‘atdagi birliklarga mosligi bo‘yicha tekshiriladi va mos keladigan birliklarga teg biriktiriladi. Ushbu yondashuv nisbatan soddaligi bilan ajralib turadi hamda muayyan matn turlari uchun, xususan, aniq kalit so‘zlar mavjud bo‘lsa, yuqori aniqlikdagi natijalarni ta’minlashi mumkin.

Shu bilan birga, lug‘atga asoslangan yondashuvning ayrim kamchiliklari mavjud. Xususan, teglash natijalarining sifat darajasi lug‘atning to‘liqligi va aniqligiga bevosita bog‘liq. Agar lug‘at barcha zaruriy birliklarni o‘z ichiga olmasa, natijaviy tahlil sifati pasayadi. Shuningdek, kontekstga bog‘liq yoki noaniq birliklarni teglashda muayyan cheklovlarga duch kelinadi. Ushbu usulning samaradorligini oshirish uchun kontekstual lug‘atlardan foydalanish, mashinani o‘qitish modellarini joriy etish yoki bir nechta lug‘atlarni birlashtirish orqali natijalarni yaxshilash mumkin.

Tabiiy tilni qayta ishlash (NLP) sohasida slenglarning avtomatik aniqlanishi va teglanishi muhim vazifalardan biri hisoblanadi. Slenglar norasmiy nutq qatlamiga tegishli bo‘lib, doimiy ravishda o‘zgarib boradi va odatiy lug‘aviy tizimlarga sig‘maydi. Shuning uchun, slenglarni samarali tarzda tasniflash va ularni to‘g‘ri teglash uchun kuchli va moslashuvchan algoritmlar talab etiladi. Slenglarni teglashda ML (mashinali o‘qitish)ga asoslangan yondashuv bu jarayonga kompyuter omilini kiritish, ya’ni mos algoritmlarni qo‘llagan holda, matn tarkibida kelgan slengni o‘ziga xos ehtimoliylik darajasiga nisbatan teglashni o‘z ichiga oladi. Bunday toifalar ierarxik tuzilishga ega bo‘lib, ML ga asoslangan usullar kompyuter lingvistikasida faol qo‘llaniladi. Tabiiy tildagi matnlarni ikki bosqichli tahlil qilishning birinchi bosqichida, lug‘atda mavjud bo‘lgan so‘z tahlil qilinayotgan matnning har bir so‘zi bilan taqqoslanadi. Agar so‘z lug‘atda mavjud bo‘lmasa u holda keying bosqichga o‘tiladi. Ikkinchi bosqichda, jarayonda shakllangan qoidalarga muvofiq mavjud bo‘lgan teglash variantlaridan yagona to‘g‘ri namunasi tanlanadi. Ushbu yondashuvning afzalligi aniqlikning ustuvor bo‘lishidir. Ammo bu jarayonda qoidalar sonining ko‘payishi aniqlikning kamayishiga sabab bo‘lishi mumkin. Avtomatik teglashning mashinali o‘qitish usullari slenglar va shunga o‘xshash noaniq so‘zlarni izohlash muammosini hal qilishda samaralidir. Bugungi kunda jahon kompyuter lingvistikasida ushbu sohada faol izlanishlar olib borilmoqda. Matnni tasniflash, undagi birliklarni teglashga doir bir qator ishlarni kuzatib, shunga amin bo‘ldikki, vaznni multiplikativ tuzatish algoritmi, yashirin semantik tahlil, transformatsiyani o‘rgatishga asoslangan algoritm, differentsial grammatika, qatorlar ro‘yxati usuli, Bayes algoritmi, perseptron usullar kabilar

noaniqlikni bartaraf etish muammosini hal eta oladi [2]. Shulardan Bayes algoritmi va uning matndagi birliklarni teglashga doir xususiyatlarini ko‘rib chiqamiz.

Bayes algoritmi obyektning ehtimolini, uning xususiyatlarini va tegishli bo‘lgan guruhini aniqlashni o‘rganuvchi algoritmdir. XVIII asr yashagan matematik olim Tomas Bayes nomi bilan ataladigan ushbu algoritm bir qator sohalarda keng tarqalgan. Mazkur algoritmning markazida Bayes xulosasi yotadi, bu usul bizga yangi dalillar yoki ma’lumotlar paydo bo‘lganda gipoteza haqidagi farazni yangilash imkonini beradi [7]. U ehtimollik tasniflagichi sifatida ham tanilgan. Masalan, so‘zni muayyan xususiyatlari, unga qo‘shilgan qo‘shimchalar va bog‘lanib kelgan so‘zlarga qarab aniq tahlil qilish qiyin, chunki xuddi shu jihatiga ko‘ra o‘xshash xususiyatlarga ega ko‘plab so‘zlar mavjud. Ammo, bu narsa haqida ehtimolli bashorat qilish mumkin va bu yerda Naive Bayes algoritmi qo‘l keladi. Bayes algoritmi, ayniqsa Naive Bayes usuli tabiiy tilni qayta ishlashda muhim rol o‘ynaydi. NLP da u orqali matnni tasniflash, sentiment tahlil, spamni aniqlash, hujjatlarni turkumlashtirish, matn tilini aniqlash, nomlangan obyektni tanish (NER), matnlarni klasterlash va modellashtirish kabi masalalarni hal qilish mumkin. Bayes nazariyasi berilgan dalillar to‘plamidan (B) muayyan gipotezaga (A) kelish ustida ishlaydi.

Slenglarni avtomatik teglashda qo‘llash mumkin bo‘lgan yana bir model Support Vector Machine (SVM) – nazorat ostidagi o‘rganish (supervised learning) asosida ishlovchi klassifikatsiya modeli bo‘lib, u matnlarni tasniflashda keng qo‘llanilishi bilan mashhur [3]. SVM modelining asosiy maqsadi – ma’lumotlar to‘plamini maksimal gipertekislik (optimal ajratuvchi chiziq) orqali ajratishdir.

SVM modeli kichik va murakkab ma’lumot to‘plamlarida yaxshi natija ko‘rsatgani, chiziqli va chiziqli bo‘lmagan tasniflash vazifalariga moslashuvchanligi [4], so‘zlarning semantik va sintaktik o‘ziga xosliklarini hisobga olgan holda slenglarni teg guruhlariga ajrata olishi kabi xususiyatlari bilan slenglarni teglashda samarali hisoblanadi:

Slenglarni teglash uchun SVM modeli quyidagi bosqichlarda ishlatiladi:

1. Ma’lumotlarni to‘plash va tayyorlash

– Sleng so‘zlar bazasini yaratish (ijtimoiy tarmoqlar, forumlar, chat loglaridan).

– Slenglar teglarini aniqlash va ularni toifalarga ajratish (masalan, yoshlar slengi, texnologik sleng, professional jargon va h.k.).

2. Xususiyatlarni (Features) aniqlash

– TF-IDF (Term Frequency-Inverse Document Frequency) yoki so‘z embedding usullari orqali har bir so‘z uchun vektorli taqdimot yaratish [6].

– Morfologik va semantik belgilarni ajratib olish.

3. SVM modelini o‘qitish va tekshirish

– Olingan xususiyatlar asosida SVM modeli o‘qitiladi.

– Model samaradorligini baholash uchun ma’lumotlar to‘plami trening va test to‘plamlariga bo‘linadi [5].

– F1-score, Precision, Recall kabi metrikalar asosida natijalarni tahlil qilinadi.

4. Natijalarni tahlil qilish va optimallashtirish

– Model natijalarini yaxshilash uchun Hyperparameter tuning (C, kernel turi, gamma) amalga oshiriladi [6].

Agar zarur bo‘lsa, boshqa NLP usullari bilan birlashtirish (masalan, LSTM yoki Word2Vec bilan kombinatsiya qilish) ham mumkin.

Xulosa qilib aytganda, til korpusida sleng birliklarini aniqlash va avtomatik teglash jarayonini to‘g‘ri yo‘lga qo‘yish – zamonaviy til texnologiyalarining eng dolzarb vazifalaridan biridir. Bu nafaqat avtomatik tarjima jarayonidagi xatoliklarni kamaytiradi, balki slenglarni oddiy leksik birliklardan farqlash orqali matnni yanada aniqroq tahlil qilish imkonini beradi. Ayniqsa, lug‘atga asoslangan yondashuv, o‘zining soddaligi va yuqori aniqligi bilan ajralib turadi. Bu yondashuv, to‘liq va dolzarb lug‘at bazasi shakllantirilgan taqdirda, tilshunoslik, kompyuter lingvistikasi hamda matnlarni avtomatik qayta ishlash sohalarida katta imkoniyatlarni ochadi.

Shu bilan birga, texnologiya rivojlanishi bilan bir qatorda, slenglarning o‘zi ham doimiy o‘zgarishda ekani, ularni muntazam yangilab borishni taqozo qiladi. Shuningdek, faqat lug‘atga tayanish yetarli emasligini anglash kerak – kombinatsiyalashgan yondashuvlardan foydalanish, ya’ni statistik, qoidaga asoslangan va mashinali o‘qitishga asoslangan usullarni ham birlashtirish orqali ancha mukammal tizimni yaratish mumkin. Bu esa, kelajakda nafaqat sleng birliklarini, balki boshqa murakkab til hodisalarini ham yanada chuqurroq o‘rganishga zamin yaratadi. Shunday ekan, slenglarni avtomatik teglash bo‘yicha olib borilayotgan izlanishlar nafaqat lingvistik nazariya uchun, balki amaliy sohalar – masalan, sun’iy intellekt yordamida matn tahlili, ma’lumotlarni avtomatik indekslash va hatto ijtimoiy tarmoqlarda til monitoringi uchun ham muhim ahamiyat kasb etadi. Mashinali o‘qitish metodlari yordamida slenglar va ularning kontekstlarini avtomatik tarzda tahlil qilish imkoniyatlarini ko‘rib chiqish mumkin. Bunda mashinali o‘qitish algoritmlari yoki neyron tarmoqlar, deep learning metodlari yordamida sleng so‘zlarini avtomatik tarzda tanib olish tizimini ishlab chiqish muhim ahamiyat kasb etadi. Shuningdek, slenglarning til tizimida qanday o‘zgarishlarni keltirib chiqarganini tahlil qilishda mashinali o‘qitish modellari katta yordam beradi. Bu boradagi tadqiqotlar esa slenglarning tilshunoslikdagi o‘rni va ijtimoiy kontekstga bog‘liq ravishda qanday rivojlanishini yanada chuqurroq o‘rganishga yo‘l ochadi.

Foydalanilgan adabiyotlar:

1. Abjalova M. (2023). Korpus birliklarini teglash va annotatsiyalash masalasi. “Zamonaviy ilm-fan va ta’lim istiqbollari” respublika konferensiyasi materiallari, (pp. 345-351). Toshkent.
2. Демская-Кульчицкая О.М., Семеренко В.Р., Ющенко Р.А. 2001: 72
3. Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. European Conference on Machine Learning, 137-142.
4. Cristianini, N., & Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press.
5. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. JMLR.
6. Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification.
7. An introduction to Naive Bayes Algorithm for beginners. (n.d.). Retrieved from Turing: http://www.turing.com
8. http://www.turing.com