SENTIMENT TAHLILI UCHUN LINGVISTIK TA'MINOTNI ANNOTATSIYALASH SXEMASI VA KO'RSATMALARI

Allanazarova Saboxat Yusupboyevna
ToshDO'TAU tayanch doktoranti
allanazarova.sabosh@gmail.com
https://doi.org/10.5281/zenodo.17060539

Annotatsiya. Ushbu maqolada sentiment tahlili uchun lingvistik ta'minotni ishlab chiqishda annotatsiyalash sxemasi va ko'rsatmalarni yaratish masalalari tahlil qilinadi. Taklif etilgan sxema matnlarning baholovchi xususiyatlarini aniqlash, ya'ni ijobiy, salbiy va neytral munosabatlarni belgilash imkonini beradi. Shuningdek, emotsional-ekspressiv birliklarning identifikatsiyasi ham ko'zda tutiladi. Annotatsiyalash jarayonini standartlashtirish uchun ishlab chiqilgan ko'rsatmalar annotatorlarning bir xil yondashuvni qo'llashiga yordam beradi va izchillikni ta'minlaydi. Tadqiqot natijalari o'zbek tili matnlarini sentiment tahlil qilishda samarali lingvistik resurslarni yaratishda qo'llanishi mumkin.

Kalit so'zlar: sentiment tahlili, lingvistik ta'minot, annotatsiya, annotatsiyalash sxemasi, emotsional birliklar

Аннотация: В статье рассматриваются вопросы разработки схемы и инструкций аннотирования при создании лингвистического обеспечения анализа сентимента. Предлагаемая схема позволяет выявлять оценочные характеристики текста, включая положительные, отрицательные и нейтральные отношения, а также идентифицировать эмоционально-экспрессивные единицы. Разработанные инструкции направлены на стандартизацию процесса аннотирования, что обеспечивает единообразие подхода аннотаторов и согласованность получаемых результатов. Результаты исследования могут быть использованы при создании эффективных лингвистических ресурсов для анализа сентимента в узбекских текстах.

Ключевые слова: анализ сентимента, лингвистическое обеспечение, аннотация, схема аннотирования, эмоциональные единицы

Abstract: This paper explores the development of an annotation scheme and guidelines for building linguistic resources to support sentiment analysis. The proposed scheme enables the identification of evaluative features in texts, including positive, negative, and neutral attitudes, as well as emotional-expressive units. The guidelines are designed to standardize the annotation process, ensuring a unified approach among annotators and improving the consistency of results. The outcomes of this research can be applied to the creation of effective linguistic resources for sentiment analysis in the Uzbek language.

Keywords: sentiment analysis, linguistic resources, annotation, annotation scheme, emotional units

Annotatsiyalangan ma'lumotlar - bu kompyuter tushunishi uchun til birliklarini tegishli ma'lumotlar bilan razmetkalash jarayoni. Ushbu ma'lumotlar lingvistik va ekstralingvistik (rasm, matn, audio yoki video) ko'rinishda bo'lishi mumkin. Ma'lumotlar izohi dastlabki bosqichda qo'lda, keyin ilg'or mashinaviy o'qitish algoritmlari va vositalaridan foydalangan holda avtomatik ravishda amalga oshirilishi mumkin[1]. Annotatsiyalash metodologiyasi, xususan qo'lda annotatsiyalash, sentiment leksikasini ishlab chiqish jarayonida muhim qadamdir. Annotatsiya jarayonida "avaylamoq" so'ziga ijobiy belgi (P), "nafrat" so'ziga esa salbiy belgi (N) qo'yildi. Bunda annotatorlar tomonidan matnlarga sentiment belgilari

qo‘shiladi. Ma’lumki, qo‘lda annotatsiyalangan matnlar ML modellarini o‘qitish uchun ishlatiladi va qo‘lda yoki yarim avtomatik tarzda ishlab chiqiladi. Qo‘lda annotatsiyalash jarayoni uzoq vaqt talab qiladigan jarayon bo‘lsa-da, mutaxassislar jalb qilinganligi sababli boshqa usullarga qaraganda aniqroq hisoblanadi[2]. Ayniqsa o‘rgatilgan leksik resurslari yo‘q tillarda hissiyotlarni tahlil qilishni o‘rganuvchilar tomonidan qo‘llaniladi. Biroq tasniflashdagi subyektiv omillar sabab, annotator xatoga yo‘l qo‘yishi mumkin[3]. Yarim avtomatik usullar qo‘lda va avtomatlashtirilgan yondashuvlarni birlashtiradi.

Qo‘lda annotatsiyalash yuqori aniqlikdagi sentiment leksikasini ishlab chiqishda juda muhim, ammo vaqt va resurs talab qilishi sababli avtomatlashtirilgan yondashuvlar bilan qo‘shib qo‘llanilishi mumkin. Qo‘lda izohlangan va avtomatik ravishda kengaytirilgan leksikalarni birlashtirish ikki xil yondashuvni birlashtirishni anglatadi: inson tomonidan qo‘lda tahrirlangan, batafsil izohlar berilgan leksikalar va avtomatik dasturlar tomonidan tahlil qilinib, kengaytirilgan leksikalar[4].

Jarayonda 6 nafar annotatorlar o‘zbek tili va uning xususiyatlarini yaxshi biladigan mutaxassislar (lingvistlar) ishtirok etgan, ular tomonidan har bir sinonimik sinset sentiment sinfi bo‘yicha belgilab chiqilgan. Bu jarayonda annatatorlarning fikri mos kelishi asosiy me’zon hisoblanadi. Tasniflash ishida birorta ham ekspert bilan fikri mos kelmagan annotator ishi qoniqarsiz hisoblanadi va chetlatiladi[5]. Shu sababli label_4 natijalari asosiy tadqiqot natijalarida aks ettirilmadi. Shu o‘rinda ta’kidlash kerakki, ma’lum sinonimik sinsetni ijobiy, salbiy yoki betaraf deb topilishi uchun kamida 3 ta annotator ovozi bo‘lishi shart.

Label_1 dan **label_6** gacha bo‘lgan ustunlar turli annotatorlar tomonidan qo‘lda berilgan hissiy belgilashlarni ifodalaydi. Bu ko‘p annotatorli yondashuv orqali inson subyektivligining o‘rganilishi ta’minlandi. **final_label** – ko‘p annotatsiyalash asosida aniqlangan yakuniy hissiy toifa (masalan: *ijobiy*, *salbiy*, *betaraf*).

1-jadval:

Annotatorlararo kelishuv darajasi

Kappa qiymati	Kelishuv darajasi:	Moslik qiymati%
0–.20	Hech qanday	0–4%
.21–.39	Minimal	4–15%
.40–.59	Zaif	15–35%
.60–.79	o‘rtacha	35–63%
.80–.90	Yuqori	64–81%
Above.90	Mukammal	82–100%

Quyida Cohen’s Kappa[1] ni hisoblash jarayoni label_1 va label_2 annotatsiyalari asosida bosqichma-bosqich tushuntiriladi:

1. Annotatorlarni tayyorlash

label_1 va **label_2** – bu ikki mustaqil baholovchi (annotator) tomonidan berilgan belgilar yoki kategoriyalar. Har bir obyekt (masalan: matn, rasm yoki boshqa ma’lumot) uchun ikkala annotator ham o‘z bahosini (kategoriya yoki belgi) beradi. Har ikki annotatorning baholari

[1] https://pmc.ncbi.nlm.nih.gov/articles/PMC3900052/

asosida ularning kelishuv darajasi aniqlanadi. Bu turli baholovchilar orasidagi subyektiv farqlarni va baholash metodikasidagi moslikni tekshirishga yordam beradi.

2. Kontingensiya jadvalini (krosjadval) tuzish

- Har bir obyekt uchun **label_1** va **label_2** qiymatlarini yozib chiqiladi.
- Jadvalning qatorlari birinchi annotatorning (label_1) kategoriyalarini, ustunlari esa ikkinchi annotatorning (label_2) kategoriyalarini ifodalaydi.
- Har bir katakchada mos keluvchi kombinatsiyaga tushgan obyektlar soni yoziladi.

Bu jadval yordamida qaysi kategoriyalarda kelishuv (diagonal elementlar) va kelishmovchiliklar (diagonaldan tashqaridagi elementlar) sodir bo‘lganini ko‘rishingiz mumkin.

3. Kuzatilgan kelishuv ehtimolligini (P_o) hisoblash

- **P_o** — bu ikkala annotator bir xil kategoriya tanlagan holatlar ulushi.
- Jadvalning **diagonal** elementlarini (ya’ni, label_1 = label_2 bo‘lgan hollarda) yig‘iladi.
- Shu yig‘indi qiymatini jami obyektlar soni (N) ga bo‘linadi:

$$P_o = \frac{\text{Diagonaldagi jami}}{N}$$

Bu qadam orqali annotatorlar orasida qanchalik ko‘p hollarda rozilik (kelishuv, moslik) borligini aniqlandi.

4. Marginal ehtimollarni hisoblash va kutilgan kelishuv ehtimolligini (P_e) aniqlash

- Har bir kategoriya uchun:

- **label_1** tomonidan tanlangan obyektlar sonining umumiy obyektlar soniga nisbati (qator ulushi) aniqlanadi.
- **label_2** tomonidan tanlangan obyektlar sonining umumiy obyektlar soniga nisbati (ustun ulushi) aniqlanadi.

- Har bir kategoriya uchun kutilgan kelishuv ehtimoli shu ikki ehtimolni ko‘paytirish orqali hisoblanadi.
- Keyin, barcha kategoriyalar bo‘yicha hosil bo‘lgan qiymatlarni qo‘shib, **P_e** ni topiladi:

$$P_e = \sum_i \left(\frac{\text{label\_1 uchun kategoriya } i \text{ soni}}{N} \times \frac{\text{label\_2 uchun kategoriya } i \text{ soni}}{N} \right)$$

Bu qadamda har ikki annotatorning mustaqil ravishda tasodifiy tanlash ehtimoli hisoblanadi. Ya’ni ular tasodifiy ravishda belgi(baho)lar berayotganida, qaysi hollarda bir xil belgi tanlanishi kutilishini aniqlanadi.

5. Cohen’s Kappa ni hisoblash

- Quyidagi formuladan foydalanib, **Cohen’s Kappa** qiymati hisoblanadi:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

P_o – kuzatilgan (observed) kelishuv ehtimolligi.
P_e – tasodifiy (expected) kelishuv ehtimolligi.

Bu formula orqali, tasodifiy kelishuv ehtimolini hisobga olib, haqiqiy kelishuv darajasi aniqlandi. Agar annotatorlar tasodifiy ravishda turli baholar bersalar, kappa qiymati 0 ga yaqin bo‘ladi; agar ular bergan baho to‘liq kelishsa (mos kelsa), kappa qiymati 1 bo‘ladi.

6. Natijalarni talqin qilish

- Hisoblangan **κ** qiymatini quyidagi mezonlarga muvofiq talqin qilishimiz mumkin:

0,00–0,20: Hech qanday yoki juda past kelishuv
0.21–0.39: Minimal kelishuv
0.40–0.59: Zaif kelishuv
0.60–0.79: O‘rtacha kelishuv
0.80–0.90: Kuchli kelishuv
0.90 dan yuqori: Deyarli mukammal kelishuv

Quyidagi jadvalda annatatorlar bergan belgilarning moslik darajasi ifodalangan, bu yerda o‘rtacha Cohen’s Kappa qiymati 0,79 ni ko‘rsatgan.

2 – jadval:

Cohen’s Kappa qiymatlari

№	label_1	label_2	label_3	label_5	label_6
label_1	1	0,72	0,77	0,83	0,83
label_2		1	0,69	0,79	0,78
label_3			1	0,79	0,77
label_5				1	0,94
label_6					1

Talqin qilish orqali annotatorlar orasidagi kelishuv darajasini aniqlab, baholash jarayonining ishonchliligi va obyektivligini baholashga erishildi.

3-jadval

Annotatsiyalangan leksemalarning so‘z turkumi bo‘yicha taqsimoti

So‘z turkumi	**Sinsetlar soni:**
Ot	1499
Sifat	1096
fe'l	249
Ravish	172
Modal	55
Yordamchi so‘z	59
Olmosh	6
Jami	3136

Birinchi 3 136 ta sinsetni qo‘lda yorliqlash jarayoni 663 ta ijobiy, 1 076 ta salbiy va 1 403 ta betaraf ma’lumotlarga olib keldi, bunda har bir so‘zning o‘zi anglatgan ma’nosi va izohiga qarab qarorlar qabul qilindi.

Xulosa qilib aytganda, sentiment tahlili uchun annotatsiyalash jarayoni lingvistik ta’minot yaratishda muhim bosqich hisoblanadi. Tadqiqot davomida 6 nafar lingvist annotator tomonidan qo‘lda yorliqlash amalga oshirildi, bunda sinonimik sinsetlar ijobiy, salbiy yoki betaraf toifalarga ajratildi. Annotatorlararo kelishuv darajasi Cohen’s Kappa koeffitsienti yordamida baholandi va o‘rtacha qiymat 0,79 ni tashkil etdi, bu esa annotatsiya jarayonining ishonchliligini ko‘rsatadi. Qo‘lda annotatsiyalash uzoq vaqt va resurs talab qilsa-da, yuqori

aniqlikni ta’minladi hamda yarim avtomatik usullar bilan qo‘shib qo‘llash orqali samaradorlik oshirildi. Tadqiqot natijasida 3 136 ta sinset annotatsiya qilinib, ularning orasida 663 tasi ijobiy, 1 076 tasi salbiy va 1 403 tasi betaraf deb belgilandi. Ushbu resurs o‘zbek tili uchun sentiment leksikonini yaratishda va mashinaviy o‘qitish modellarini tayyorlashda muhim amaliy ahamiyatga ega.

References:

1. Du, J., Wang, D., Lin, B., He, L., Huang, L. C., Wang, J., Manion, F. J., Li, Y., Cossrow, N., & Yao, L. (2025). Use of deep learning-based NLP models for full-text data elements extraction for systematic literature review tasks. *Scientific reports*, *15*(1), 19379. https://doi.org/10.1038/s41598-025-03979-5
2. Amusat, O. O., Hegde, H., Mungall, C. J., Giannakou, A., Byers, N. P., Gunter, D., Fagnan, K., & Ramakrishnan, L. (2024). Automated annotation of scientific texts for ML-based keyphrase extraction and validation. *Database : the journal of biological databases and curation*, *2024*, baae093. https://doi.org/10.1093/database/baae093
3. Mozetič, I., Grčar, M., & Smailović, J. (2016). Multilingual Twitter Sentiment Classification: The Role of Human Annotators. *PloS one*, *11*(5), e0155036. https://doi.org/10.1371/journal.pone.0155036
4. Samreen, A., & Ali, S. A. (2023). Dataset construction to detect human behavior with the help of emotions, sentiments and mood for Roman Urdu. *Data in brief*, *52*, 109906. https://doi.org/10.1016/j.dib.2023.109906
5. Yang, F., Zamzmi, G., Angara, S., Rajaraman, S., Aquilina, A., Xue, Z., Jaeger, S., Papagiannakis, E., & Antani, S. K. (2023). Assessing Inter-Annotator Agreement for Medical Image Segmentation. *IEEE access : practical innovations, open solutions*, *11*, 21300–21312. https://doi.org/10.1109/access.2023.3249759