TIL KORPUSINI TEGLASH JARAYONIDA ANNOTATORLARNING ROLI, IAA O‘LCHOVLARI VA TEGLASH SIFATINI TA’MINLASH METODLARI

Xusainova Zilola Yuldashevna,
f.f.f.d. (PhD), dotsent v.b.
xusainovazilola@navoiy-uni.uz
ToshDO‘TAU

Annotatsiya. Tabiiy tilni qayta ishlash (NLP) sohasida ishonchli teglangan ma’lumotlar mashinali o‘qitish (ML) va modellarni baholash uchun muhim ahamiyatga ega. Ushbu maqolada teglash jarayonining bosqichlari keltiriladi va annotatorlararo kelishuv (IAA) mezonlari asosida teglash izchilligi baholanadi. Matnli korpus tayyorlash, uni teglash ko‘rsatmalarini ishlab chiqish, hamda bir nechta annotatorlar tomonidan mustaqil ravishda matnni teglash bosqichlari bayon etiladi. Annotatorlar mustaqil tarzda so‘z turkumlarini teglash (POS) va nomlangan obyektlarni aniqlash (NER) kabi NLP vazifalari bo‘yicha matnni teg bilan belgilaydi. Annotatorlararo kelishuv darajasi Kohen Kappasi (ikki annotator o‘rtasidagi kelishuv), Fleiss Kappasi (bir nechta annotatorli holatda) va Krippendorff Alpha (umumiy kelishuv o‘lchovi) kabi statistik ko‘rsatkichlar yordamida baholanadi. Olingan natijalar o‘rtacha 0,75 atrofidagi Kappa qiymatini va yuqori kelishuv darajasini ko‘rsatadi. Bu esa teglash natijasining yuqori ekanligini tasdiqlaydi. Shuningdek, annotatorlar o‘rtasidagi tafovutlar teglash bo‘yicha ko‘rsatmalarni yanada takomillashtirish zarurligini ko‘rsatadi. Teglash ko‘rsatmalari qanchalik aniq ishlab chiqilgan bo‘lsa, kelishuv darajasi shunchalik yuqori bo‘ladi. Teglash jarayonida IAA kelishuv ko‘rsatkichlarini qo‘llash ma’lumotlar sifatini oshirishga xizmat qiladi.

Kalit so‘zlar: *Teglash jarayoni, annotatorlararo kelishuv, Inter-Annotator Agreement (IAA), Kohen Kappa, Fleiss Kappa, Krippendorff Alpha, tabiiy tilni qayta ishlash (NLP), o‘lchov metrikalari.*

Kirish. Tabiiy tilni qayta ishlash sohasida ma’lumotlarni **qo‘lda teglash (annotatsiya)** jarayoni muhim o‘rin tutadi. Teglash – bu matn, tasvir yoki boshqa ma’lumotlarga inson mutaxassislari tomonidan teglar yoki kategorial belgi qo‘yish jarayoni bo‘lib, mashinali o‘qitish (ML) modellari uchun ma’lumotlar to‘plamini tayyorlashga xizmat qiladi. Teglash sifatli amalga oshirilsa, modelni o‘qitish va baholash jarayonlari aniqroq bo‘ladi. Aksincha, noto‘g‘ri teglangan ma’lumot NLP tizimi natijalarining yomonlashuviga olib keladi[11]. Shuning uchun annotatorlar, ya’ni teglashni amalga oshiruvchi mutaxassislar hal qiluvchi ahamiyatga ega.

Annotatorlar jamoasi odatda soha bo‘yicha mutaxassislar (Subject Matter Experts)dan iborat bo‘lib, ular maxsus **yo‘riqnomalar (guidelines)** asosida teglashni amalga oshiradi[6:155-158]. Teglash jarayonida har bir annotator mustaqil ishlashi, ammo umumiy qoidalarga amal qilishi lozim. Annotatorlar o‘rtasida bir xil ma’lumotga bir xil teg berish *annotatsiya sifatining muhim ko‘rsatkichi* hisoblanadi. Annotatorlarning qarorlari o‘zaro qanchalik mos kelishini **Inter-Annotator Agreement (IAA),** ya’ni *annotatorlararo kelishuv ko‘rsatkichi* orqali o‘lchash qabul qilingan [1: 302]. IAA aslida bir xil ma’lumotlar ustida bir necha annotator mustaqil ishlaganda, ularning natijalari necha foiz hollarda o‘zaro mos tushishini ifodalaydi[2: 586-591]. Ma’lumotlarni teglash natijasi past bo‘lsa, bu annotatsiya jarayonida muammo borligini va ma’lumotlar sifati pastligini anglatadi. IAA darajasi yuqori bo‘lsa, aksincha, teglash ishonchliroq va *“oltin standart”*ga yaqin bo‘ladi [5: 42-45].

Ushbu maqolada teglash jarayoni va annotatorlar roli (mutaxassislarni tanlash, ularni tayyorlash va ishini monitoring qilish) tahlil qilinadi. Shuningdek, Inter-Annotator Agreement (IAA) jarayonidagi *o‘lchov metrikalari – bevosita kelishuv* (*Observed Agreement*), *Kohen Kappasi[5], Fleiss Kappasi* va *Krippendorff Alpha[15]* formulalar bilan ko‘rsatiladi. So‘ngra *teglash sifatini ta’minlash metodlari* – konsensusga erishish, annotatorlarni o‘qitish, bir nechta annotator

ishtirokida sinovlar va tafovutlarni hal etish jarayoni haqida so‘z yuritiladi. Natijada, teglash sifati va IAA ko‘rsatkichlari o‘rtasidagi bog‘liqlik tahlil qilinadi. Teglash sifati past bo‘lganda, mashinali o‘qitish modeli natijalariga ta’siri tajribalar va ilmiy manbalar asosida muhokama qilinadi.

Annotatorlarni tanlash, tayyorlash va monitoring qilish

Teglash jarayonini samarali tashkil qilish maqsadida annotatorlarni to‘g‘ri tanlash va ularni teglash vazifasiga tayyorlash zarur. Amaliyotda, avvalo, teglash talab qilinayotgan soha bo‘yicha yetarli bilim va ko‘nikmaga ega mutaxassislar jalb etiladi[11]. Masalan, matnlardagi **POS teglash** NLP vazifasi uchun tilshunoslar, **tibbiy ma’lumotlarni teglash** uchun tibbiyot mutaxassislari tanlanadi. Agar vazifa murakkab bo‘lsa, annotatorlar maxsus *o‘quv-treningda* o‘qitilishi tavsiya etiladi.

Teglash yo‘riqnomasi – har bir annotator amal qilishi lozim bo‘lgan qoidalar to‘plami bo‘lib, oldindan ishlab chiqiladi va mutaxassislarga tushuntiriladi. Aniq va batafsil ko‘rsatmalar annotatorlarning vazifani bir xil tushunib, izchil teglashiga yordam beradi. Teglash jarayonida annotatorlar ishini monitoring qilish nihoyatda muhim bosqich hisobalanib, quyidagi ikki xil yondashuv asosida amalga oshiriladi:

1) *Individual taqsimlash.* Har bir annotator ma’lumotlar to‘plamining alohida (o‘zaro ustma-ust tushmagan) qismini teglaydi. Bunda, annotatorlar natijalari to‘g‘ridan-to‘g‘ri solishtirib bo‘lmaydi, shuning uchun **sifat nazorati** uchun har bir annotator teglagan ma’lumotlarning bir qismini alohida ekspert tomonidan ko‘rib chiqish talab etiladi. Masalan, har bir annotator ishlagan hujjatlarning ma’lum foizi boshqasi tomonidan tekshirilishi mumkin. Bu orqali *yopiq ko‘rinishdagi* xatolar aniqlanadi.

2) *Ustma-ust tushuvchi taqsimlash.* Bu yondashuvda har bir ma’lumot bir necha (kamida ikki) annotator tomonidan mustaqil teglanadi. Bunda, biror matn bo‘yicha annotatorlar orasidagi tafovutlarni aniqlash mumkin bo‘ladi va to‘plam bo‘yicha **o‘rtacha IAA**ni hisoblash imkoniyati paydo bo‘ladi. Annotatorlar kelishuv

darajasi past bo‘lsa, loyiha rahbarlari yo‘riqnomani qayta ko‘rib chiqishi yoki annotatorlarni qayta o‘qitishi zarur. Amalga oshirilgan tajribalar shuni ko‘rsatadiki, katta *hajmdagi korpusni teglashni boshlashdan oldin* kichikroq sinov bosqichida IAA yuqori qiymatga erishilishiga ta’minlash kerak. Bir necha iterativ teglash va muhokamalar orqali jamoa yo‘riqnomani takomillashtirishi va annotatorlar bir xil tushunishini ta’minlashi lozim. Shundan so‘ng, asosiy korpusni teglash tavsiya etiladi.

Annotatorlar faoliyatini nazorat qilish davomida *doimiy fikr-mulohaza (feedback)* va *treninglar o‘tkazish* kerak. Tadqiqotlar shuni ko‘rsatadiki, *muntazam tazrda annotatorlarni o‘qitish va aloqalar teglash sifatini sezilarli darajada oshiradi*. Annotatorlar ishidagi nomuvofiqliklar va qiyinchiliklar muhokama qilinib borilsa, ular keyingi vazifalarda xatolarini tuzatadi va izchillik ortadi. Annotatorlarning kelishuv ko‘rsatkichlari (masalan, kappa)ni davriy ravishda hisoblab borish orqali *qaysi annotator ko‘proq xato teglayotganligi*, qaysi kategoriyalar noaniqlik keltirayotgani aniqlanadi. Shu tariqa, teglash jarayonini **real vaqt rejimida monitoring** qilib, sifati pasayib ketishining oldi olinadi.

Inter-Annotator Agreement (IAA) va uning o‘lchovlari

Annotatorlararo kelishuv – bir ma’lumot ustida ishlagan ikki yoki undan ortiq annotatorlar tomonidan teglangan ma’lumotlarning o‘zaro mos kelish darajasi. IAA qiymati qanchalik yuqori bo‘lsa, *ma’lumotlarni teglash sifati* shunchalik yaxshi hisoblanadi. IAAni turli usullar bilan miqdoriy ifodalash mumkin. Eng sodda usul – bevosita kelishuv foizini topish, ya’ni annotatorlar necha hollarda bir xil teg qo‘yganini aniqlash. Biroq, kelishuv foizi ba’zan xato bo‘lishi mumkin, chunki annotatorlarning *tasodifan bir xil javob tanlashi ehtimoli* ham mavjud. Shu sabab, tadqiqotlarda *“chance-corrected” (tasodifiy moslikka tuzatilgan) ko‘rsatkichlar* qo‘llaniladi[21:935-938]. Quyida IAAni o‘lchashda eng ko‘p qo‘llaniladigan metrikalar keltirilgan:

1. Bevosita kelishuv (Observed Agreement, P_o) – Bu ko‘rsatkich annotatorlar nechta obyekt bo‘yicha to‘liq bir xil teg qo‘yganini foizi sifatida aniqlanadi. Masalan, ikki annotator 100 ta holatdan 85 tasida bir xil teg qo‘ygan bo‘lsa, P_o=0.85 (ya’ni 85%).

$$P_o = \frac{annotatorlar\ bir\ xil\ qaror\ chiqargan\ holatlar\ soni}{umumiy\ holatlar\ soni}$$

P_o ko‘rsatkich tasodifiy moslikni hisobga olmaydi va shuning uchun yuqori qiymatlar doimo yuqori sifatni anglatmasligi mumkin. Masalan, kam kategoriyali yoki notekis taqsimlangan ma’lumotlarda tasodifan kelishuv yuqori chiqishi mumkin.

2. Kohen kappasi (Cohen’s Kappa, κ) – ikki annotator uchun kelishuv ko‘rsatkichi bo‘lib [4:300-302], Kohen Kappa statistikasi ikki annotator *bergan teglarning mosligi tasodifiy mos kelish ehtimolidan aniq farq qiladimi* degan savolga javob beradi. Kappa qiymati quyidagi formula orqali hisoblanadi[19:276-280]:

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

bu yerda $\boldsymbol{P_o}$ – yuqoridagi kuzatilgan (bevosita) kelishuv, $\boldsymbol{P_e}$ esa tasodifan kelishib qolish ehtimoli. $\boldsymbol{P_e}$ qiymatni hisoblash uchun har bir annotatorning teg taqsimotlari olinib, ixtiyoriy holatda bir xil teg berilish ehtimoli topiladi. Kappaning qiymati [-1..1] oralig‘ida bo‘lib, $\boldsymbol{\kappa = 1}$ to‘liq kelishuvni, $\boldsymbol{\kappa = 0}$ tasodifiy darajadagi moslikni, $\boldsymbol{\kappa < 0}$ esa tasodifdan *yomon* va *sistematik nomuvofiqlik* borligini bildiradi. Amaliyotda kappa **0.80 dan yuqori** bo‘lsa, *“deyarli mukammal”* kelishuv, [**0.60..0.80**] oralig‘ida bo‘lsa, *“yuqori”* yoki *“yetarli”* kelishuv, [**0.40..0.60**] oralig‘ida *“o‘rtacha”* kelishuv, bundan past qiymat esa jiddiy muammo borligini ko‘rsatadi (Landis va Koch mezonlari) [16:363-374]. Kohenning Kappa qiymati faqat ikki annotator va nominal klassifikatsiya uchun mo‘ljallangan.

3. Fleiss Kappasi. Ushbu metrika Cohen kappasining N ta annotator ishtirokidagi umumiy holga tatbiq qilingan shaklidir[18:]. Fleiss kappa tasodifiy

moslikka tuzatilgan kelishuvni baholaydi va *kategorik ma’lumotlar* uchun qo‘llanadi. Bunda, har bir obyekt bo‘yicha annotatorlar orasidagi kelishuv darajasi avval topiladi, so‘ngra butun korpus bo‘yicha bu qiymatlar o‘rtacha olinadi. Buni $\boldsymbol{P}$ deb belgilaymiz. Xuddi shu tarzda, har bir kategoriya bo‘yicha annotatorlarning umumiy tanlash ehtimoli kvadratlari yig‘indisi hisoblanadi. Bu tasodifiy kelishuv kutilmasi $\boldsymbol{P_e}$ bo‘ladi[10]. Fleiss kappasi formulasi ham Kohen kappa kabi:

$$\kappa_F = \frac{P - P_e}{1 - P_e}$$

Ushbu metrikani bo`yicha baholash Kohen kappaga o‘xshash tarzda amalga oshiriladi. Fleiss kappa metrikasida har bir obyekt bir xil sondagi annotatorlar tomonidan baholashga mo‘ljallangan. Masalan, har bir hujjatni 3 tadan annotator ko‘rgan bo‘lsa, foydalanish mumkin.

4. Krippendorff Alpha (α) – universal kelishuv ko‘rsatkichi hisoblanadi. IAAni o‘lchash uchun eng moslashuvchan va mukammal ko‘rsatkichlardan biri bo‘lib, u *istalgan sonli annotatorlar, turli tipdagi ma’lumotlar* (nominal, ordinal, interval, nisbiy) va hatto *to‘liq belgilarga ega bo‘lmagan ma’lumotlar* (masalan, ba’zi obyektlar faqat bitta annotator tomonidan baholangan holatlar) uchun ham hisoblash imkonini beradi. Krippendorff alpha metrikasida juftliklar orasidagi nomuvofiqlik (disagreement)ni tahlil qilinib, uni tasodifiy nomuvofiqlik bilan solishtiriladi:

$$\alpha = 1 - \frac{D_o}{D_e}$$

bu yerda $\boldsymbol{D_o}$ – kuzatilgan nomuvofiqlik darajasi (annotatorlar orasidagi farqlar yig‘indisi), $\boldsymbol{D_e}$ esa tasodifiy kutilgan nomuvofiqlik. Shuningdek, $\boldsymbol{\alpha}$ ham (kappa kabi) kuzatilgan va kutilgan qiymatlarning nisbatiga asoslanadi, lekin bu yerda to‘g‘ridan-to‘g‘ri moslik emas, balki farqlar ko‘rib chiqiladi. Nominal ma’lumotlar uchun $\boldsymbol{\alpha}$ ning hisoblash natijasi Kohen kappa va Fleiss kappa bilan bir xil natija berishi mumkin[12], biroq $\boldsymbol{\alpha}$ metrikasi *ko‘p qirrali*: masalan, ordinal reytinglar

uchun farqlar kvadratik og‘irlik bilan olinib, **Weighted Kappa**ga teng natija beradi. $\boldsymbol{\alpha}$ qiymatining interpretatsiyasi ham kappa kabi: **1** – mukammal kelishuv, **0** – tasodifiy, **manfiy** – kelishmovchilik kuchli. Ko‘p hollarda teglash natijasini baholashda *Krippendorff’s alpha metrikasi tavsiya etiladi.* Chunki u turli hollarda qo‘llaniladi va bir necha turdagi kelishuv ko‘rsatkichlarini umumlashtiradi.

Teglash sifatini baholashga qaratilgan tadqiqotlarda boshqa IAA ko‘rsatkichlari ham mavjud. Masalan, **Scott’s Pi** [20:321-325] Kohen kappa metrikasiga o‘xshash metrika bo‘lib, ikki annotator uchun bir xil formulaga ega (faqat P_e boshqacha hisoblanadi). Shuningdek, **Gwet’s AC1/AC2, Pearson va Spearman korrelyatsiyalari** kabi o‘lchovlar muayyan hollarda qo‘llaniladi. Biroq, yuqorida sanab o‘tilgan to‘rtta metrikalaridan ko‘p hollarda foydalaniladi. Kelishuv ko‘rsatkichlari vositasida teglash jarayonining **ishonchliligini miqdoran baholash** mumkin bo‘ladi.

Teglash sifatini ta’minlash usullari

Teglash sifatini yuqori darajada ta‘minlash uchun bir qator metodik yondashuvlar qo‘llaniladi. Quyida asosiy usullar keltirilgan:

1. *Aniq yo‘riqnoma.* Teglash jarayonini boshlash uchun **aniq va to‘liq yozma yo‘riqnoma** ishlab chiqiladi. Unda har bir kategoriya yoki tegning ta’rifi, misollar, chegaraviy holatlar (edge cases) bo‘yicha ko‘rsatmalar bo‘ladi. Annotatorlar ushbu ko‘rsatmalarni to‘liq tushunganiga ishonch hosil qilish uchun birgalikdagi trening o‘tkaziladi. Shuningdek, teglash jarayoni davomida yangi noaniq holatlar yuzaga kelganda, yo‘riqnoma *dinamik tarzda yangilanib* boradi.

2. *Bir necha annotator va consensus.* Muhim ma’lumotlar albatta **kamida ikki nafar annotator** tomonidan teglanishi va ularning natijalari solishtirilishi kerak. Agar natijalar orasida farq bo‘lsa, oddiy konsensus algoritmlari qo‘llaniladi: masalan, ko‘pchilikning fikri bo‘yicha final yorliq tanlanadi yoki muhokama orqali umumiy qarorga kelinadi[3:2940-2945]. Konsensusli teglash individual annotator

subyektivligining ta’sirini kamaytiradi va aniqroq natijani topishga xizmat qiladi. Ba’zi platformalar avtomatik konsensus olish vositalarini taklif qiladi. Masalan, agar 3 annotatordan 2 nafari bir xil teg qo‘ysa, shu teg qabul qilinadi, agar annotatirlar uch xil teg qo‘ysa, bu obyekt *“nizo”* sifatida belgilanib, keyingi tekshiruvga uzatiladi.

3. *Adjudikatsiya (nizolarni hal etish).* Bir nechta annotatorlar bergan teglarida tafovut yuzaga kelganda, **adjudikator** deb ataluvchi tajribali mutaxassis yakuniy hakamlikni amalga oshiradi[13]. Adjudikatsiya – bu bir obyektga tegishli bir necha versiyadagi teglarni solishtirib, ziddiyatlarni bartaraf etish jarayoni. Ko‘pincha adjudikator jamoaning katta tajribaga ega a’zosi bo‘ladi yoki annotatorlar guruhi birgalikda muhokama qilib, bahsli holatlar yuzasidan kelishilgan qarorga keladi. Masalan, matnning bir qismi bitta annotator tomonidan *“neytral”*, boshqasi tomonidan *“salbiy”* deb teglangan bo‘lsa, adjudikator matnni qayta ko‘rib, yo‘riqnoma mezonlariga muvofiq to‘g‘ri tegni belgilaydi va shu bilan **oltin standart**ni shakllantiradi. Adjudikatsiya jarayoni ayniqsa teglash jarayoni boshlanishida muhim hisoblanib, jamoa birgalikda murakkab holatlarni tahlil qilib, kelgusida qanday teglash borasida bir qarashga ega bo‘ladi.

4. *Davriy nazorat va tahlil (review cycle).* Teglash davomida sifatni oshirish uchun **doimiy nazorat** va tahlillar o‘tkazib turiladi. Tajribali annotatorlar *yangi teglangan ma’lumotlarni tekshiradi, bir-birining ishini ko‘rib chiqadi* va *nomuvofiqliklarni muhokama qiladi*. Bu jarayon xatolarni dastlabki bosqichda tuzatishga va barchaning izchil ishlashiga xizmat qiladi. Masalan, har hafta jamoa to‘planib, murakkab teglangan misollarni (difficult cases) muhokama qilishi lozim. Teglangan ma’lumotlarni birgalikda ko‘rib chiqish annotatorlar o‘rtasida muloqotni kuchaytiradi va tushunmovchiliklarni bartaraf etadi.

5. *Avtomatlashtirilgan sifat nazorati.* Katta hajmdagi ma’lumotlarni teglashda inson nazorati bilan birga **avtomatik sifat tekshiruvlari (quality screens)** joriy

etilishi mumkin. Bunda teglangan ma’lumotlardagi **mantiqiy bog‘liqsizliklar** yoki qoidalarga zid holatlar dastur vositasida aniqlanadi. Masalan, agar annotator matnning biror segmentiga ikki xil teg qo‘ysa yoki noto‘g‘ri formatda kiritsa, tizim buni alohida nazoratga oladi. Yana bir usul – oldindan ma’lum “oltin” namunalarni teglashga qo‘shib, annotatorning ulardagi aniqligini kuzatish. Agar annotator bu nazorat namunalarida xato qilsa, uni ogohlantirish yoki qo‘shimcha trening o‘tkazilishi kerak**.** Sifat nazorati yordamida teglash jarayonidagi aniqlik va ishonchlilik saqlab qolinadi.

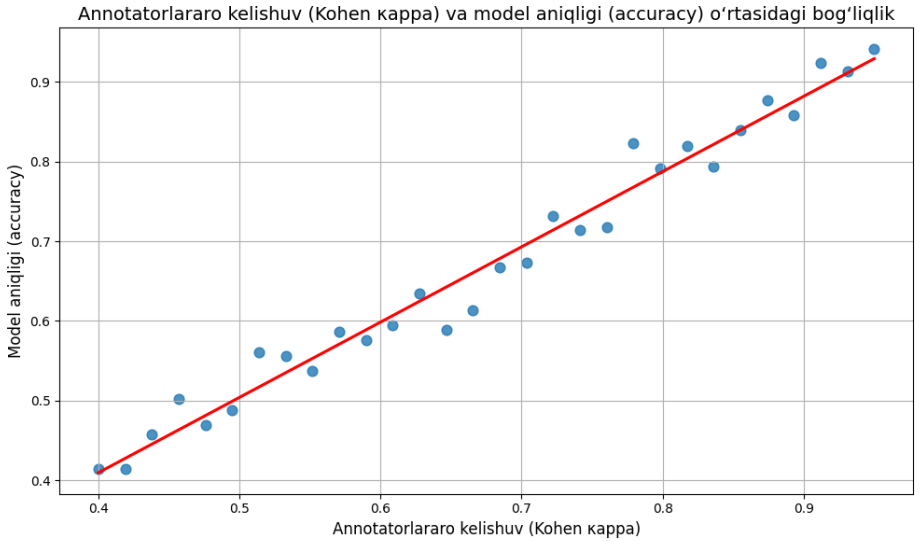
6. *Baholash:* Teglash sifati muntazam ravishda IAA metrikalari orqali baholanib boriladi. Masalan, har bir yangi 100 ta teglangan obyekt uchun jamoa Kohen yoki Fleiss kappasini hisoblab ko‘rishi mumkin. Agar kutilmaganda IAA pasaysa, darhol sabablarini o‘rganish lozim. Ba’zi hollarda, qo‘shimcha baholash vazifalari ham joriy etilishi mumkin. Masalan, vaqti-vaqti bilan annotatorlarga sinov testlari berilib, ularning qanchalik to‘g‘ri teglayotgani aniqlanadi. Bu natijalar bo‘yicha har bir annotator uchun sifat ko‘rsatkichi hisoblanib, eng yaxshi va eng yomon ko‘rsatkichga ega annotatorlar aniqlanadi. Natijada umumiy jarayon sifati yuqori darajada ushlab turiladi.

Yuqoridagi chora-tadbirlarning barchasi *ma’lumotlar to‘plamini izchil va ishonchli teglangan holatda shakllantirish*ga qaratilgan. **Yuqori sifatli teglash** – kelgusida quriladigan ML modellari muvaffaqiyatining poydevori hisoblanadi. Sifatli teglangan korpuslardan o‘rgatilgan modellar aniq va ishonchli natijalar beradi, noaniq teglangan ma’lumotlar esa model samaradorligini pasaytiradi.

Annotatsiya sifati va IAA o‘lchovlari o‘rtasidagi bog‘liqlik

Annotatorlararo kelishuv (IAA) teglash sifatining bilvosita ko‘rsatkichi bo‘lib xizmat qiladi. Umuman olganda, teglash sifati oshgan sari IAA ham oshadi va aksincha[6:158]. Ushbu bog‘liqlik, o‘z navbatida, mashinali o‘qitish modellari natijalariga ham ta’sir qiladi. Yuqori sifatli teglangan ma’lumotlar bilan o‘qitilgan

model odatda yuqori aniqlikka erishadi, past sifatli (kelishuvsiz) teglangan ma’lumotlar bilan o‘qitilgan model esa past natija ko‘rsatadi. Quyidagi 1-rasmda IAA o‘lchovlari o‘rtasidagi bog‘liqlik tavsiflangan.



1-rasm. IAA o‘lchovlari o‘rtasidagi bog‘liqlik.

Yuqoridagi 1-rasmdan ko‘rinib turibdiki, IAA oshgan sari modelning samaradorligi ham oshmoqda. Odatda, IAA va modelning aniqligi o‘rtasida ijobiy korrelyatsiya mavjud. Olib borilgan tadqiqot natijalariga ko‘ra bu korrelyatsiya **r≈0.48** atrofida ekanligi ko‘rsatilgan. ya’ni sezilarli darajada musbat bog‘liqlik borligi statistik jihatdan tasdiqlangan[19:275-284]. Agar IAA qiymati yuqori bo‘lsa, model uchun maksimal erishish mumkin bo‘lgan aniqlik darajasi ham baland bo‘lishi zarur. Quyidagi 1-jadvalda *teglash sifati pasayishi IAA metrikalarida va model natijasida qanday aks etishi* keltirilgan.

1-jadval. Teglash sifati turli holatlarida IAA ko‘rsatkichlari va model aniqligi

Annotatsiya sifati holati	**P_o (kelishuv)**	**Kohen κ**	**Fleiss κ**	**Krippendorff α**	**Taxminiy model aniqligi**
A’lo (deyarli mukammal)	1.00	1.00	1.00	1.00	0.99 (99%)
Yaxshi (yuqori)	0.90	0.80	0.78	0.85	0.90 (90%)
O‘rtacha	0.75	0.50	0.48	0.55	0.75 (75%)

Past (kelishmovchilik mavjud)	0.60	0.20	0.18	0.25	0.60 (60%)
Juda yomon (tasodifiy)	0.50	0.00	0.00	0.00	0.50 (50%)

Teglash sifati yuqori bo‘lsa, annotatorlar deyarli barcha holatda kelishadi (P_o va kappa 1 ga teng) va model aniqligi yuqori bo‘ladi. Teglash sifati pasaysa, kelishuv foizi va kappa keskin tushadi, model aniqligi ham shunga mos ravishda yomonlashadi. 50% kelishuv – bu taxminan tasodifiy teg berish darajasi bo‘lib, bunday ma’lumotdan o‘rgangan model ham 50% aniqlikdan oshmaydi.

Muayyan ML modellarini ishlab chiqishda **model aniqligi faqat IAA ga bog‘liq emas**. Modelning arxitekturasi, ma’lumot hajmi va murakkabligi kabi omillar ham muhim rol o‘ynaydi. Ammo teglash sifati past bo‘lsa, hatto eng kuchli model ham yuqori natija bermaydi. Chunki model “shovqinli” ma’lumotni o‘rganadi. Masalan, *teglarda xatolik yoki kelishmovchilik* mavjud bo‘lsa, model noto‘g‘ri qolip (shablon)larni o‘zlashtirib olishi mumkin. Bu esa model sifatiga salbiy ta’sir qiladi. Shu bois, **teglash sifati past bo‘lsa, modelning xatosi ham ortadi**. Bu amalga oshirilgan tajribalarda ko‘p kuzatilgan holatdir.

Biroq, ayrim hollarda *model IAAdan yuqori natija ko‘rsatishi* ham mumkin. Bu, odatda, annotatorlar o‘rtasidagi ba’zi farqlar tizimli bo‘lmagan yoki tasodifiy bo‘lganda yuzaga keladi. Model ko‘p ma’lumot ko‘rib, ikkala annotator qolipini ham qisman o‘rganadi va ba’zi hollarda har ikkisidan yaxshiroq bashorat qilishi mumkin. Masalan, ikki annotator baholagan matnlar to‘plamida annotatorlar 80% holatda kelishgan bo‘lsa ($\kappa \approx 0.6$), model ba’zan 85% to‘g‘ri bashorat qila olishi mumkin. Bu holat ayniqsa annotatorlar *o‘zaro bir-biridan biroz farqli xatolar* qilganda kuzatiladi. Model bu xatolarni *“o‘rtacha qilib”*, asl mantiqni topishi mumkin. Shunday bo‘lsa-da, IAA model aniqligi uchun yuqori chegarani belgilamaydi, lekin ishonchli indikator vazifasini bajaradi. Tajribalar shuni

ko‘rsatadiki, *IAA va model aniqligi odatda bir-biriga mos ravishda o‘zgaradi.* Shuning uchun IAA kelishuvni oshirish ustida ishlash bevosita model natijalarini ham yaxshilaydi.

Teglash jarayonining real misollari

Hozirgi kunda turli tillar bo‘yicha katta hajmli teglangan korpuslar yaratilmoqda. Ingliz tilida Brown Corpus[14], Penn Treebank kabi dastlabki ishlar mashhur bo‘lsa [17:313-330], o‘zbek tilida ham so‘nggi yillarda **teglangan korpuslar** yaratishga qaratilgan tadqiqotlar paydo bo‘lmoqda [9]. Masalan, Elov va boshq. (2024) o‘zbek tilining morfologik[8] va sintaktik[7] teglangan korpusini yaratish ustida ishlaganlar. Quyidagi 2-jadvalda o‘zbek tili korpusi matnlarini CoNLL-2012 kengaytirilgan formatidagi teglash namunalari keltirilgan.

2-jadval. O‘zbek tilidagi koreferensiya korpusidan CoNLL-2012 formatidagi namunalar (kengaytirilgan format)

Hujjat ID	Part №	Gap №	Token №	Token	POS	Parse bitimi	Lemma	NE	Coreference
texts_001	0	1	0	Navoiy	N	NP	Navoiy	PER	(1
texts_001	0	1	1	15-asrda	N	ADVP	15-asrda	DATE	-
texts_001	0	1	2	yashagan	VB	VP	yashamoq	-	-
texts_001	0	1	3	buyuk	JJ	ADJP	buyuk	-	-
texts_001	0	1	4	o‘zbek	N	NP	o‘zbek	NORP	-
texts_001	0	1	5	shoiri	N	NP	shoir	TITLE	1)
texts_001	0	1	6	.	.	O	.	-	-
texts_001	0	2	7	U	P	NP	u	-	1
texts_001	0	2	8	Hirotda	N	PP	Hirot	LOC	(2
texts_001	0	2	9	tug‘ilgan	VB	VP	tug‘ilmoq	-	-
texts_001	0	2	10	.	.	O	.	-	2)
science_015	0	1	0	Suv	NN	NP	suv	-	(3
science_015	0	1	1	0°C	CD	ADVP	0°C	-	-
science_015	0	1	2	da	IN	PP	da	-	-
science_015	0	1	3	muzga	NN	NP	muz	-	-
science_015	0	1	4	aylanadi	VBZ	VP	aylanmoq	-	3)
science_015	0	1	5	.	.	O	.	-	-
science_015	0	2	6	Bu	DT	NP	bu	-	(4
science_015	0	2	7	jarayon	NN	NP	jarayon	-	4)
science_015	0	2	8	juda	RB	ADVP	juda	-	-
science_015	0	2	9	muhimdir	JJ	ADJP	muhim	-	-
science_015	0	2	10	.	.	O	.	-	-
news_100	0	1	0	Toshkent	NNP	NP	Toshkent	LOC	(5

news_100	0	1	1	shahrida	NN	NP	shahar	LOC	5)
news_100	0	1	2	yangi	JJ	NP	yangi	-	(6
news_100	0	1	3	stadion	NN	NP	stadion	FAC	6)
news_100	0	1	4	qurildi	VBN	VP	qurilmoq	-	-
news_100	0	1	5	.	.	O	.	-	-
news_100	0	2	6	Ushbu	DT	NP	ushbu	-	(6
news_100	0	2	7	stadion	NN	NP	stadion	FAC	6)
news_100	0	2	8	50	CD	NP	50	-	-
news_100	0	2	9	ming	CD	NP	ming	-	-
news_100	0	2	10	odam	NN	NP	odam	-	-
news_100	0	2	11	sig‘imga	NN	NP	sig‘im	-	-
news_100	0	2	12	ega	JJ	VP	ega	-	-
news_100	0	2	13	.	.	O	.	-	-

Ushbu jadvalda:

– **POS:** So‘z turkumlari (N – ot, P – olmosh, VB – fe’l va hokazo).

– **Parse bitimi:** Soddalashtirilgan sintaktik struktura ko‘rsatkichi.

– **Lemma:** So‘zning asosiy shakli (lemmalar).

– **NE:** Named Entity, nomlangan obyektlar turi (LOC – joy nomi, PER – shaxs, ORG – tashkilot va hokazo).

– **Coreference:** Koreferensiya zanjirlari identifikatorlari. Ochuvchi qavs “(” yangi zanjirning boshlanishini, yopuvchi qavs “)” esa yakunlanishini bildiradi. Yakka holda eslatma (n) shaklida ifodalanadi.

Ushbu jadvaldagi matnlar CoNLL-2012 standart formatida teglangan tarzda keltirilgan bo‘lib, amaliyotda aynan shunday tarzda koreferensiya aniqlash modellari tomonidan foydalaniladi.

Bunda ToshDO‘TAU kompyuter lingvistikasi va raqamli texnologiyalar kafedrasi mutaxassislari tomonidan maxsus tegger ishlab chiqilib, bir nechta annotatorlarning ishlashini qulaylashtirish yo‘lga qo‘yilgan. Lingvist annotatorlar jamoasi (doktorant, magistrant va talabalar) tuzilib, ular uchun batafsil yo‘riqnoma ishlab chiqilgan va teglash jarayonida doimiy aloqa o‘rnatilgan. Natijada, korpusda har bir so‘zning morfologik va sintaktik bog‘lanishlari teglangan bo‘lib, *bir necha*

bor tekshiruv va tahrirlar orqali konsensusga keltirilgan (adjudikatsiya qilingan). Hozirda ushbu korpusdan foydalanib, o‘zbek tili uchun turli NLP modellar (masalan, *tuzilmaviy tahlil va lug‘at boyligini o‘rganish* modellari) o‘qitilmoqda.

Korpus bo‘yicha annotatorlararo kelishuv yuqori (taxminan 0.85 atrofida kappa) bo‘lib, bu ma’lumot sifatli teglanganini ko‘rsatadi. Ushbu korpus asosida o‘zbek tilida avtomatik NER modellar sinovdan o‘tkazilganda, ularning aniqligi yuqori darajada bo‘lgan. Bu esa, yuqori sifatli teglash jarayoni yakunlangach, modellarning samaradorligi oshishidan dalolat beradi.

Aksincha, teglash sifati past bo‘lgan holatlarda model natijasi qanday yomonlashishini ham olib borilgan ilmiy tadqiqotlar tasdiqlaydi. Masalan, ba’zi o‘tkazilgan tajribalarda annotatorlarning kelishuvi atigi $\boldsymbol{\kappa \approx 0.3}$ bo‘lgan holda, o‘qitilgan modelning aniqligi ham 55–60% dan oshmagan (yaxshisi 0.6 F1 atrofida). Annotatorlarni qayta o‘qitib, yo‘riqnoma yangilab, kelishuvni $\kappa \approx 0.6$ ga yetkazilgach, modelning F1 ko‘rsatkichi ham 80% gacha o‘sgan. Demak, teglardagi *“shovqin”*ni kamaytirish (aniqlikni oshirish) bevosita modelning o‘rganishini yaxshilaydi.

Bundan tashqari, turli vazifalarning tabiiy murakkabligi ham IAA ko‘rsatkichlariga ta’sir qiladi. Masalan, matnlardagi imlo xatolarni aniqlash kabi NLP vazifalarda annotatorlar deyarli bir xil qarorga kelishadi (kelishuv yuqori); POS teglash ham qoidalarga asoslanganligi tufayli odatda $\boldsymbol{\kappa > 0.8}$ darajada bo‘ladi. Ammo matnning mazmunini baholash, sentiment (ijobiy/salbiy) kabi subyektiv komponentli vazifalarda annotatorlar qarorida tafovut ko‘proq bo‘lishi mumkin. Bunday hollarda ham yechim *yo‘riqnomaning imkon qadar aniqlashtirilishi, ko‘proq misollar keltirilishi* va *jamoaviy muhokamalar orqali tushunchalarning bir xilda shakllantirilishidir*.

Yuqoridagi tahlillar va misollar shuni ko‘rsatadiki, teglash jarayonining sifati NLP loyihalarining muvaffaqiyati uchun asos bo‘lib xizmat qiladi. Annotatorlarni

to‘g‘ri tanlash va tayyorlash, aniq yo‘riqnomalar, sifatni nazorat qilishning turli usullarini joriy etish orqali teglarning ishonchliligi maksimal darajaga yetkaziladi. Annotatorlararo kelishuv (IAA) esa ushbu ishonchlilikni o‘lchashga imkon beradigan muhim ko‘rsatkichdir. IAA qiymatlariga qarab, til korpusi **qay darajada ishonchli** ekanligi baholanadi va zarur bo‘lganda, qo‘shimcha chora ko‘riladi (masalan, $\boldsymbol{\kappa < 0.5}$ bo‘lsa, ehtimol butun jarayonni qayta ko‘rib chiqish lozim).

Shuni alohida ta’kidlash joizki, IAA – bu sifat kafolati emas, balki indikator. Agar annotatorlar hammasi bir xil noto‘g‘ri yo‘lni tanlasa (ya’ni, ma’lumotlar izchil lekin noto‘g‘ri teglansa), IAA yuqori bo‘lishi mumkin, lekin teglarning o‘zi noto‘g‘ri bo‘ladi. Shu sabab, annotatorlar kelishib noto‘g‘ri teglamasligi uchun **adjudikatsiya va ekspert nazorati** muhim sanaladi. Ya’ni, ishlab chiqilgan yakuniy “oltin standart”ni real ma’lumotlarga iloji boricha moslashtirish kerak. IAA ko‘rsatkich faqat annotatorlarning bir-biri bilan kelishuvini ko‘rsatadi. Shuning uchun, agar imkon bo‘lsa, ma’lumotning bir qismi bo‘yicha haqiqiy natija mavjud bo‘lishi va annotatorlar shunga erishishga intilishi lozim. Masalan, tibbiy diagnostika ma’lumotlarini teglashda bir guruh shifokorlar qarorini o‘zaro solishtirishdan tashqari, ba’zan **laboratoriya tahlil natijasi** kabi mustaqil mezon bilan solishtirilsa, yanada to‘g‘ri baho beriladi.

Kelishuvni oshirishga qaratilgan harakatlar (treninglar, qo‘shimcha tushuntirishlar, konsensus) aslida teglash sifatini oshiradi va shu bilan **modelning yanada yaxshi o‘rganishiga** zamin yaratadi. Zamonaviy katta til modellari (LLM) ham o‘rgatiladigan ma’lumotlar sifatiga ta’siri katta bo‘lib, agar ma’lumotlarda xatolik mavjud bo‘lsa, model mantiqsiz yoki noto‘g‘ri xulosa chiqaradi.

Teglash jarayonini tashkil qilishda *optimal strategiyani tanlash* ham muhimdir. Masalan, cheklangan resurslarda barcha ma’lumotni ikki marta teglash murakkab bo‘lsa, tanlanma tarzda *ustma-ust teglash* va *audit tekshiruvlarini o‘tkazish* mumkin. Yoki anonim annotatorlar jalb qilinganda, albatta *bir nechta*

annotatorlar va ko‘pchilik ovozi tamoyilidan foydalanish lozim. Aks holda, bir kishining xatosi butun ma’lumotni buzishi mumkin. Turli NLP loyiha talablariga ko‘ra sifatni ta’minlash metodlaridan mos kombinatsiya tanlanadi.

Xulosa

Annotatorlar jamoasi NLP ilovasini poydevorini quruvchi ustunlari hisoblanadi. Ularning roli nafaqat ma’lumotga teg qo‘yish, balki bilvosita ravishda modelning chegara shartlarini belgilashdan iborat. Shu sababli, ularni to‘g‘ri boshqarish, mehnatini muvofiqlashtirish va natijalarini baholab borish – NLP tizimini ishlab chiqish muvaffaqiyati garovi. IAA esa bu jarayonda “termometr” vazifasini o‘taydi, ya’ni bizga teglash “salomatligi” haqida signal beradi. Bugungi kunda ToshDO‘TAU mutaxassislari tomonidan o‘zbek tilida yanada ko‘proq sifatli teglangan korpuslar yaratilishi, ularda teglash jarayoni va kelishuv darajasi batafsil hujjatlashtirilishi bo‘yicha amaliy ishlar olib borilmoqda. Bu esa, o‘z navbatida, NLP modellarining sifatini oshirib, ilmiy tadqiqotlarda yangi marralarga erishishga xizmat qiladi.

Foydalanilgan adabiyotlar ro‘yxati

1. Artstein R. Inter-annotator agreement // Handbook of Linguistic Annotation. – Dordrecht: Springer, 2017. – P. 297–313.

2. Artstein R., Poesio M. Inter-coder agreement for computational linguistics // Computational Linguistics. – 2008. – Vol. 34, № 4. – P. 555–596.

3. Baledent A., Mathet Y., Widlöcher A., Couronne C., Manguin J.L. Validity, agreement, consensuality and annotated data quality // Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022). – Marseille, France, 2022. – P. 2940–2948.

4. Boguslav M., Cohen K.B. Inter-annotator agreement and the upper limit on machine performance: evidence from biomedical natural language processing //

MEDINFO 2017: Precision Healthcare through Informatics. – Amsterdam: IOS Press, 2017. – P. 298–302.

5. Cohen J. A coefficient of agreement for nominal scales // Educational and Psychological Measurement. – 1960. – Vol. 20, № 1. – P. 37–46.

6. Decker Z. Collaboration in Adjudication // Canon Law Society of America Proceedings. – 2016. – Vol. 78. – P. 158–170.

7. Elov B., Abdullayeva O. O‘zbek tili korpusini sintaktik teglash masalasi // Computer Linguistics: Problems, Solutions, Prospects. – 2024. – Vol. 1, № 1.

8. Elov B., Xudayberganov N. O‘zbek tili korpusi matnlarini POS teglash usullari // Computer Linguistics: Problems, Solutions, Prospects. – 2024. – Vol. 1, № 1.

9. Elov B., Xusainova Z. Til korpuslarini lingvistik teglash bosqichlari // Computer Linguistics: Problems, Solutions, Prospects. – 2024. – Vol. 1, № 1

10. https://datatab.net/tutorial/fleiss-kappa

11. https://medium.com/@jorgecp/the-adjudication-process-in-collaborative-annotation-61623c46b700

12. https://www.surgehq.ai/blog/inter-rater-reliability-metrics-an-introduction-to-krippendorffs-alpha

13. Islamaj R., Kwon D., Kim S., Lu Z. TeamTat: a collaborative text annotation tool // Nucleic Acids Research. – 2020. – Vol. 48, № W1. – P. W5–W11.

14. Kholkovskaia O. Role of the Brown Corpus in the History of Corpus Linguistics // Poster Proceedings. – Prague, 2017.

15. Krippendorff K. Content Analysis: An Introduction to Its Methodology. – 4th ed. – Thousand Oaks: Sage Publications, 2018. – 472 p.

16. Landis J.R., Koch G.G. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers // Biometrics. – 1977. – Vol. 33, № 2. – P. 363–374.

17. Marcus M., Santorini B., Marcinkiewicz M.A. Building a large annotated corpus of English: The Penn Treebank // Computational Linguistics. – 1993. – Vol. 19, № 2. – P. 313–330.

18. Randolph J.J. Free-Marginal Multirater Kappa (Multirater K [free]): An Alternative to Fleiss’ Fixed-Marginal Multirater Kappa. – Joensuu: University of Joensuu, 2005.

19. Richie R., Grover S., Tsui F.R. Inter-annotator agreement is not the ceiling of machine learning performance: Evidence from a comprehensive set of simulations // Proceedings of the 21st Workshop on Biomedical Language Processing. – Dublin, Ireland, 2022. – P. 275–284.

20. Scott W.A. Reliability of content analysis: The case of nominal scale coding // Public Opinion Quarterly. – 1955. – Vol. 19, № 3. – P. 321–325.

21. Skjærholt A. A chance-corrected measure of inter-annotator agreement for syntax // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL). – Baltimore, Maryland, 2014. – Vol. 1. – P. 934–944.