SENTIMENT TAHLIL UCHUN KATTA HAJMDAGI DATASETNI YARATISH BOSQICHLARI

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

Abdullayev Abdulla Quranbayevich,
NTM ta’limni kredit tizimini boshqarish bo‘lim boshlig‘i
abdulla_abdullayev9270@mail.ru
Urganch innovatsion university

Annotatsiya. O‘zbek tilida sentiment tahlilni rivojlantirishning asosiy to‘siqlaridan biri – lingvistik xususiyatlar va yetarli hajmdagi sifatli ma’lumotlar to‘plamining yo‘qligidir. Ushbu maqolada sentiment tahlil uchun katta hajmdagi datasetni yaratish jarayoni uch bosqichli tizim asosida tahlil qilinadi. Avvalo, birinchi bosqichda yaratiladigan datasetning maqsadi, sifat mezonlari va foydalanish sohasi aniqlanib, ilmiy talablar ishlab chiqiladi. Ikkinchi bosqichda, turli axborot manbalari – ijtimoiy tarmoqlar, forumlar, bloglar, onlayn sharhlar va yangilik portallaridan xom ma’lumotlar yig‘ilib, ularning dolzarbligi va ishonchliligi tahlil qilinadi. Uchinchi bosqichda esa, yig‘ilgan ma’lumotlar tozalash, normalizatsiya, tokenizatsiya va lemmatizatsiya kabi ishlov berish operatsiyalari yordamida mos va strukturallashgan dataset shakllantiriladi. Shuningdek, annotatsiya va balanslash jarayonlari orqali har bir sentiment kategoriyasining teng va to‘liq aks etishi ta’minlanadi. Ushbu tizimli yondashuv ma’lumotlarning lingvistik, semantik va statistik jihatlarini hisobga olib, sentiment tahlil modellari uchun yuqori aniqlikdagi korpus yaratishga xizmat qiladi. Tadqiqot natijalari yaratilgan datasetning sifatli va ishonchli ekanligini, shuningdek, sentiment tahlil algoritmlarining samaradorligini oshirishga hissa qo‘shishini ko‘rsatadi. Ushbu tadqiqot sentiment tahlil uchun katta hajmdagi datasetni yaratishning ilmiy asoslangan bosqichlarini taqdim etib, sohada innovatsion yechimlarni ilgari suradi.

Kalit so‘zlar: *Sentiment tahlil, NLP, dataset yaratish, ma’lumotlarni tayyorlash, katta hajmdagi dataset, lingvistik annotatsiya, Inter-annotator kelishuv.*

Kirish

Tabiiy tilni qayta ishlash sohasida sentiment tahlilning yuqori aniqlik va umumlashuv qobiliyatiga erishishda **katta hajmdagi, yuqori sifatli** datasetning yaratilishi nihoyatda muhim ahamiyatga ega. Avvalo, turli axborot manbalaridan – ijtimoiy tarmoqlar, forumlar, onlayn sharhlar va yangilik portallaridan – xom ma’lumotlar yig‘ilib, korpusning asosiy poydevori shakllantiriladi. Keyin, yig‘ilgan

ma’lumotlar tozalash jarayoni orqali HTML teglar, maxsus belgilar, ortiqcha raqamlar va keraksiz so‘zlardan xoli holga keltiriladi, bu esa keyingi bosqichlar uchun ishonchli ma’lumot zaxirasini ta’minlaydi. Shu bilan birga, tokenizatsiya, normalizatsiya va lemmatizatsiya kabi boshlang‘ich ishlov berish operatsiyalari qo‘llanilib, matnlarning lingvistik va semantik tuzilishi soddalashtiradi. Annotatsiya bosqichida esa, har bir matn sentiment bo‘yicha ijobiy, salbiy yoki neytral deb belgilanadi, bu usul yarim avtomatik va qo‘lda amalga oshirilgan annotatsiya tizimlari yordamida bajariladi. Ma’lumotlar sifatini yanada oshirish maqsadida, dataset balanslash bosqichida turli sentiment kategoriyalarining teng taqsimlanishi statistik metodlar orqali ta’minlanadi. So‘nggi bosqichda, annotatsiyalangan va balanslangan ma’lumotlar trening, test va validatsiya to‘plamlariga ajratilib, modelni o‘rgatish uchun mustahkam asos yaratiladi. Modelni o‘rgatish jarayonida, katta hajmdagi ma’lumotlar yordamida algoritm treningi amalga oshirilib, gradient tushishi, kross-validatsiya va boshqa optimallashtirish strategiyalari orqali model parametrlarida yaxshilanishlar kiritiladi. Keyinchalik, modelni baholash bosqichida aniqlik, precision, recall va F1-score kabi metrikalar asosida eksperimental sinovlar o‘tkazilib, yaratilgan modelning amaliy qo‘llanilishidagi samaradorligi aniqlanadi. Natijada, yaratilgan dataset va o‘rgatilgan model real sharoitlarda sentiment tahlilini avtomatlashtirish imkonini yaratib, yuqori aniqlik va barqarorlikka erishishni ta’minlaydi. Shuningdek, modelni joriy qilish bosqichi orqali tizimga integratsiyalashgan yechim real vaqtda sentiment tahlilini amalga oshirishga imkon yaratadi. Ushbu tizimli yondashuv, ma’lumotlarni tayyorlash, tozalash, annotatsiya, balanslash, modelni o‘rgatish, baholash va joriy qilish bosqichlarining o‘zaro bog‘liqligini ishonchli tarzda namoyish etadi.

Yuqorida ta’kidlanganidek, zamonaviy sentiment tahlil tizimlarining samaradorligi, qaysi algoritmdan foydalanilmasin, asosan uning o‘qitiladigan ma’lumotlar **bazasining hajmi, sifati** va **representativligiga** bog‘liq. O‘zbek tilidagi matnlar uchun katta hajmdagi ma’lumotlar to‘plamini yaratish sohasi nisbatan yangi bo‘lib, tilning agglutinativ tuzilishi, kodli almashish (rus/ingliz tillaridan so‘zlar qo‘shilishi) kabi lingvistik xususiyatlar tufayli qiyinchiliklar tug‘diradi.

Xalqaro tajribalar (Hovy va Lavid, 2010[1]; Cambria, 2016[2]) shuni ko‘rsatadiki, ma’lumotlar to‘plamini shakllantirishda muvozanatli tanlov, to‘g‘ri annotatsiya va turli kontekstlarni qamrab olish algoritmlarning adolatli ishlashi uchun hal qiluvchi ahamiyatga ega. O‘zbek tiliga oid mavjud korpuslar (masalan, “UzbCorpus”) ko‘pincha cheklangan hajmga ega yoki ma’lum bir soha (ijtimoiy media, yangiliklar) bilan chegaralangan, bu esa ko‘p qirrali sentiment tahlilni qiyinlashtiradi[3]. Ma’lumotlarni yig‘ish bosqichida manbalarning xilma-xilligi (forumlar, ijtimoiy tarmoqlar, rasmiy hujjatlar) va demografik tarqalishi (yosh, mintaqaviy dialektlar) to‘plamning universalligini ta’minlashda muhim rol

o‘ynaydi. Elov va boshq. (2024) ta’kidlaganidek, o‘zbek tilida morfologik normalizatsiya (lemmatizatsiya, stemming) va tokenizatsiyani to‘g‘ri amalga oshirish matnlarni kompyuter tushunishi uchun moslashtirishning asosiy sharti hisoblanadi[4]. Tovush yozuvi, imlo xatolari va qisqartmalar kabilarni tozalash (Kuryozov va boshq., 2019) esa ma’lumotlarni vektorlashtirish jarayonidagi shovqinlarni kamaytiradi[5]. Annotatsiya bosqichida mahalliy til mutaxassislari ishtiroki (Mengliyev va boshq., 2023) emotsional nuanslarni to‘g‘ri belgilash imkonini beradi[6]. Bundan tashqari, turli annotatorlar orasidagi kelishuv darajasini (inter-annotator agreement) o‘lchash (Bobicev va boshq., 2017) ma’lumotlar sifatini baholashda muhim metrikadir[7]. Ma’lumotlarni tasodifiy bo‘linishi (train-test split), sinov to‘plamlarining diversifikatsiyasi (Hamdamov va boqsh., 2024) modelning overfittingdan himoyalanishiga yordam beradi[8]. Yana bir muammo – ma’lumotlar to‘plamini dinamik yangilab borish, chunki til tirik tizim bo‘lib, yangi so‘zlar va iboralar doimiy paydo bo‘ladi. Xalqaro standartlarga moslab ma’lumotlarni ochiq manbali ravishda taqdim etish (masalan, Kaggle, Hugging Face) ham O‘zbek tilidagi NLP hamjamiyatining rivojiga hissa qo‘shadi. Xulosa qilib, keng ko‘lamli, sifatli va diversifikatsiyalangan ma’lumotlar to‘plamini yaratish – bu nafaqat sentiment tahlil, balki boshqa NLP vazifalari uchun ham o‘zbek tilining raqamli resurslarini boyitishga qaratilgan muhim ilmiy-amaliy harakatdir.

Katta hajmdagi datasetni yaratishda quyidagi bosqichlarni amalga oshirish lozim:

1. Dataset uchun asosiy talablarni ishlab chiqish.

Ushbu bosqichda, avvalo, yaratiladigan datasetning maqsadi, qo‘llanilishi sohalari va ilmiy talablarini aniq belgilash lozim. Tadqiqot kontekstida, dataset tarkibida kutilayotgan semantik, sintaktik va lingvistik xususiyatlar, shuningdek, sentimentning aniqligi va granulyatsiyasi (masalan, ijobiy, salbiy, neytral yoki nozik hissiyot darajalari) aniqlanadi. Shuningdek, ma’lumotlarning to‘liqligi, dolzarbligi, aniqligi va moslashuvchanligi kabi sifat mezonlari belgilab olinadi. Ushbu talablar asosida, keyingi bosqichlarda ishlatiladigan annotatsiya protokollari, sinov metrikalari va modelning o‘lchov ko‘rsatkichlari ham ishlab chiqiladi, bu esa yaratilayotgan datasetning ilmiy jihatdan mustahkam poydevorini tashkil etadi.

Sentiment tahlil uchun dataset yaratishning dastlabki bosqichi – maqsadga mos funksional va texnik talablarni aniqlashdir. Bu bosqichda quyidagilar ko‘rib chiqiladi: **Maqsadni aniqlash:** Dataset sentiment tahlilning qaysi turi (binary: ijobiy/salbiy, yoki ko‘p toifali: neytral, aralash, emotsional darajalar) uchun mo‘ljallangan? **Lingvistik talablar:** O‘zbek tilining agglutinativligi, dialektlar (Farg‘ona, Toshkent shevalari), kodli almashish (rus/ingliz tillaridan so‘zlar) kabi xususiyatlar hisobga olinadi. **Hajm va diversifikatsiya:** Kamida 100000 ta annotatsiyalangan matn, turli janrlar (ijtimoiy media, rasmiy matnlar, nutq) va

demografik guruhlar (yosh, jins) qamrovini ta’minlash. **Etika va qonuniy talablar:** Shaxsiy ma’lumotlarni olib tashlash (GDPR), mualliflik huquqi (CC BY litsenziya) va anonymizatsiya. **Annotatsiya protokoli:** Inter-annotator kelishuv (Cohen’s kappa > 0.8) va mahalliy ekspertlarni jalb qilish. Datasetni shakllantirishda quyidagi talablarga rioya qilish kerak:

- **Gaplar**: O‘zbek tilidagi real gaplardan iborat bo‘lishi lozim.
- **Sentiment teglari**: Datasetdagi har bir gap "ijobiy", "salbiy", yoki "neytral" kabi teglar vositasida teglanishi lozim.
- **Dataset hajmi**: Katta hajmdagi dataset yaxshi natijalar olishga yordam beradi. "Ijobiy", "salbiy", yoki "neytral" kabi har bir kategoriyadan kamida **1000-5000 ta gap** bo‘lishi maqsadga muvofiq.

2. Dataset manbalari.

Ushbu bosqichda, kerakli ma’lumotlarni yig‘ish uchun foydalaniladigan axborot manbalari aniqlanadi va ularning ilmiy asosda tahlil qilinishi amalga oshiriladi. Tadqiqot doirasida, ijtimoiy tarmoqlar, forumlar, bloglar, onlayn sharhlar va yangilik portallari kabi turli manbalar orqali xom ma’lumotlar to‘planadi. Har bir manbaning ishonchliligi, dolzarbligi va ma’lumotlarning strukturaliligi baholanib, ularning datasetga qo‘shilishi mumkinligi aniqlanadi. Shuningdek, manbalar orasida lingvistik va madaniy farqlarni inobatga olish zarur, chunki bu sentiment tahlilning yakuniy natijalariga sezilarli ta’sir ko‘rsatadi. Manbalarni tanlash jarayonida, mavjud ilmiy adabiyotlar, statistik tahlillar va avvalgi tajribalar asosida optimal manba to‘plami shakllantiriladi. Matnlar **turli manbalardan**, **turli mavzulardan** va **turli kontekstlardan** bo‘lishi muhim. Datasetni quyidagi manbalardan shakllantirish mumkin:

- **Ijtimoiy tarmoqlar**: Telegram, Twitter, Facebook, yoki Instagram ijtimoiy tarmoqlardagi o‘zbek tilidagi postlar yoki izohlar (odamlarning fikrlari va sharhlari).
- **Sharhlar**: Mahsulot sharhlari (masalan, Oson, Uzum, ZoodMall kabi onlayn do‘konlar yoki restoran sharhlari) va kino sharhlarini shakllantirish (ijobiy va salbiy fikrlarni ajratish uchun).
- **Yangilik saytlari**: kun.uz, daryo.uz, gazeta.uz kabi onlayn nashrlar (neytral matnlar uchun ideal manba).
- **Bloglar va maqolalar**: O‘zbek tilidagi bloglardan yoki yangiliklar saytlaridan gaplarni to‘plash zarur (odamlar haqqoniy fikr bildiradigan manba).
- **YouTube izohlari**: Video sharhlari orqali real his-tuyg‘ularni aniqlash.
- **Sun’iy yozilgan gaplar**: Agar haqiqiy ma’lumotlar yetarli bo‘lmasa oddiy gaplar yozib, ularga sentiment teg berish kerak.

3. Datasetni shakllantirish.

Ushbu bosqichda, yig‘ilgan xom ma’lumotlar asosida datasetni strukturaviy, semantik va statistik jihatdan boyitish jarayoni amalga oshiriladi. Avvalo, xom ma’lumotlar tozalash, normalizatsiya, tokenizatsiya va lemmatizatsiya kabi boshlang‘ich ishlov berish jarayonlari orqali mos va yagona formatga keltiriladi. Keyin, annotatsiya bosqichida, har bir matn sentiment bo‘yicha (ijobiy, salbiy, neytral yoki yanada nozik darajalar) qo‘lda yoki yarim avtomatik tarzda belgilab qo‘yiladi. Ushbu jarayon uchun inter-annotator agreement kabi sifat nazorati mexanizmlari qo‘llaniladi. Shuningdek, datasetni balanslash bosqichi orqali, har bir sentiment kategoriyasining teng taqsimlanishini ta’minlash uchun statistik metodlar, masalan, over-sampling yoki under-sampling usullari qo‘llaniladi. Natijada, tayyor dataset trening, test va validatsiya to‘plamlariga ajratilib, keyingi modellashtirish bosqichlariga tayyor holda ilmiy asosda shakllantiriladi. Datasetni shakllantirishda qo‘lda yoki avtomatik vositalardan foydalanish mumkin. Bu bosqichda quyidagi amallar bajariladi:

- **Ma’lumotlarni yig‘ish**: Saytlar va platformalardan ma’lumotlarni yig‘ish uchun **web scraping** usulidan foydalanish mumkin (BeautifulSoup, Selenium, Twitter API, YouTube API).
- **Ma’lumotlarni tozalash**: Avtomatik top`langan matnlarni tozalash lozim. Bu jarayonda matnlardan noto‘g‘ri belgilar, html teglar, va ortiqcha so‘zlar olib tashlanadi.
- **Ma’lumotlarni teglash**: Har bir gap uchun sentimentni aniqlash va uni teglash. Bu jarayon ham qo‘lda yoki avtomatik tarzda amalga oshiriladi. Odatda kichik dataset uchun matnlar qo‘lda teglanadi. Kattaroq hajmdagi dataset matnlari dastlab avtomatik teglanadi va keyin qo‘lda qayta ko‘rib chiqiladi.
- **Matnlarning xilma-xilligi. Mavzu bo‘yicha muvozanatni ta`minlash uchun** matnlar turli sohalarni qamrab olishi kerak. Masalan, mahsulot sharhlari – elektronika, kiyim-kechak, kitoblar; xizmatlar – ta’lim, mehmonxonalar, restoranlar; siyosiy va ijtimoiy sharhlar.
- **Tasnif bo‘yicha ajratish (Sentiment bo‘yicha muvozanat)**: Har bir kategoriya uchun ijobiy, salbiy va neytral gaplarni teng (tekis) taqsimlash lozim. Masalan quyidagicha tasniflash usulidan foydalanish mumkin:
  - *juda ijobiy (20-25%)*
  - *ijobiy (20-25%)*
  - *neytral (20-25%)*
  - *salbiy (20-25%)*
  - *juda salbiy (5-10%)*
- **Matn uzunligi:** Datasetni shakllantirishda matn uzuligi ham muhim ahamiyat kasb etadi. Masalan quyidagicha tasniflash usulidan foydalanish mumkin:

- *qisqa matnlar (1-2 gap) (30-35%)*
- *o‘rta matnlar (3-5 gap) (30-35%)*
- *uzun matnlar (5+ gap) (30-35%)*

- **Saqlash**: Datasetni .csv yoki .json formatida yoki ma`lumotlar bazasi jadvallarida saqlash lozim.

Ushbu tizimli yondashuv orqali yaratilgan dataset, sentiment tahlilning yuqori aniqligi va ishonchliligini ta’minlashda muhim omil bo‘lib, kelgusidagi ilmiy tadqiqotlar va amaliy ilovalar uchun mustahkam metodologik poydevor sifatida xizmat qiladi. Datasetni ochiq manbali platformalar (Hugging Face, Zenodo) ga joylashtirish va litsenziya (CC BY-SA 4.0) bilan ta’minlash mumkin.

Dataset namunasi

1-jadval. O‘zbek tili gaplarini teglash formati

№	Gap	Sentiment
.	Bu kitob juda qiziqarli ekan.	ijobiy
.	Xizmat sifati meni qoniqtirmadi.	salbiy
.	Bugungi ob-havo juda yaxshi edi.	neytral
.	Bu film vaqtni behuda ketkazish edi.	salbiy
.	Ovqat juda mazali edi.	ijobiy
.	Pulimni qaytarishmadi, xizmat yomon!	salbiy
.	Yetkazib berish kechikdi, lekin mahsulot yaxshi.	salbiy

Xulosa

O‘zbek tilida sentiment tahlilni rivojlantirishda katta hajmdagi datasetni yaratishning asosiy bosqichlari (talablarni belgilash, manbalarni tanlash, shakllantirish) tilning lingvistik xususiyatlari va amaliy muammolarni hisobga olgan holda ishlab chiqilishi kerak. Tadqiqot shuni ko‘rsatdiki, ma’lumotlarni yig‘ishda ijtimoiy tarmoqlar, mavjud korpuslar va demografik diversifikatsiya muhim rol o‘ynaydi. Datasetni shakllantirish bosqichida morfologik normalizatsiya (lemmatizatsiya, stemming) va kodli almashishni hisobga oladigan tozalash algoritmlari samaradorlikni oshirishga yordam beradi. Annotatsiya jarayonida mahalliy ekspertlar ishtiroki va inter-annotator kelishuv (Cohen’s kappa >0.8) dataset sifatini ta’minlaydi. Yaratilgan datasetning muvaffaqiyati, shuningdek, sinov to‘plamlari orqali tekshirilgan statistik muvozanat va dinamik yangilash mexanizmlariga bog‘liq. Natijada, o‘zbek tiliga mo‘ljallangan keng ko‘lamli dataset nafaqat sentiment tahlil, balki chat-botlar, avtomatik tarjima kabi NLP vazifalari uchun asos bo‘lishi mumkin. Biroq, dialektlar, yangi so‘zlar va semantik noaniqliklar kabi muammolar datasetni doimiy yangilab borishni talab qiladi. Kelajakda ko‘p tilli transformer modellarni (masalan, BERT) o‘zbek tiliga

moslashtirish va ochiq manbali hamjamiyatni rivojlantirish sohaga yangi imkoniyatlar yaratadi. Xulosa qilib, ushbu ish o‘zbek tilining raqamli resurslarini boyitishga qaratilgan muhim qadam hisoblanadi.

Foydalanilgan adabiyotlar:

1. Hovy, E., & Lavid, J. (2010). Towards a ‘science’of corpus annotation: a new methodological challenge for corpus linguistics. International journal of translation, 22(1), 13-36.
2. Cambria, E., Hazarika, D., Poria, S., Hussain, A., & Subramanyam, R. B. V. (2018). Benchmarking multimodal sentiment analysis. In Computational Linguistics and Intelligent Text Processing: 18th International Conference, CICLing 2017, Budapest, Hungary, April 17–23, 2017, Revised Selected Papers, Part II 18 (pp. 166-179). Springer International Publishing.
3. Elov, B., & Xusainova, Z. (2024). Til korpuslarini lingvistik teglash bosqichlari. Computer Linguistics: problems, solutions, prospects, 1(1).
4. Boltayevich, E. B., Yuldashevna, X. Z., Mamurjonovna, U. S., Ermamatovich, N. S., qizi, A. Sh. A., & Shavkatjon, M. (2024, October). Algorithms for Parsing Roots and Stems of Words in Uzbek Language. In 2024 9th International Conference on Computer Science and Engineering (UBMK) (pp. 126-130). IEEE.
5. Kuriyozov, E., Matlatipov, S., Alonso, M. A., & Gómez-Rodríguez, C. (2019, May). Construction and evaluation of sentiment datasets for low-resource languages: The case of Uzbek. In Language and Technology Conference (pp. 232-243). Cham: Springer International Publishing.
6. Mengliev, D. B., Akhmedov, E. Y., Barakhnin, V. B., Hakimov, Z. A., & Alloyorov, O. M. (2023, November). Utilizing lexicographic resources for sentiment classification in Uzbek language. In 2023 IEEE XVI International Scientific and Technical Conference Actual Problems of Electronic Instrument Engineering (APEIE) (pp. 1720-1724). IEEE.
7. Bobicev, V., & Sokolova, M. (2017). Inter-annotator agreement in sentiment analysis: machine learning perspective. In Recent Advances in Natural Language Processing (pp. 97-102).
8. Hamdamov, O‘., Elov, B., & Alayev, R. (2024). Ma’lumotlar to‘plamini o‘qitish, baholash va test to‘plamlariga ajratish usullari. Digital transformation and artificial intelligence, 2(6), 107-116.