O‘ZBEK TILIDAGI ELEKTRON MATNLARNING SEMANTIK XUSUSIYATLARINI CHUQUR IFODALASHNING KO‘P BOSQICHLI KONSEPTUAL MODELI

Karaxanova Shirin Azizjon qizi

Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti, Toshkent, O‘zbekiston
E-mail: shirinak23@gmail.com

K E Y W O R D S

o‘zbek tili, tabiiy tilni qayta ishlash, tematik tasniflash, sentiment tahlil, semantik tahlil, CNN, BiLSTM, gibrid neyron tarmoq, BPE, matnni normallashtirish

A B S T R A C T

Maqolada o‘zbek tilidagi elektron matnlarning semantik xususiyatlarini chuqur ifodalash maqsadida tematik va sentiment tahlil vazifalarini ketma-ket hamda o‘zaro uzviy bog‘liq holda bajaruvchi ko‘p bosqichli konseptual model taklif etilgan. Modelning matematik asosi matnni uzluksiz semantik fazodagi chiziqli bo‘lmagan trayektoriya sifatida qarashga, semantik fazoni esa tematik va sentiment ko‘pxilliklarining ko‘paytmasi ko‘rinishida ifodalashga tayanadi. Modelni amalga oshirish uchun CNN va BiLSTM arxitekturalarini birlashtirgan gibrid neyron tarmoq ishlab chiqilgan: CNN lokal n-gramm xususiyatlarini ajratadi, BiLSTM ikki yo‘nalishli kontekstni modellashtiradi, yakuniy qatlamda sentiment sinflari uchun Softmax, ko‘p yorliqli tematik tasniflash uchun esa sigmoid funksiyasi qo‘llaniladi. Lotin, kirill va yangi lotin alifbosidagi yozuvlarni yagona shaklga keltiruvchi idempotent normallashtirish operatori ishlab chiqilgan. Tajribalarda normallashtirishsiz berilgan kirill yozuvidagi matnlarda aniqlik 53–80% ga pasayishi, operator qo‘llanilganda esa sifat yo‘qotilmasligi aniqlandi; BPE tokenizatsiyasi sharoitida stop-so‘zlarni olib tashlash va lemmatizatsiya aniqlikni sezilarli oshirmadi. Taklif etilgan yondashuv klassik TF-IDF usulidan yuqori, Transformer modellari bilan teng F1-score hamda yozuv tizimi o‘zgarishiga nisbatan barqarorlikni namoyish etadi.

KIRISH

Raqamli texnologiyalarning rivojlanishi natijasida o‘zbek tilidagi elektron matnlar hajmi ijtimoiy tarmoqlar, ommaviy axborot vositalari, elektron murojaatlar va davlat organlarining raqamli hujjatlari hisobiga tez o‘sib bormoqda. Bunday katta hajmdagi tuzilmalashtirilmagan ma’lumotlar oqimini qo‘lda qayta ishlash imkonsiz bo‘lgani sababli, matnning mavzusini (tematik sinfini) hamda unda ifodalangan munosabatni (sentimentni) avtomatik aniqlash tizimlari jamoatchilik fikrini monitoring qilish, murojaatlarni yo‘naltirish va boshqaruv qarorlarini qo‘llab-quvvatlash uchun muhim ahamiyat kasb etmoqda [6, 7]. So‘nggi yillarda bu vazifalar chuqur o‘rganish va Transformer arxitekturalari asosida sezilarli rivojlandi [1, 2], biroq erishilgan yutuqlar asosan resurslari boy tillarga tegishli.

O‘zbek tili tabiiy tilni qayta ishlash nuqtai nazaridan resurslari cheklangan tillar qatoriga kiradi. Uning agglyutinativ tabiati, ya’ni so‘z o‘zagiga ketma-ket qo‘shimchalar qo‘shilishi orqali juda ko‘p so‘z shakllarining hosil bo‘lishi lug‘at hajmining kengayishiga va ma’lumotlarning siyrakligiga (data sparsity) olib keladi [8, 25]. So‘z tartibining erkinligi va polisemiya matndagi yashirin mazmuniy munosabatlarni an’anaviy statistik usullar bilan qamrab olishni yanada qiyinlashtiradi. Bundan tashqari, raqamli makonda o‘zbekcha matnlar lotin va kirill yozuvlarida hamda yangi lotin alifbosida uchraydi, bir xil harf yoki tutuq belgisi esa turli Unicode belgilari bilan yozilishi mumkin [19]. Natijada mazmunan bir xil matn modelga turlicha ko‘rinishda kelib, tasniflash aniqligi keskin pasayadi.

Matnni tasniflashning klassik yondashuvlari so‘zlar chastotasiga asoslangan TF-IDF tasviri [5] va an’anaviy mashinali o‘qitish algoritmlariga tayanadi; o‘zbekcha yangiliklar maqolalarini ko‘p sinfli tasniflashda ham shunday usullar qo‘llangan [12]. Bunday tasvirlar so‘zlar tartibi va kontekstni hisobga olmaydi hamda morfologik jihatdan boy tillarda lug‘at hajmining o‘sishidan zarar ko‘radi.

(cc) BY

So‘z embeddinglari (Word2Vec) [3] so‘zlarni zich vektor fazosida tasvirlash imkonini berdi, fastText esa subso‘z axborotidan foydalanib morfologik jihatdan boy tillarda tasvir sifatini yaxshiladi [8]. Lug‘atdan tashqari so‘zlar muammosini hal qilishda BPE [25] va SentencePiece [10] kabi subso‘z tokenizatsiyasi algoritmlari keng qo‘llanilmoqda.

Neyron tarmoq arxitekturalari orasida konvolyutsion tarmoqlar matndagi lokal n-gramm xususiyatlarini samarali ajratadi [15, 17], LSTM [4] va uning ikki yo‘nalishli varianti BiLSTM [13] esa ketma-ketlikdagi uzoq masofali bog‘liqliklarni modellashtiradi. Ikkala yondashuvni birlashtiruvchi C-LSTM [14] va rekurrent konvolyutsion tarmoqlar [16] matnni tasniflashda raqobatbardosh va ko‘pincha yuqori natijalarni namoyish etgan.

Transformer arxitekturasi [1] hamda BERT [2] va XLM-R [9] kabi oldindan o‘qitilgan modellar matnni tasniflashda yuqori aniqlikka erishdi. O‘zbek tili uchun ham BERT turidagi oldindan o‘qitilgan modellar, jumladan Tahrirchi-BERT [23] mavjud. Biroq bunday modellar katta hisoblash resurslarini talab qiladi, bu esa real vaqt rejimidagi monitoring tizimlari uchun qo‘shimcha cheklov yaratadi.

O‘zbek tili bo‘yicha amalga oshirilgan tadqiqotlar sentiment ma’lumotlar to‘plamlarini yaratish va baholash [11], emojilarning fikr tasniflashiga ta’sirini o‘rganish [18], yangiliklar matnlarini tematik tasniflash [12] hamda lotin va kirill alifbolari o‘rtasida avtomatik transliteratsiya vositalarini ishlab chiqish [19] yo‘nalishlarida to‘plangan.

Mavjud ishlarning tahlili quyidagi bo‘shliqlarni ko‘rsatadi: birinchidan, tematik va sentiment tahlil, odatda, mustaqil vazifalar sifatida hal etiladi va ular o‘rtasidagi bog‘liqlik hisobga olinmaydi; ikkinchidan, matn ko‘pincha faqat bitta tematik sinfga tegishli deb qaraladi, vaholanki real matnlar bir nechta mavzuga tegishli bo‘lishi mumkin; uchinchidan, yozuv tizimi (lotin, kirill, yangi alifbo) o‘zgarishiga modelning barqarorligi kam o‘rganilgan.

Tadqiqotning maqsadi o‘zbek tilidagi elektron matnlarning semantik xususiyatlarini chuqur ifodalash uchun tematik va sentiment tahlilni birlashtiruvchi ko‘p bosqichli konseptual model hamda uning asosida gibrid CNN+BiLSTM neyron tarmoq algoritmini ishlab chiqishdan iborat. Maqsadga erishish uchun quyidagi vazifalar hal etildi:

1. matnning semantik fazosi va ko‘p bosqichli strategiyaning matematik ifodasini shakllantirish;
2. tizimning kirish va chiqish parametrlarini aniqlash;
3. CNN va BiLSTM qatlamlariga asoslangan gibrid modelning funksional arxitekturasini ishlab chiqish;
4. turli yozuv tizimlarini yagona shaklga keltiruvchi normallashtirish operatorini va dastlabki ishlov berish bosqichlarini formallashtirish;
5. modelni o‘rgatish va baholash metodikasini belgilash hamda dastlabki ishlov berish bosqichlarining ta’sirini tajribada tekshirish.

Tadqiqotning ilmiy yangiligi shundan iboratki, tematik va sentiment tahlil vazifalari ketma-ket va o‘zaro uzviy bog‘liq holda bajariladi, matnning bir nechta tematik sinfga tegishliligi hisobga olinadi hamda yozuv tizimi o‘zgarishiga barqaror normallashtirish operatori model tarkibiga kiritiladi.

Maqolaning qolgan qismi quyidagicha tuzilgan: 2-bo‘limda taklif etilayotgan konseptual model, gibrid arxitektura, dastlabki ishlov berish va o‘rgatish jarayonlari bayon etilgan; 3-bo‘limda tajriba sharoiti, natijalar va ularning muhokamasi keltirilgan; 4-bo‘limda xulosalar va kelgusi tadqiqot yo‘nalishlari berilgan.

TADQIQOT METODOLOGIYASI

Elektron matnli ma’lumotlarni intellektual tahlil qilish va ularning semantik mazmunini chuqur ifodalash hozirgi vaqtda kompyuter lingvistikasi va sun’iy intellekt yo‘nalishlarining, xususan, o‘zbek tili tabiiy tilini qayta ishlash (NLP) tizimlarining eng dolzarb muammolaridan biri hisoblanadi. O‘zbek tilining agglyutinativ tabiati, boy morfologik strukturasi, so‘z tartibining erkinligi hamda bir so‘zning ko‘p ma’noliligi (polisemiya) kabi xususiyatlari matn tarkibidagi yashirin mazmuniy munosabatlarni an’anaviy statistik usullar bilan qamrab olish imkonini bermaydi. Shu bois, ushbu tadqiqotda o‘zbek

(cc) BY

tilidagi matnlarni kompleks semantik tahlil qilishga mo‘ljallangan, ko‘p bosqichli strategiyaga asoslangan tabiiy tilni qayta ishlov berish modeli va uning konseptual asosi ishlab chiqildi. Ushbu modelning fundamental o‘ziga xosligi - tematik va sentiment tahlil vazifalarini ketma-ket hamda o‘zaro uzviy bog‘liq holda bajarish orqali matnlarning yashirin semantik mazmunini chuqurroq ifodalash hamda tematik sinf va sentiment bo‘yoq o‘rtasidagi bog‘liqlikni tahlil jarayonida hisobga olishga xizmat qilishidir.

Modelning umumiy matematik ifodasi. Taklif etilayotgan ko‘p bosqichli konseptual strategiyaning fundamental matematik ifodasi ierarxik semantik strukturaviylik, ko‘p o‘lchovli kontinuallik va kontekstual invariantlik tamoyillariga asoslanadi. Matn shunchaki diskret simvollar yoki leksemalar ketma-ketligi emas, balki cheksiz o‘lchovli va murakkab geometriyaga ega bo‘lgan continuous (uzluksizlik) semantik fazosidagi chiziqli bo‘lmagan trayektoriya sifatida qaraladi.

O‘zbek tilidagi bir nechta tematik sinfga tegishli va semantik jihatdan murakkab matnlar korpusi uchun umumiy semantik fazo S) quyidagicha modellashtiriladi:

$$S = M_{thematic} \times M_{sentiment} \text{ (2.1)}$$

Bu yerda $M_{thematic}$ va $M_{sentiment}$ mos ravishda matnning tematik va sentiment abstraksiya darajalarini ifodalovchi, o‘zaro chiziqli bo‘lmagan bog‘liqlikka ega bo‘lgan differensiallanuvchi ko‘pxilliklar (manifolds) tizimidir. Ko‘p bosqichli strategiyada ushbu ikki faza alohida emas, balki bir-birini shartlovchi va korrelyatsiya qiluvchi komponentlar sifatida ishlaydi.

Modelning matematik mohiyati topologik ma’lumotlar tahlili (TDA) hamda gibrid chuqur neyron tarmoqlarning latent transfer operatorlari apparatiga tayanadi. An’anaviy usullardagi masofa funksiyalari o‘zbek tilining semantik murakkabligini to‘liq aks ettira olmagani sababli, ushbu modelda ikki konseptual birlik o‘rtasidagi masofa (metrika) kontekstual variatsiyalarni hisobga oluvchi chiziqli bo‘lmagan akslantirishlar orqali optimallashtiriladi:

$$f: \chi \longrightarrow \mathcal{H}_{hybrid} \text{ (2.2)}$$

Bu yerda χ - o‘zbek tilidagi boshlang‘ich diskret matnlar to‘plami, $\mathcal{H}_{\text{hybrid}}$ -gibrid arxitektura yordamida ajratib olingan oliy tartibli latent xususiyatlarning Hilbert fazosidir. Gibrid neyron tarmoq modeli o‘zbek tilidagi elektron matnlarni avtomatik ravishda tematik sinflar bo‘yicha yuqori aniqlikda tasniflash, ya’ni ularni aniq tematik sinflarga ajratish masalasini samarali hal qiladi.

Takomillashtirilgan ko‘p bosqichli tasniflash algoritmi yordamida har bir $x \in \chi$ matn elementi uchun uning ko‘p yorliqlilik darajasi hisobga olinadi. Matnning har bir o‘lchovdagi tematik vazni va sentiment spektri quyidagi ehtimollik taqsimoti operatori orqali ketma-ket aniqlanadi:

$$P(Tasnif|x) = \Phi\left(g_{emot}\left(f_{them}(x)\right)\right) \text{(2.3)}$$

Bu yerda f_{them} - ko‘p bosqichli tasniflashga asoslangan tematik tahlil operatori bo‘lib, u murakkab va bir nechta tematik sinfga tegishli matnlarni modellashtirish sifatini oshiradi; g_{emot} - olingan tematik kontekst asosida ishlovchi sentiment tahlil operatori; Φ esa yakuniy kompleks semantik qaror qabul qiluvchi gibrid neyron tarmoq funksiyasidir.

Tizimning kirish va chiqish parametrlari.

Ko‘p bosqichli konseptual strategiyaning funksional arxitekturasi aniq belgilangan kirish va chiqish parametrlari tizimiga ega bo‘lib, o‘zbek tilining lingvistik va strukturaviy qonuniyatlarini to‘liq qamrab oladi.

1. Kirish parametrlari (Input Parameters).

Modelning hisoblash qatlamiga uzatiladigan boshlang‘ich va dinamik ma’lumotlar oqimi quyidagi parametrlardan iborat:

T_{uz_raw} (O‘zbek tilidagi xom matnlar to‘plami): Ijtimoiy tarmoqlar, ommaviy axborot vositalari, elektron murojaatlar yoki raqamli hujjatlardan olingan, tarkibida morfo-sintaktik o‘zgarishlar bo‘lgan, qisman yoki butunlay strukturalashmagan tabiiy til matnlari korpusi.

(cc) BY

D_{subj} (Tematik sinflar to‘plami): Matnlarni aniq tematik sinflar bo‘yicha guruhlash uchun xizmat qiluvchi o‘zbekcha domen lug‘atlari, terminologik ontologiyalar va lingvistik qoidalar bazasi.

H_{param} (Gibrid neyron tarmoq giperparametrlari): Ketma-ket va parallel qatlamlarning latent o‘lchamlari (d), tasniflash silsilasining ko‘p bosqichli chuqurlik darajasi hamda tematik-sentiment integratsiya koeffitsiyenti (α).

2. Chiqish parametrlari (Output Parameters).

Ko‘p bosqichli matematik va neyro-tarmoqli konvertatsiyalardan so‘ng, tizim amaliy qarorlar qabul qilish va boshqaruv tizimlariga integratsiya qilish uchun tayyor bo‘lgan quyidagi formal ko‘rsatkichlarni shakllantiradi:

$$C_{themztic} = \{c_1, c_2, \dots, c_k\}$$

(Ko‘p bosqichli tematik profillar koeffitsiyenti): Matnning faqatgina bitta tematik sinfga emas, balki bir nechta murakkab tematik sinflarga tegishlilik ehtimolliklari va uning aniq tematik sinflari.

$$E_{sentiment} = \{e_{pos}, e_{neg}, e_{neu}\}$$

(Sentiment vektori): Matndagi subyektiv munosabat, his-tuyg‘u (ijobiy, salbiy, neytral) va ularning intensivlik darajalari.

R_{struct}(Strukturalashtirilgan semantik interpretatsiya): Matnning kompyuter tushunadigan formatga keltirilgan konseptual xaritasi bo‘lib, u klassik TF-IDF yondashuviga nisbatan yuqoriroq, Transformer modellari bilan esa teng aniqlik (F1-score) hamda yozuv tizimi o‘zgarishlariga nisbatan barqarorlik (stability) ko‘rsatkichlarini namoyish etadi.

Ushbu taklif etilayotgan ko‘p bosqichli strategiya va ishlab chiqilgan intellektual tematik hamda sentiment tahlil qilish algoritmi o‘zbek tilidagi katta hajmli ma’lumotlar oqimini (ijtimoiy tarmoq postlari, OAV xabarlari, elektron murojaatlar va davlat organlarining raqamli hujjatlari) real vaqt rejimida monitoring qilish va chuqur semantik tahlil etish imkonini amaliy jihatdan to‘liq asoslaydi.

Gibrid neyron tarmoq modelining tarkibiy qismlari va funksional arxitekturasi.

O‘zbek tilidagi matnlarni tematik sinflar (domenlari) bo‘yicha avtomatik tasniflash hamda ularning semantik chuqurligini va sentimentini aniqlash uchun ushbu tadqiqot doirasida CNN va BiLSTM neyron tarmoq arxitekturalarining kombinatsiyasiga asoslangan gibrid model ishlab chiqildi. Mazkur gibrid yondashuv o‘zbek tilidagi matnning lokal xususiyatlarini, jumladan, so‘zlar va so‘z birikmalari darajasidagi muhim belgilarni CNN yordamida ajratib olish hamda ketma-ketlikning ikki yo‘nalishdagi kontekstual bog‘lanishlarini BiLSTM yordamida modellashtirish imkonini berish orqali yuqori tasniflash aniqligini ta’minlaydi.

Taklif etilayotgan gibrid CNN+BiLSTM modelining ierarxik strukturasi va funksional qatlamlari quyidagi tizimli ketma-ketlikda tashkil etilgan:

1. Kirish va embedding qatlami (Input & Embedding Layer). Tabiiy tildagi xom (dastlabki ishlov berilmagan) matnli ma’lumotlar dastlab raqamli formatga (zich vektor ko‘rinishiga) keltiriladi. O‘zbek tilining agglyutinativlik xususiyatini inobatga olgan holda, so‘z shakllari va ularning semantik kontekstini ifodalash uchun ikki xil yondashuv qo‘llanildi: birinchisida subso‘z (BPE) tokenlari uchun model bilan birgalikda o‘qitiladigan 128 o‘lchamli vektorlar ishlatildi (yengil konfiguratsiya, CNN+BiLSTM-Lite), ikkinchisida esa o‘zbek tili uchun oldindan o‘qitilgan Tahrirchi-BERT modelining 768 o‘lchamli kontekstual embeddinglaridan foydalanildi (Gibrid CNN+BiLSTM (BERT-emb) konfiguratsiyasi). Uzunligi *N* ga teng bo‘lgan matn quyidagi ko‘rinishdagi matritsa sifatida ifodalanadi:

$$\boldsymbol{X} = [\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_N] \in \mathbb{R}^{N \times d} \quad (2.4)$$

Bu yerda$e_t \in R^d$- matndagi t - so‘zning d-o‘lchovli zich semantik vektori.

2. Konvolyutsion qatlamlar (CNN Layers) [15]. Matnli embeddinglar matritsasi o‘lchamlari har xil bo‘lgan chiziqli bo‘lmagan

skanerlovchi filtrlar (konvolyutsiya darchalari) orqali o‘tkaziladi. CNN qatlamlarining asosiy vazifasi - matndagi n-gramm va lokal semantik xususiyatlarni (so‘zlar va iboralar o‘rtasidagi yaqin aloqalarni) ajratib olishdir. ω o‘lchamli filtr yordamida olingan lokal xususiyat $c_i$quyidagicha hisoblanadi:

$$c_i = f(\boldsymbol{W} \cdot \boldsymbol{X}_{i:i+w-1} + b) \ (2.5)$$

Bu yerda W - filtrning og‘irlik koeffitsiyentlari matritsasi, b - chetlanish (bias) parametri, f chiziqli bo‘lmagan aktivatsiya funksiyasi (ReLU).

3. Normallashtirish va Max Pooling qatlamlari. Konvolyutsiya qatlamidan olingan xususiyatlar xaritasi (feature map) qatlam bo‘yicha normallashtirish (Layer Normalization) [24] amali orqali me’yorlashtiriladi va uning ketma-ketlik uzunligi qisqartirilmagan holda BiLSTM qatlamiga uzatiladi. Bu so‘zlar tartibi haqidagi axborotni keyingi qatlamda saqlab qolish imkonini beradi. Maksimal pooling operatori esa BiLSTM qatlamining chiqishida qo‘llanilib, har bir o‘lcham bo‘yicha eng yuqori intensivlikka ega semantik signallarni saqlab qoladi:

$$h_{doc} = max_{1 \le t \le L} h_t \ (2.6)$$

Bu jarayon matn uzunligining o‘zgaruvchanligiga qaramay, eng muhim kontekstual xususiyatlarni invariant shaklda yakuniy tasniflash qatlamiga uzatadi.

4. Ikki yo‘nalishli rekurrent qatlam (BiLSTM Layer) [13]. CNN qatlamlari tomonidan ajratib olingan va normallashtirilgan lokal semantik xususiyatlar ketma-ketligi global kontekstni modellashtirish uchun BiLSTM qatlamlariga kirish parametri sifatida uzatiladi. Oddiy rekurrent tarmoqlardan farqli ravishda, BiLSTM matnning ketma-ketlik tuzilishini to‘g‘ri (forward) va teskari (backward) yo‘nalishlarda parallel ravishda o‘rganadi.

To‘g‘ri yo‘nalishdagi yashirin holat:

$$\vec{\boldsymbol{h}}_t = LSTM(\hat{c}_t, \vec{\boldsymbol{h}}_{t-1}) \ (2.7)$$

Teskari yo‘nalishdagi yashirin holat:

$$\overleftarrow{\underset{t}{\boldsymbol{h}}} = LSTM\left(\hat{c}_t, \overleftarrow{\underset{t+1}{\boldsymbol{h}}}\right) \ (2.8)$$

Yakuniy birlashtirilgan kontekst vektori har ikki yo‘nalishdagi holatlarning konkatenatsiyasi orqali shakllanadi:

$$\boldsymbol{h}_t = \left[\vec{\boldsymbol{h}}_t \parallel \overleftarrow{\underset{t}{\boldsymbol{h}}}\right] \ (2.9)$$

Bu esa matndagi so‘zlarning nafaqat o‘zidan oldingi, balki o‘zidan keyingi semantik birliklar bilan ham uzoq masofali bog‘liqliklarini to‘liq qamrab olish imkonini beradi.

5. Yakuniy tasniflash qatlami (Dense & Softmax Layer). BiLSTM qatlamidan olingan yuqori tartibli global va lokal xususiyatlar majmuasi to‘liq bog‘langan (Fully Connected/Dense) qatlamga yo‘naltiriladi. Tematik sinflar va sentiment sinflar bo‘yicha ehtimollik taqsimotini olish uchun yakuniy chiqish qiymati sentiment sinflar uchun Softmax, ko‘p yorliqli tematik tasniflash uchun esa har bir tematik sinf bo‘yicha alohida sigmoid funksiyasi orqali normalizatsiya qilinadi:

$$\hat{\boldsymbol{y}} = softmax(\boldsymbol{W}_y \cdot \boldsymbol{h}_t + b_y) \ (2.10)$$

Bu yerda $\hat{\mathbf{y}}$ - matnning tegishli tematik sinflarga tegishlilik ehtimolliklari vektoridir.

O‘zbek tilidagi matnlarga dastlabki ishlov berish bosqichlarining algoritmik va matematik tavsifi.

Gibrid CNN+BiLSTM modeliga matnli ma’lumotlarni uzatishdan oldin, xom ma’lumotlar tarkibidagi shovqinlarni (inglizcha “noise”) kamaytirish, hisoblash resurslarini optimallashtirish va modelning umumiy aniqligini oshirish maqsadida dastlabki ishlov berish (inglizcha “preprocessing”) bosqichi amalga oshiriladi. O‘zbek tilining agglyutinativ tabiati, ya’ni so‘z o‘zagiga ketma-ket qo‘shimchalar qo‘shilishi orqali yangi so‘z shakllarining hosil bo‘lishi, ushbu bosqichda maxsus algoritmik yondashuvlarni talab etadi.

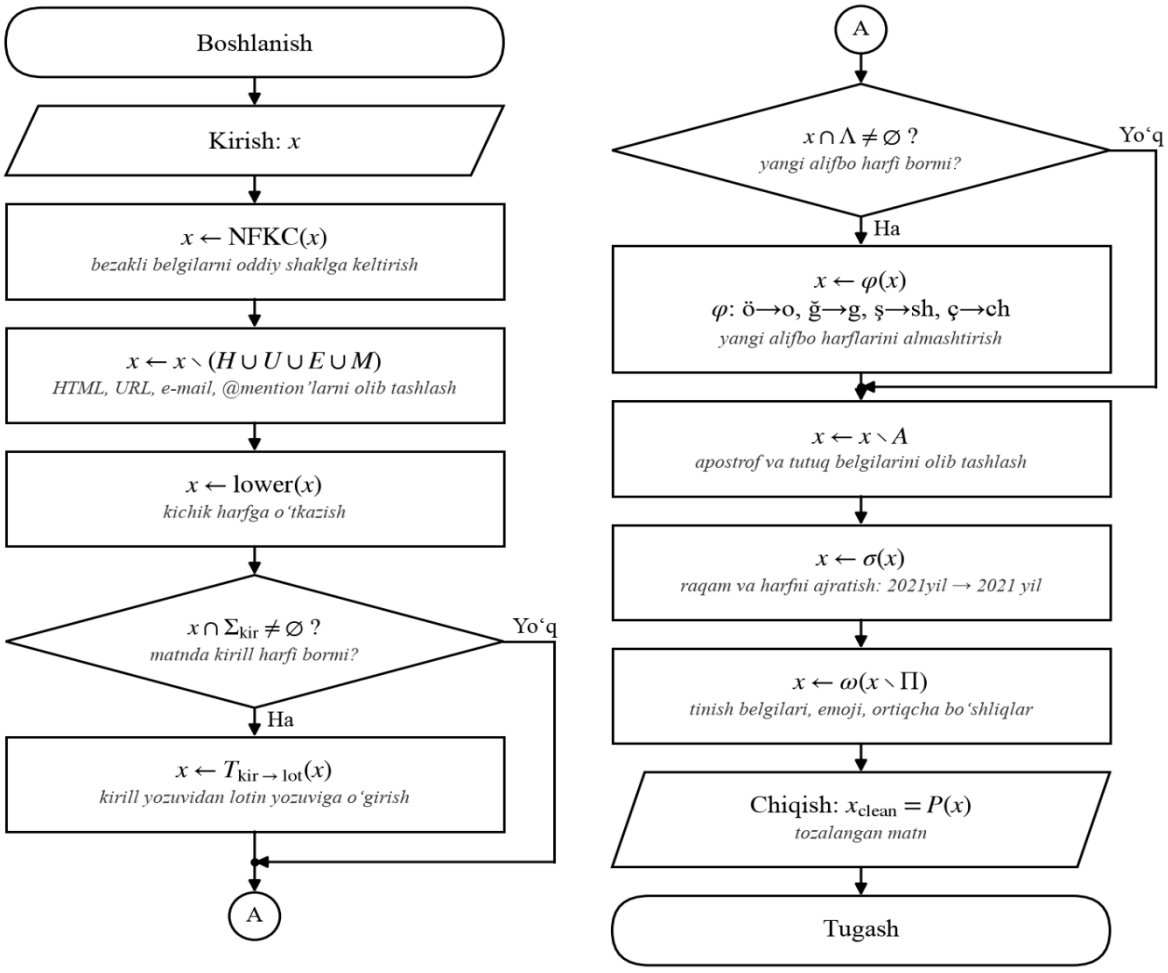
Tokenizatsiyadan oldin matnning yozuv shakli yagona ko‘rinishga keltiriladi. Raqamli makonda o‘zbek tilidagi matnlar lotin va kirill yozuvlarida, shuningdek, yangi tahrirdagi lotin alifbosida (ö, ğ, ş, ç harflari bilan) uchraydi, bir xil harf esa turli Unicode belgilari bilan yozilishi mumkin (masalan, o‘, o‘, o‘ yoki o`). Shu sababli normallashtirish operatori matnni Unicode NFKC

(cc) BY

shakliga keltiradi, HTML teglari va havolalarni olib tashlaydi, kirill harflarini lotin yozuviga transliteratsiya qiladi, yangi alifbo harflarini mos ravishda o, g, sh, ch ko‘rinishiga o‘tkazadi hamda apostrof va tutuq belgilarining barcha variantlarini bir xil qayta ishlaydi. Operator idempotentlik xossasiga ega, ya’ni $P(P(x)) = P(x)$: allaqachon normallashtirilgan matnga qayta qo‘llanilganda u o‘zgarmaydi.

O‘tkazilgan tajribalarda kirill yozuvidagi matnlar normallashtirishsiz berilganda modellarning aniqligi 53–80% ga pasaygani, normallashtirish qo‘llanilganda esa sifat yo‘qotilishi kuzatilmagani aniqlandi. Yangi alifbodagi matnlarda ham normallashtirish operatori modellarni qayta o‘qitmasdan 1,4 foiz punktlik pasayishni to‘liq bartaraf etdi.

Normallashtirish algoritmining blok-sxemasi 2.1-rasmda keltirilgan. Blok-sxemada quyidagi belgilashlar qabul qilingan: $x \cap S \neq \emptyset$ — x matnida S to‘plamining kamida bitta belgisi bor; $x \setminus S$ — S elementlarini x dan olib tashlash; H, U, E, M — HTML teglari, URL, e-mail manzillari va @mention’lar; Σ_{kir} — o‘zbek kirill alifbosi harflari; $T_{\text{kir}\to\text{lot}}$ — kirill yozuvidan lotin yozuviga transliteratsiya; Λ — yangi lotin alifbosi harflari, φ — ularni korpus yozuviga almashtirish; A — apostrof va tutuq belgilari; lower — kichik harfga o‘tkazish; σ — raqam va harf chegarasiga bo‘shliq qo‘yish; Π — harf, raqam va bo‘shliq bo‘lmagan belgilar (tinish belgilari, emoji); ω — ketma-ket bo‘shliqlarni bittaga keltirish; A — ustunlar orasidagi ulagich. Kichik harfga o‘tkazish transliteratsiyadan oldin bajariladi: kirill va yangi alifbo harflari jadvallari katta va kichik harflarni birday qamragani uchun bu tartib natijani o‘zgartirmaydi.



2.1-rasm. Matnni normallashtirish algoritmining blok-sxemasi

(cc) BY dtai.uz

Dastlabki ishlov berish jarayoni formal ravishda quyidagi ketma-ket funksional operatorlar zanjiri ko‘rinishida ifodalanadi:

$$P: T_{uz_raw} \longrightarrow \chi_{clean} \quad (2.11)$$

Ushbu transformatsiya jarayoni matematik va algoritmik jihatdan quyidagi uchta asosiy mikrobosqichni o‘z ichiga oladi:

1. Matnni segmentatsiyalash va tokenizatsiya (Tokenization) qilish.

Tokenizatsiya - uzluksiz belgilar ketma-ketligidan iborat bo‘lgan xom matnni T_{uz_raw} alohida semantik birliklar - tokenlar (so‘zlar, morfemalar yoki subso‘zlar) to‘plamiga ajratish jarayonidir. O‘zbek tilidagi matnlar uchun oddiy probellar orqali ajratish yetarli bo‘lmagani sababli (masalan, qo‘shma so‘zlar yoki chiziqcha bilan yoziladigan so‘zlar mavjudligi bois), tokenizatsiya operatori T quyidagicha aniqlanadi:

$$T\left(T_{uz_{raw}}\right) = \{t_1, t_2, \dots, t_N\}, t_i \in V \quad (2.12)$$

Bu yerda V - tizimning umumiy lug‘at tarkibi (vocabulary). Amaliyotda o‘zbek tili lug‘at fondining cheksiz kengayib ketishini (inglizcha Out-Of-Vocabulary muammosini) oldini olish maqsadida Byte-Pair Encoding (BPE) [25] yoki WordPiece algoritmlari qo‘llaniladi. Ushbu algoritmlar so‘zlarni eng kichik ma’noli qismlarga (subso‘zlarga) parchalash orqali ishlaydi, masalan:

T(“kitoblarimizdan”) = {“kitob”, “##lar”, “##imiz”, “##dan”}

2. Stop-so‘zlarni filtrlash va shovqinlarni yo‘qotish (Stop-words Removal).

Matnda tez-tez takrorlanadigan, biroq uning tematik yoki sentiment mazmunini aniqlashda muhim semantik yuklamaga ega bo‘lmagan yordamchi so‘zlar (masalan: va, bilan, uchun, ham, ammo, esa, chunki va h.k.), punktuatsiya belgilari, raqamlar va gipersilkalar matndan olib tashlanadi.

Formal ravishda, tizimda oldindan shakllantirilgan va o‘zbek tilining o‘ziga xosliklarini hisobga oluvchi stop-so‘zlar to‘plami S_{stop} mavjud deb qaralsa, filtrlash operatori F quyidagi shart asosida ishlaydi:

$$F(\{t_1, t_2, \dots, t_N\}) = \{t_i | t_i \notin S_{stop} \wedge is_{alpha(t_i)}\} \quad (2.13)$$

Biroq o‘tkazilgan tajribalar shuni ko‘rsatdiki, subso‘z (BPE) tokenizatsiyasi qo‘llanilganda stop-so‘zlarni oldindan olib tashlash tasniflash aniqligini oshirmaydi: neyron tarmoq kam axborotli birliklarning ta’sirini o‘qitish jarayonida o‘zi kamaytiradi, inkor va modal so‘zlarning olib tashlanishi esa sentiment tahlil uchun muhim belgilarning yo‘qolishiga olib kelishi mumkin. Shu sababli ishlab chiqilgan dasturiy vositada ushbu operator ixtiyoriy bosqich sifatida saqlab qolindi.

3. Stemming va lemmatizatsiya (Stemming & Lemmatization).

O‘zbek tilida so‘z o‘zgartiruvchi va so‘z yasovchi qo‘shimchalar juda ko‘p bo‘lganligi sababli, birgina o‘zak o‘nlab turlicha shakllarda kelishi mumkin (masalan: ish, ishchi, ishladi, ishlashmoqda). Model ushbu so‘zlarni bir-biri bilan bog‘liq semantik invariant sifatida tushunishi uchun ularni boshlang‘ich lug‘at shakliga (lemmaga) keltirish zarur.

Ushbu tadqiqotda o‘zbek tilining qat’iy agglyutinativ qoidalariga asoslangan, kontekstni hisobga oluvchi matematik lemmatizatsiya operatori Lem ishlab chiqilgan. Har bir token t_i shartli ravishda o‘zak (lemma) l_i va affikslar (qo‘shimchalar) a_{ij} ketma-ketligi yig‘indisi sifatida qaraladi:

$$t_i = l_i + \sum_{j=1}^{m} a_{ij} \quad (2.14)$$

Lemmatizatsiya algoritmi matndagi so‘zning turkumini va uning atrofidagi so‘zlar (kontekst) bilan bog‘liqligini tahlil qilib, optimal o‘zakni aniqlaydi:

$$Lem\ (t_i \mid kontekst) = l_i, \quad l_i \in \mathcal{V}_{lexicon} \quad (2.15)$$

Masalan, ushbu algoritm yordamida Lem(“maktablarimizda”) = "maktab" ko‘rinishida o‘zak ajratiladi.

Dastlabki ishlov berish bosqichlarining barchasidan muvaffaqiyatli o‘tgan toza va invariant tokenlar ketma-ketligi $\chi_{\text{clean}} = \{l_1, l_2, \dots, l_k\}$ keyingi bosqichda zich semantik vektorlar (Embedding) matritsasiga aylantirilib,

(cc) BY

bevosita CNN+BiLSTM gibrid neyron tarmog‘ining kirish qatlamiga uzatiladi. Bu esa o‘zbek tilidagi bir nechta tematik sinfga tegishli hamda semantik jihatdan murakkab matnlarni modellashtirish sifatini tubdan oshirishga xizmat qiladi.

Shuni ta’kidlash lozimki, subso‘z tokenizatsiyasi so‘z o‘zagi va qo‘shimchalarni alohida birliklar sifatida ajratgani bois, tajribalarda lemmatizatsiya operatorining qo‘shimcha qo‘llanilishi tasniflash aniqligini sezilarli o‘zgartirmadi. Shu sababli yakuniy modelda so‘z shakllarining xilma-xilligi asosan subso‘z tokenizatsiyasi orqali hisobga olindi, lemmatizatsiya operatori esa ixtiyoriy bosqich sifatida saqlab qolindi.

Modelni o‘rgatish (Training) va sinovdan o‘tkazish (Testing) jarayonlari.

Gibrid CNN+BiLSTM modelining parametrlari va og‘irlik koeffitsiyentlarini optimallashtirish katta hajmdagi o‘zbek tilidagi matnlar korpusi (ijtimoiy tarmoqlar, OAV va elektron murojaatlar) asosida amalga oshirildi. O‘rganish jarayoni quyidagi matematik bosqichlarni o‘z ichiga oladi:

- Yo‘qotish funksiyasini hisoblash (Loss Function). Model tomonidan bashorat qilingan natija ($\hat{y}$) bilan haqiqiy etalon sinf (y) o‘rtasidagi farqni minimallashtirish uchun ko‘p sinfli kross-entropiya (Categorical Cross-Entropy) yo‘qotish funksiyasidan foydalaniladi:

$$\mathcal{L} = -\sum_{i=1}^{M} y_i \, log(\hat{\boldsymbol{y}}_i) \quad (2.16)$$

Bu yerda (M) - tasniflanishi kerak bo‘lgan jami tematik yoki sentiment sinflar soni.

- Gradientni orqaga tarqatish (Backpropagation). Yo‘qotish funksiyasining gradientlari $\nabla\mathcal{L}$ tarmoq qatlamlari bo‘yicha zanjir qoidasiga ko‘ra orqaga uzatiladi va og‘irlik koeffitsiyentlari AdamW optimizatori va o‘rganish tezligini bir siklli (OneCycle) boshqarish usuli [26] yordamida adaptiv tarzda yangilanadi.

Model o‘rgatilgandan so‘ng, u ilgari neyron tarmoqqa ko‘rsatilmagan test (sinov) ma’lumotlar to‘plami yordamida tekshiriladi. Modelning umumiy barqarorligi va aniqlik ko‘rsatkichlarini baholash uchun quyidagi standart metrikalar tizimi qo‘llaniladi:

Aniqlik (Accuracy). To‘g‘ri tasniflangan matnli hujjatlar sonining umumiy sinovdan o‘tkazilgan hujjatlar miqdoriga nisbati.

Recall (Takrorlash). Modelning berilgan haqiqiy sinf yoki tematik sinfga tegishli hujjatlarni qidirib topish va to‘g‘ri aniqlashdagi chuqurlik samaradorligi.

F1-o‘lchov (F1-score). Aniq mantiqiy xulosaga kelish uchun Precision (aniqlik darajasi) va Recall ko‘rsatkichlarining garmonik o‘rtacha qiymati bo‘lib, o‘zbek tili korpusidagi nomutanosib (imbalanced) sinflarni xolis baholash imkonini beradi.

Yuqorida keltirilgan ko‘rsatkichlar quyidagi formulalar orqali hisoblanadi [22]:

$$Aniqlik = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.17)$$

$$Precision = \frac{TP}{TP+FP} \quad (2.18)$$

$$Recall = \frac{TP}{TP+FN} \quad (2.19)$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall} \quad (2.20)$$

$$F1_{makro} = \frac{1}{M}\sum_{i=1}^{M} F1_i \quad (2.21)$$

Bu yerda *TP*, *TN*, *FP* va *FN* mos ravishda to‘g‘ri musbat, to‘g‘ri manfiy, noto‘g‘ri musbat va noto‘g‘ri manfiy tasniflashlar soni; $F1_i$ — i-sinf uchun F1-o‘lchov; M — sinflar soni. Ko‘p yorliqli tematik tasniflashda har bir tematik sinf alohida ikkilik vazifa sifatida baholanadi, so‘ng natijalar sinflar bo‘yicha o‘rtalashtiriladi (makro-o‘rtacha, (2.21)). Sinflar nomutanosib bo‘lganda makro-F1 kam uchraydigan sinflarning ham hissasini teng hisobga oladi, shu sababli u faqat aniqlikka nisbatan xolisroq baho beradi.

Algoritmning barcha qadamlarini bajarib, model matnli ma’lumotlarni tasniflashda yuqori natijalarga erishadi. CNN qatlamlari matndagi lokal xususiyatlarni aniqlaydi, BiLSTM esa matnning ketma-ketlikdagi global kontekstini tushunishga yordam beradi. Bunday gibrid

(cc) BY

yondashuv matnni to‘liq va samarali tahlillash imkonini beradi.

CNN+BiLSTM gibrid strategiyasining konseptual afzalliklari.

Taklif etilgan ushbu gibrid neyron tarmoq modeli o‘zbek tili tabiiy tilini qayta ishlashda mavjud standart (baza) modellarga nisbatan quyidagi konseptual afzalliklarga ega:

- Lokal va global xususiyatlarni birgalikda o‘rganish. CNN matnning morfologik qismlari, o‘zak-qo‘shimcha munosabatlari va lokal frazalarini samarali aniqlasa, BiLSTM ushbu lokal belgilarni umumiy gap va matn konteksti silsilasida ierarxik bog‘laydi;

- Kontekstual ikki yo‘nalishli tahlil. O‘zbek tilida erkin so‘z tartibi mavjudligi sababli, gapdagi ega, kesim yoki aniqlovchining o‘rni o‘zgarishi mumkin. BiLSTMning ikki yo‘nalishli arxitekturasi matn mazmunini ham oldinga, ham orqaga qarab zanjirsimon tahlil qilgani bois, kontekstual mantiqning yo‘qolishiga yo‘l qo‘ymaydi;

- Yuqori aniqlik va barqarorlik. Ushbu gibrid tizim semantik jihatdan murakkab, bir nechta tematik sinfga tegishli hamda sentiment yuklamaga ega bo‘lgan elektron xabarlarni, rasmiy raqamli hujjatlarni hamda ijtimoiy tarmoqlardagi xalqchil nutq ko‘rinishlarini tematik sinflar bo‘yicha yuqori barqarorlik bilan differensiyalash imkonini beradi.

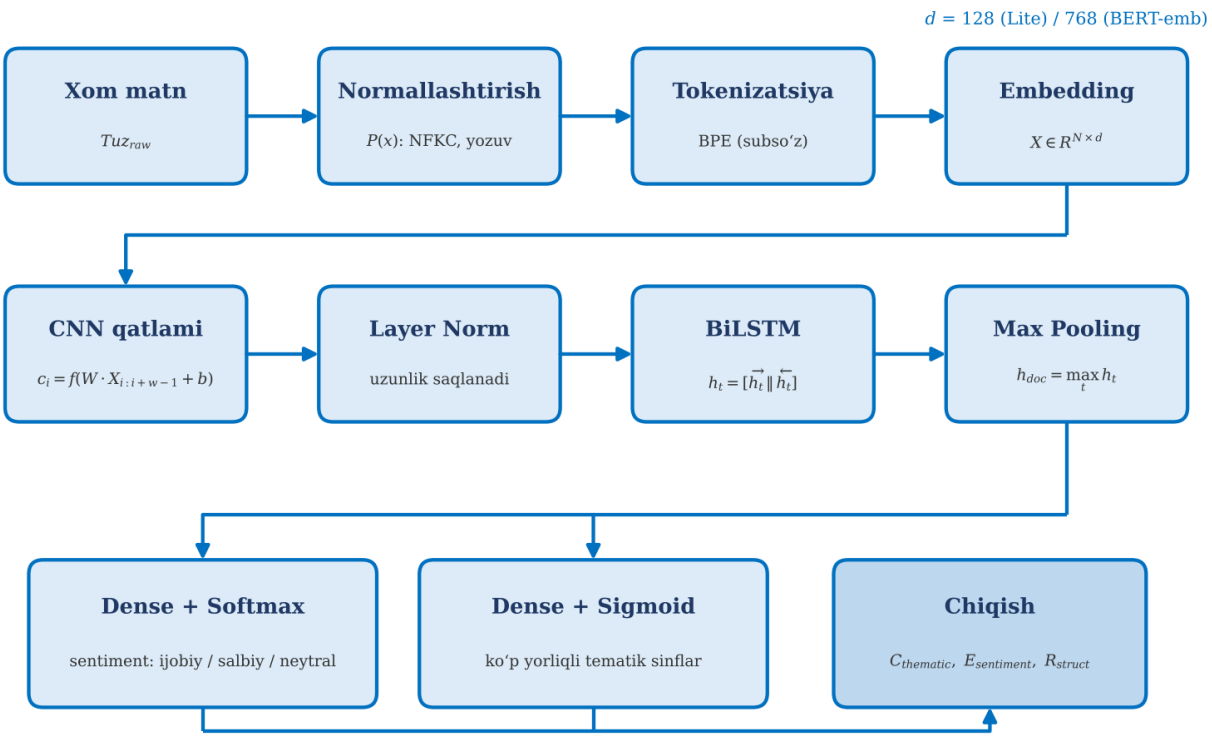
Gibrid modelning tarkibiy qismlari, ularning matematik ifodasi va bajaradigan vazifalari 1-jadvalda umumlashtirilgan, modelning dastlabki ishlov berishdan yakuniy tasniflashgacha bo‘lgan umumiy arxitekturasi esa 2.2-rasmda keltirilgan. Rasmda ko‘rinib turibdiki, BiLSTM qatlami chiqishidagi yagona hujjat vektori ikkita mustaqil chiqish boshiga — sentiment va ko‘p yorliqli tematik tasniflash boshlariga — uzatiladi.

1-jadval.

Gibrid CNN+BiLSTM modelining tarkibiy qismlari va ularning vazifalari

Qatlam	Matematik ifoda	Vazifasi
Embedding	$X = (e_1, \ldots, e_N) \in R^{N\times d}$ (2.4)	So‘zlarni zich semantik vektorlarga o‘tkazish: 128 o‘lchamli BPE embeddinglari yoki 768 o‘lchamli Tahrirchi-BERT embeddinglari
CNN	$c_i = f(W\cdot X_{i:i+w-1} + b)$ (2.5)	Lokal n-gramm va yaqin semantik aloqalarni ajratib olish
Layer Normalization	—	Xususiyatlar xaritasini me’yorlashtirish; ketma-ketlik uzunligi qisqartirilmaydi va so‘zlar tartibi saqlanadi
BiLSTM	$h_t = [\rightarrow h_t \,\|\, \leftarrow h_t]$ (2.7)–(2.9)	Ikki yo‘nalishli global kontekst va uzoq masofali bog‘liqliklarni modellashtirish
Max Pooling	$h_{doc} = \max h_t$ (2.6)	Eng intensiv semantik signallarni matn uzunligiga invariant tarzda ajratish
Dense + Softmax / sigmoid	$y = \text{softmax}(W_y\cdot h + b_y)$ (2.10)	Sentiment sinflari uchun Softmax, ko‘p yorliqli tematik tasniflash uchun har bir sinf bo‘yicha alohida sigmoid

(cc) BY



2.2-rasm. Taklif etilayotgan gibrid CNN+BiLSTM modelining umumiy arxitekturasi

NATIJALAR VA MUHOKAMA

Tajribalar ijtimoiy tarmoqlar, ommaviy axborot vositalari va elektron murojaatlardan yig‘ilgan o‘zbek tilidagi matnlar korpusida o‘tkazildi. Korpus o‘rgatish va sinov qismlariga ajratilgan bo‘lib, sinov to‘plami modelga o‘qitish jarayonida ko‘rsatilmagan hujjatlardan iborat. Modelning ikki konfiguratsiyasi — yengil CNN+BiLSTM-Lite va oldindan o‘qitilgan Tahrirchi-BERT embeddinglaridan foydalanuvchi Gibrid CNN+BiLSTM (BERT-emb) — o‘rganildi; ularning asosiy parametrlari 2-jadvalda keltirilgan. Baholash uchun Aniqlik, Precision, Recall va F1-score ko‘rsatkichlaridan, nomutanosib sinflar uchun esa makro-F1 dan foydalanildi (2.17)–(2.21) [22]. Parametrlar Adam [21] ning vazn so‘nishi ajratilgan varianti AdamW [20] yordamida optimallashtirildi. Taqqoslash uchun TF-IDF tasviriga asoslangan klassik yondashuv va Transformer modellari baza (baseline) sifatida olindi.

2-jadval.

Modelning ikki konfiguratsiyasi va o‘rgatish parametrlari

Parametr	CNN+BiLSTM-Lite	Gibrid CNN+BiLSTM (BERT-emb)
Kirish tasviri	BPE tokenlar uchun model bilan birga o‘qitiladigan embedding	Tahrirchi-BERT modelining kontekstual embeddinglari
Embedding o‘lchami (d)	128	768
Qatlamlar ketma-ketligi	CNN → Layer Norm → BiLSTM → Max Pooling → Dense	CNN → Layer Norm → BiLSTM → Max Pooling → Dense
Chiqish funksiyasi	Sentiment — Softmax; tematik sinflar — alohida sigmoid	Sentiment — Softmax; tematik sinflar — alohida sigmoid
Yo‘qotish funksiyasi	Kross-entropiya (2.16)	Kross-entropiya (2.16)
Optimizator	AdamW [20], OneCycle o‘rganish tezligi [26]	AdamW [20], OneCycle o‘rganish tezligi [26]
Dastlabki ishlov	Normallashtirish operatori P (2.11)	Normallashtirish operatori P (2.11)

(cc) BY

Dastlabki ishlov berishning har bir operatori alohida tekshirildi; kuzatilgan natijalar 3-jadvalda umumlashtirilgan. Eng muhim natija yozuv tizimini normallashtirish bilan bog‘liq: kirill yozuvidagi matnlar normallashtirishsiz berilganda modellarning aniqligi 53–80% ga pasaygan, operator qo‘llanilganda esa sifat yo‘qotilishi kuzatilmagan. Yangi alifbodagi matnlarda ham normallashtirish operatori modellarni qayta o‘qitmasdan 1,4 foiz punktlik pasayishni to‘liq bartaraf etdi. Bu operatorning idempotentligi ($P(P(x)) = P(x)$) uni har qanday kirish ma’lumotiga xavfsiz qo‘llash imkonini beradi.

3-jadval.

Dastlabki ishlov berish bosqichlarining ta’siri bo‘yicha kuzatilgan natijalar

Bosqich	Tajribada kuzatilgan natija	Qaror
Normallashtirish (kirill yozuvi)	Normallashtirishsiz aniqlik 53–80% ga pasaygan; operator qo‘llanilganda sifat yo‘qotilmagan	Majburiy bosqich
Normallashtirish (yangi alifbo)	1,4 foiz punktlik pasayish qayta o‘qitishsiz to‘liq bartaraf etilgan	Majburiy bosqich
Stop-so‘zlarni olib tashlash (BPE bilan)	Aniqlik oshmagan; inkor va modal so‘zlar yo‘qolishi mumkin	Ixtiyoriy bosqich
Lemmatizatsiya (BPE bilan)	Aniqlik sezilarli o‘zgarmagan	Ixtiyoriy bosqich

Stop-so‘zlarni filtrlash va lemmatizatsiya bo‘yicha natijalar boshqacha bo‘ldi. BPE tokenizatsiyasi qo‘llanilganda stop-so‘zlarni oldindan olib tashlash aniqlikni oshirmadi, chunki neyron tarmoq kam axborotli birliklarning ta’sirini o‘qitish jarayonida o‘zi kamaytiradi; inkor va modal so‘zlarning olib tashlanishi esa sentiment tahlil uchun muhim belgilarning yo‘qolishiga olib kelishi mumkin. Lemmatizatsiyaning qo‘shimcha qo‘llanilishi ham aniqlikni sezilarli o‘zgartirmadi, sababi subso‘z tokenizatsiyasi so‘z o‘zagi va qo‘shimchalarni alohida birliklar sifatida ajratadi. Shu bois ikkala operator ham ixtiyoriy bosqich sifatida saqlab qolindi.

Taqqoslash natijalariga ko‘ra, taklif etilgan yondashuv klassik TF-IDF usuliga nisbatan yuqoriroq aniqlikka, Transformer modellari bilan esa teng aniqlikka (F1-score) erishadi, shu bilan birga yozuv tizimi o‘zgarishiga nisbatan barqarorlik (stability) ko‘rsatkichi yuqori saqlanadi. TF-IDF tasvirida so‘zlar tartibi va kontekst yo‘qoladi, CNN+BiLSTM esa lokal va global bog‘liqliklarni birgalikda hisobga oladi; bu ustunlikning asosiy sababi hisoblanadi. Transformer modellari bilan tenglikka esa nisbatan sodda arxitektura bilan erishilgani amaliy ahamiyatga ega: xususan, 128 o‘lchamli embeddingli CNN+BiLSTM-Lite konfiguratsiyasi 768 o‘lchamli kontekstual embeddinglarga nisbatan kichik vektor o‘lchamiga ega bo‘lgani uchun hisoblash yukini kamaytiradi va real vaqt rejimidagi monitoring uchun qulay.

Muhokama

Olingan natijalar o‘zbek tilining lingvistik xususiyatlari bilan izohlanadi. Erkin so‘z tartibi sababli ega, kesim yoki aniqlovchining o‘rni o‘zgarishi mumkin; BiLSTM matnni ham oldinga, ham orqaga qarab tahlil qilgani uchun bunday o‘zgarishlarda kontekstual mantiq saqlanadi. Agglyutinativ morfologiya esa CNN va subso‘z tokenizatsiyasi yordamida o‘zak va qo‘shimchalar darajasida qamrab olinadi. Layer Normalization natijasining BiLSTMga uzunligi qisqartirilmagan holda uzatilishi so‘zlar tartibi haqidagi axborotni saqlaydi, yakuniy Max Pooling esa matn uzunligining o‘zgaruvchanligiga qaramay eng muhim signallarni invariant shaklda ajratadi.

Dastlabki ishlov berish tajribalari amaliy xulosa beradi: subso‘z tokenizatsiyasi qo‘llanilgan tizimlarda murakkab morfologik tahlil (lemmatizatsiya) va stop-so‘zlarni filtrlash mavjud foydaning kichik qismini beradi, yozuv tizimini normallashtirish esa muhim bo‘lib qoladi. Real ma’lumot oqimlarida lotin, kirill va yangi alifbodagi matnlar aralash uchraydi, shu sababli normallashtirishsiz tizimning aniqligi keskin pasayishi mumkin.

(cc) BY

Tematik va sentiment tahlilni ketma-ket bajarishning (2.3) ahamiyati shundaki, bir xil so‘z mavzuga qarab turlicha emotsional yuklamaga ega bo‘lishi mumkin. Masalan, “o‘sish” so‘zi iqtisodiy ko‘rsatkichlar haqidagi xabarda ijobiy, jinoyatchilik yoki kasallanish haqidagi xabarda esa salbiy ma’no kasb etadi. Sentiment operatorining tematik kontekst asosida ishlashi bunday holatlarda xatolarni kamaytirishga xizmat qiladi.

Tadqiqotning cheklovlari ham mavjud. Birinchidan, ijtimoiy tarmoq matnlarida sarkazm, dialekt shakllari va o‘zbek-rus tillari aralashuvi (code-switching) tasniflashni qiyinlashtiradi. Ikkinchidan, tematik sinflar soni va ularning nomutanosibligi kam uchraydigan mavzular bo‘yicha natijalarni pasaytirishi mumkin. Uchinchidan, tahlil o‘zbek tili korpusi bilan cheklangan va boshqa agglyutinativ tillarga ko‘chirish alohida tekshiruvni talab etadi. Kelgusi tadqiqotlarda korpus hajmini kengaytirish, qaror qabul qilish jarayonining izohlanuvchanligini (masalan, diqqat mexanizmi orqali) oshirish, terminologik ontologiyalarni (D_{subj}) modelga chuqurroq integratsiya qilish hamda real vaqt rejimida ishlash uchun modelni siqish yo‘nalishlari rejalashtirilgan.

XULOSA

Tadqiqot natijasida o‘zbek tilidagi elektron matnlarning semantik xususiyatlarini chuqur ifodalashga mo‘ljallangan ko‘p bosqichli konseptual model va uning asosidagi gibrid CNN+BiLSTM algoritmi ishlab chiqildi. Asosiy xulosalar quyidagilardan iborat:

1) Matn uzluksiz semantik fazodagi trayektoriya sifatida, semantik fazo esa tematik va sentiment ko‘pxilliklarining ko‘paytmasi sifatida ifodalandi; tematik va sentiment tahlil ketma-ket hamda o‘zaro bog‘liq holda bajariladi va bu ikki sinf o‘rtasidagi bog‘liqlikni hisobga olish imkonini beradi.

2) CNN lokal xususiyatlarni, BiLSTM esa ikki yo‘nalishli global kontekstni modellashtiruvchi gibrid arxitektura ishlab chiqildi; model ikki konfiguratsiyada — CNN+BiLSTM-Lite (128 o‘lchamli BPE embeddinglari) va Gibrid CNN+BiLSTM (768 o‘lchamli Tahrirchi-BERT embeddinglari) shaklida — amalga oshirildi.

3) Lotin, kirill va yangi lotin alifbosidagi matnlarni yagona shaklga keltiruvchi idempotent normallashtirish operatori kirill yozuvidagi matnlarda kuzatilgan 53–80% aniqlik pasayishini va yangi alifbodagi 1,4 foiz punktlik pasayishni qayta o‘qitishsiz bartaraf etdi.

4) BPE tokenizatsiyasi sharoitida stop-so‘zlarni olib tashlash va lemmatizatsiya aniqlikni sezilarli oshirmadi, shu sababli ular ixtiyoriy bosqich sifatida saqlab qolindi.

5) Taklif etilgan yondashuv klassik TF-IDF usulidan yuqori, Transformer modellari bilan teng F1-score hamda yozuv tizimi o‘zgarishiga nisbatan barqarorlikni namoyish etadi va ijtimoiy tarmoq postlari, OAV xabarlari, elektron murojaatlar hamda davlat organlarining raqamli hujjatlari oqimini real vaqt rejimida monitoring qilish va chuqur semantik tahlil etish uchun amaliy asos bo‘ladi.

ADABIYOTLAR RO‘YXATI

1. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. Attention is all you need // Advances in Neural Information Processing Systems 30 (NeurIPS 2017). – 2017. – P. 5998–6008.
2. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding // Proceedings of NAACL-HLT 2019. – 2019. – P. 4171–4186.
3. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space // arXiv:1301.3781. – 2013.
4. Hochreiter S., Schmidhuber J. Long short-term memory // Neural Computation. – 1997. – Vol. 9, No. 8. – P. 1735–1780.
5. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval // Information Processing & Management. – 1988. – Vol. 24, No. 5. – P. 513–523.
6. Liu B. Sentiment Analysis and Opinion Mining. – San Rafael: Morgan & Claypool Publishers, 2012. – 167 p.
7. Pang B., Lee L. Opinion mining and sentiment analysis // Foundations and Trends

(cc) BY

in Information Retrieval. – 2008. – Vol. 2, No. 1–2. – P. 1–135.
8. Bojanowski P., Grave E., Joulin A., Mikolov T. Enriching word vectors with subword information // Transactions of the Association for Computational Linguistics. – 2017. – Vol. 5. – P. 135–146.
9. Conneau A., Khandelwal K., Goyal N., et al. Unsupervised cross-lingual representation learning at scale // Proceedings of ACL 2020. – 2020. – P. 8440–8451.
10. Kudo T., Richardson J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing // Proceedings of EMNLP 2018: System Demonstrations. – 2018. – P. 66–71.
11. Kuriyozov E., Matlatipov S., Alonso M.A., Gómez-Rodríguez C. Construction and evaluation of sentiment datasets for low-resource languages: The case of Uzbek // Human Language Technology. Challenges for Computer Science and Linguistics (LTC 2019). – Cham: Springer, 2020. – P. 232–243.
12. Rabbimov I.M., Kobilov S.S. Multi-class text classification of Uzbek news articles using machine learning // Journal of Physics: Conference Series. – 2020. – Vol. 1546. – Art. 012097.
13. Schuster M., Paliwal K.K. Bidirectional recurrent neural networks // IEEE Transactions on Signal Processing. – 1997. – Vol. 45, No. 11. – P. 2673–2681.
14. Zhou C., Sun C., Liu Z., Lau F. A C-LSTM neural network for text classification // arXiv:1511.08630. – 2015.
15. Kim Y. Convolutional neural networks for sentence classification // Proceedings of EMNLP 2014. – 2014. – P. 1746–1751.
16. Lai S., Xu L., Liu K., Zhao J. Recurrent convolutional neural networks for text classification // Proceedings of the 29th AAAI Conference on Artificial Intelligence. – 2015. – P. 2267–2273.
17. Zhang X., Zhao J., LeCun Y. Character-level convolutional networks for text classification // Advances in Neural Information Processing Systems 28 (NIPS 2015). – 2015. – P. 649–657.
18. Rabbimov I., Mporas I., Simaki V., Kobilov S. Investigating the effect of emoji in opinion classification of Uzbek movie review comments // Speech and Computer (SPECOM 2020), LNCS 12335. – Cham: Springer, 2020. – P. 435–445.
19. Salaev U., Kuriyozov E., Gómez-Rodríguez C. A machine transliteration tool between Uzbek alphabets // arXiv:2205.09578. – 2022.
20. Loshchilov I., Hutter F. Decoupled weight decay regularization // International Conference on Learning Representations (ICLR 2019). – 2019.
21. Kingma D.P., Ba J. Adam: A method for stochastic optimization // International Conference on Learning Representations (ICLR 2015). – 2015.
22. Sokolova M., Lapalme G. A systematic analysis of performance measures for classification tasks // Information Processing & Management. – 2009. – Vol. 45, No. 4. – P. 427–437.
23. Tahrirchi. tahrirchi-bert-base: pretrained BERT model for the Uzbek language [Electronic resource]. – Hugging Face. – URL: https://huggingface.co/tahrirchi/tahrirchi-bert-base
24. Ba J.L., Kiros J.R., Hinton G.E. Layer normalization // arXiv:1607.06450. – 2016.
25. Sennrich R., Haddow B., Birch A. Neural machine translation of rare words with subword units // Proceedings of ACL 2016. – 2016. – P. 1715–1725.
26. Smith L.N., Topin N. Super-convergence: Very fast training of neural networks using large learning rates // arXiv:1708.07120. – 2017.

(cc) BY