

AVTOMATLASHTIRILGAN TIL TAHLILI: SLENGNI ANIQLASH VA TEGLASH MUAMMOLARI

Botir Elov Boltayevich,
Texnika fanlari falsafa doktori, dotsent
elov@navoiy-uni.uz
ToshDO‘TAU

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

Annotatsiya. Ushbu maqolada zamonaviy matnlarda keng tarqalgan sleng birliklarini avtomatik aniqlash va teglash muammolari yoritilgan. Tilshunoslik va sun'iy intellekt kesishmasida olib borilayotgan tadqiqotlar tahlil qilinib, amaliy misollar keltiriladi. Tadqiqot natijalari til texnologiyalari rivojida muhim ahamiyat kasb etishi mumkinligi ta'kidlanadi. Muammoning dolzarbligi, mavjud yondashuvlar, duch kelinayotgan qiyinchiliklar hamda kelajakdagi imkoniyatlar ko'rib chiqiladi. Maqola matnlarda slenglarni avtomatik tahlil qilish uchun zamonaviy metodlar va sun'iy intellekt tizimlarining imkoniyatlarini o'rganishga qaratilgan.

Abstract. This article discusses the problems of automatic detection and labeling of common slang units in modern texts. Research conducted at the intersection of linguistics and artificial intelligence is analyzed and practical examples are given. It is emphasized that the results of the research can be of great importance in the development of language technologies. The relevance of the problem, existing approaches, challenges encountered, and future opportunities are considered. The article is aimed at studying the capabilities of modern methods and artificial intelligence systems for automatic analysis of slang in texts.

Аннотация. В статье рассматриваются проблемы автоматического обнаружения и разметки общеупотребительных сленговых единиц в современных текстах. Будут проанализированы исследования на стыке лингвистики и искусственного интеллекта, а также приведены практические примеры. Подчеркивается, что результаты исследования могут иметь существенное значение в развитии языковых технологий. Обсуждаются актуальность проблемы, существующие подходы, стоящие перед ней проблемы и будущие возможности. Целью статьи является исследование возможностей современных методов и систем искусственного интеллекта для автоматического анализа сленга в текстах.

Kalit so'zlar: *sleng, avtomatlashtirish, teglash, sun'iy intellekt, til tahlili, NLP.*

Global raqamli makonda foydalanuvchilar sonining ortishi matnli ma'lumotlarning mislsiz darajada ko'payishiga olib keldi. Xususan, ijtimoiy tarmoqlar, bloglar va messenjerlar orqali har kuni milliardlab matnlar yaratilmoqda. Ushbu matnlarning muhim xususiyatlaridan biri – norasmiylik va slenglarning keng qo'llanilishi. Sleng tili tez o'zgaruvchan, kontekstga bog'liq va ko'pincha rasmiy lug'atlardan chetda qoladi. Shu bois, slengni avtomatik tarzda aniqlash va teglash masalasi sun'iy intellekt tizimlari uchun dolzarb muammolardan biri hisoblanadi.

Sleng – bu ma'lum ijtimoiy guruh yoki jamiyat qatlamida qo'llaniladigan, norasmiy, obrazli va emotsional boyitilgan so'z va iboralar tizimi. Sleng odatda kundalik muloqotda, yoshlar tilida, professional jargonda yoki internet muhitida keng qo'llaniladi[1]. Sleng birliklarining asosiy xususiyatlari: normativ bo'lmaganlik, tez yangilanish, kontekstga bog'liqlik, emotsional yuklama va hokazo.

Sleng birliklari ko'pincha qisqartirishlar, metaforalar, leksik yangi hosilalar, fonetik o'zgartirishlar orqali yaratiladi. Bu esa ularni aniqlashni murakkablashtiradi. Matnlarda slengni aniqlash bir nechta nazariy yondashuvlar orqali amalga oshiriladi. Ulardan eng birinchisi, lug'atga asoslangan yondashuv bo'lib, bu usulda oldindan tuzilgan sleng lug'atlaridan foydalaniladi. Lug'atdagi birliklar matnda qidirilib, aniqlansa, teg qo'yiladi. Kamchiligi – lug'at doimiy yangilanib turishi kerak va yangi paydo bo'lgan slenglarni qamrab olmaydi. Ikkinchi nazariy yondashuv qoidaga asoslangan yondashuvdir. Bu yondashuv sleng birliklarini fonetik yoki morfologik xususiyatlariga ko'ra aniqlashga qaratiladi. Misol uchun, qisqartmalar yoki ba'zi o'zgacha yozuv shakllari orqali. Shu bilan birga, slenglarni aniqlashda statistik modellardan foydalaniladi. Masalan, n-gram modellari yoki TF-IDF usuli orqali matndagi odatiy bo'lmagan birliklar aniqlanadi [1]. Bundan tashqari, mashinaviy o'rganish algoritmlari – logistika regressiyasi, Naive Bayes, SVM yoki chuqur o'rganish (Deep Learning) modellari qo'llaniladi.

So'nggi yillarda slengni aniqlashda transformer arxitekturasida asosidagi modellar samaradorlikni oshirdi. Bu modellar kontekstni chuqur tushunishga asoslanganligi sababli, yangi va kutilmagan sleng shakllarini ham muvaffaqiyatli aniqlash imkonini beradi.

Matnda slengni aniqlagach, uni teg (ya'ni belgi) bilan belgilash jarayoni muhim bosqich hisoblanadi. Teglash orqali matndagi sleng birliklari yanada aniqlik bilan tasniflanadi. Bu esa keyingi til tahlilida (masalan, sentiment tahlili yoki matnni normallashtirishda) qo'l keladi.

Slengni teglashda quyidagi muammolar kuzatiladi:

1. Ko'pma'nolilik (polisemiya). Slenglarning ba'zilar tildagi boshqa leksik hodisalar: metafora, o'xshatish, sinekdoxa kabi hodisalarga o'xshab ketadi. Ammo

shunday bo'lsa ham mazkur hodisalar bir-biridan farqli holatda qarashni talab etadigan ayrim xususiyatlarga ega. So'zlarning ma'no ko'chish hodisalari, umuman ko'chma ma'noda qo'llangan birliklar har bir so'zlovchi tomonidan, nutq vaziyatidan kelib chiqqan holda individual hosil qilinishi mumkin. Ammo slenglarda bunday xususiyat yo'q, ular an'anaviy foydalanishda bo'ladi.

2. Kontekstga bog'liqlik. Kontekstual semantika muhim hisoblanadi. Slengni aniqlash uchun faqat so'zning o'zi emas, uning atrofidagi kontekst ham tahlil qilinishi kerak. Har qanday slengning mazmun-mohiyati, u muayyan so'zlar bilan birgalikda kelganda, kontekstda anglashiladi. Nutqdan ajratilgan holatda ko'pchilik o'zbekcha slenglar slenglik ma'nosini yo'qotadi.

3. Tez yangilanish va eskirish. Tilda mavjud slenglar tez-tez yangilanib turadi. Bunga chet tillardan kirib keladigan slenglarni misol qilib keltirish mumkin. Masalan, “lol” so'zi kontekstdan kelib chiqib kulgu ifodasi yoki boshqa ma'no berishi mumkin. Teglash modellarining ustunligi shundaki, ular kontekstni inobatga olib, to'g'ri teglash imkonini yaratadi.

4. Slengning aniqligi va chegaralari. Aksariyat slenglar rasmiy lug'atlarga kirmagan bo'lishi mumkin va u kontekstdan kontekstga o'zgaradi.

5. Slenglar diaxron (vaqt o'tishi bilan o'zgaruvchan) va diatopik (mintaqaga bog'liq) bo'ladi. Masalan: “otvet qaytarish” – oddiy javob berish emas, balki ko'proq “keskin munosabat bildirish” ma'nosini anglatadi.

6. Korpus muammosi (ma'lumot yetishmovchiligi). Ko'pchilik NLP (Natural Language Processing) modellari rasmiy, toza matnlar (masalan, yangiliklar, ilmiy maqolalar) asosida o'rgatilgan. Ma'lumotlar yetarli bo'lmagani uchun “low-resource language” muammosi yuzaga kelishi mumkin. Modellar (masalan, BERT, RoBERTa) kontekstni yaxshi tushunsa ham, slengda bu har doim yetarli bo'lmaydi.

7. Yangi slenglar va dinamik rivojlanish. Sleng tezda yangilanadi. Yangi so'zlar ijtimoiy tarmoqlar, trendlar orqali tez tarqaladi.

Slenglarni avtommatik teglash (ijtimoiy tarmoq, chat, forumlar asosida), kontekstga moslashtirilgan modellarini o'rganish, slenglar bazasini kengaytirish (xalqdan to'plangan misollar bilan) orqali yanada samarali natijalar berish mumkin. Slenglarni teglash (ya'ni, ularni belgilash, ajratish yoki aniqlash) tilshunoslikda, tarjima, NLP (natural language processing), hamda madaniyatshunoslikda muhim ahamiyatga ega. Slenglar – bu tilning norasmiy, ko'proq og'zaki shaklida ishlatiladigan elementlari. Ularni teglash orqal sleng va adabiy til o'rtasidagi farq aniqlanadi, til o'zgarishi va rivojlanishini kuzatish osonlashadi, muayyan yosh guruhlari yoki submadaniyatlar tilini o'rganish imkoniyati yaratiladi. Shuningdek, slenglar ko'p hollarda kontekstga bog'liq bo'ladi va ularni noto'g'ri tarjima qilish

mazmunni buzib qo'yadi. Slengni aniqlash to'g'ri ekvivalent topishga yordam beradi. So'zning ohangini va uslubini saqlab qolishga xizmat qiladi. Masalan, chatbotlar, avtomatik tarjimonlar, yoki ijtimoiy tarmoqdagi kontentni tahlil qiluvchi tizimlar uchun to'g'ri ma'nosini tushunish hamda kontekstga mos javob qaytarish orqali slengni aniqlay olish muhim. Qaysi sleng qaysi mintaqada, yosh guruhida yoki submadaniyatda keng ishlatilishini aniqlash mumkin. Qolaversa, ijtimoiy tarmoqlardagi til o'zgarishlarini tahlil qilish qulaylashadi.

Yuqorida keltirganimiz singari slenglarni aniqlashda statistik modellardan foydalanish keng tarqalgan. Masalan, n-gram modellari yoki TF-IDF (Term Frequency-Inverse Document Frequency) usuli orqali matndagi odatiy bo'lmagan birliklar ajratib olinadi. n-gram modellar – bu matndagi so'zlar ketma-ketligini o'rganishga asoslangan statistik model. Ular matndagi so'zlarni qismlarga ajratib, o'xshashliklarni aniqlashda yordam beradi. n – bu ketma-ketlikdagi so'zlar sonini bildiradi [1], masalan:

- 1) 1-gram (unigram): har bir alohida so'z.
- 2) 2-gram (bigram): ikki so'zli ketma-ketlik (masalan, “yaxshi kitob”).
- 3) 3-gram (trigram): uch so'zli ketma-ketlik (masalan, “men kitob o'qiyman”).

n-gram modellarining afzalligi shundaki, ular kontekstga asoslanib, o'xshash so'z birikmalarini aniqlashda yaxshi natijalar beradi. n-gram modelida slengni aniqlashda, odatiy bo'lmagan so'zlar yoki qisqartmalarning yuqori chastotaga ega bo'lgan kombinatsiyalarini topish mumkin. Shuningdek, slenglarning tez-tez ishlatiladigan shakllari yoki yangi paydo bo'lgan shakllari n-gram modellarini qo'llash orqali aniqlanadi.

Kelgusida slenglarni aniqlash va teglash sohasida quyidagi yo'nalishlarda ilmiy izlanishlar olib borish bilan bog'liq vazifalar mavjud:

T/R	Yo'nalish	Mexanizm
1	Til korpusini kengaytirish	Sleng birliklarini doimiy ravishda yangilab borish orqali
2	Ko'p tillilikni qo'llab-quvvatlash hamda tildagi rang-baranglikni saqlash	Sleng tilni jonlantiradi. Teglash orqali bu boylikni yo'qotmasdan ko'rsatish mumkin. Ko'p tillilikda bu juda ham muhim.
3	Kontekstga asoslangan dinamik teglash tizimlarini yaratish	Tarjima yoki tilni avtomatlashtirish tizimi kontekstni to'g'ri tushunadi
4	Real vaqt rejimida ishlaydigan slengni aniqlash botlarini ishlab chiqish	Til o'rganish jarayonida yoki avtomatik tarjima qilinayotgan tizimlar slenglarni ajrata oladigan bo'lsa, ular noto'g'ri tushunish xavfidan saqlanadi.
5	Ko'p modalli yondashuv	Sleng ko'pincha matn bilan cheklanmaydi – videolar, audio formatlarda ham keng qo'llaniladi.

		Slengni aniqlashni faqat matnga emas, balki boshqa formatlarga ham kengaytirish.
6	Slengni aniqlash jarayonini tezlashtirish va zamonaviy o'zgarishlarga moslashuvchan qilish	Slenglarni tezkor aniqlash uchun umumiy kontekstni tushunish modellarini qo'llash orqali.

Xulosa qilib aytganda, matnlarda slenglarni avtomatik aniqlash va teglash zamonaviy tilshunoslik va sun'iy intellekt kesishmasida dolzarb tadqiqot yo'nalishidir. Bundan tashqari, ijtimoiy tarmoqlar monitoringi, brend reputatsiyasini boshqarish, ijtimoiy tahlil va psixolingvistik tadqiqotlar uchun ham slengni avtomatik aniqlash muhim ahamiyat kasb etadi. Amaliy natijalar shuni ko'rsatadiki, chuqur o'rganishga asoslangan modellar yuqori aniqlikni ta'minlaydi. Kelgusida algoritmlarni takomillashtirish va ma'lumotlar bazasini kengaytirish orqali ushbu tizimlarni yanada samarali qilish mumkin.

Foydalanilgan adabiyotlar:

1. Devlin J., Chang M. W., Lee K., & Toutanova K. (2019). “BERT: Pre-training of deep bidirectional transformers for language understanding.” NAACL.
2. Eisenstein, J. (2013). “What to do about bad language on the internet.” Proceedings of NAACL.
3. Zhang Y., and Yang Q. (2018). “An overview of multi-task learning in deep neural networks.” IEEE Transactions.
4. Joachims T. – Text categorization with support vector machines: Learning with many relevant features. 1998, European Conference on Machine Learning, 137-142 p.
5. Матюшенко Е. Современный молодежный сленг: формирование и функционирование. автореф. дисс. кандидата филологических наук – Волгоград, 2007. – 4 б.
6. Muhamedova S. Kompyuter lingvistikasi (metodik qo'llanma). – Toshkent, 2007.; Po'latov A., Muhamedova S. Kompyuter lingvistikasi (o'quv qo'llanma). – Toshkent, 2008 – 98-b.