MATNLARDAN TOKSIK (ZARARLI) KONTENTNI ANIQLASH UCHUN SUN'IY INTELLEKT MODELINI YARATISH

Toshpulatov Javoxir Xamidulla o‘g‘li[1]

[1] O‘zbekiston Respublikasi Ekologiya va iqlim o‘zgarishi milliy qo‘mitasi, Sun’iy intellekt texnologiyalari, raqamlashtirish va ma’lumotlarni boshqarish bo‘limi bosh mutaxassisi
[1] E-mail: M25029@tiue.uz
DOI: 10.61587/mmit.tiue.uz.v1i1.279

Annotatsiya. Ushbu maqolada matnlardagi toksik va zararli kontentni avtomatik aniqlash muammosi ko'rib chiqiladi. Ijtimoiy tarmoqlar va onlayn platformalarda toksik mazmunning ko'payishi nafaqat foydalanuvchilarga, balki jamiyatga ham salbiy ta'sir ko'rsatmoqda. Tadqiqotda gibrid sun'iy intellekt arxitekturasi taklif etilgan bo'lib, u BERT asosidagi transformerlar va ikki yo'nalishli LSTM tarmoqlarini birlashtirib ishlatadi. Taklif etilgan model O'zbekiston ijtimoiy tarmoqlaridan yig'ilgan o'zbek va rus tillardagi aralash matnlar asosida mashq qildirildi. Tajribalar natijasida model 94,7% aniqlik (accuracy), 93,2% to'liqlik (recall) va 94,0% F1-ko'rsatkich qiymatlariga erishdi. Bundan tashqari, maqolada ma'lumotlar to'plami tayyorlash, modelni baholash mezonlari va real vaqt rejimida qo'llash imkoniyatlari tahlil qilingan.

Kalit so‘zlar: *toksik kontent, tabiiy tilni qayta ishlash, BERT, ikki yo'nalishli LSTM, deep learning, matn klassifikatsiyasi, ijtimoiy tarmoqlar moderatsiyasi, O'zbek NLP.*

DEVELOPMENT OF AN ARTIFICIAL INTELLIGENCE MODEL FOR DETECTING TOXIC (HARMFUL) CONTENT IN TEXTS

Toshpulatov Javoxir Xamidulla ogli[1]

[1] Chief Specialist, Artificial Intelligence Technologies, Digitalization and Data Management Department, National Committee of the Republic of Uzbekistan on Ecology and Climate Change
[1] E-mail: M25029@tiue.uz

Abstract. This paper addresses the problem of automatic detection of toxic and harmful content in texts. The proliferation of toxic content on social networks and online platforms has a negative impact not only on users but also on society as a whole. The study proposes a hybrid artificial intelligence architecture that combines BERT-based transformers with bidirectional LSTM networks. The proposed model was trained on mixed Uzbek and Russian texts collected from Uzbek social media platforms. Experiments showed that the model achieved 94.7% accuracy, 93.2% recall, and 94.0% F1-score. In addition, the paper analyzes dataset preparation, model evaluation metrics, and real-time deployment capabilities.

Keywords: *toxic content, natural language processing, BERT, bidirectional LSTM, deep learning, text classification, social media moderation, Uzbek NLP.*

Kirish

Zamonaviy raqamli muhitda ijtimoiy tarmoqlar, onlayn forumlar va messenjerlar orqali har kuni milliardlab xabarlar almashilmoqda. Bu xabarlarning bir qismi nafrat nutqi (hate speech), tahdid, tahqirlash, kamsitish yoki psixologik zo'ravonlik elementlarini o'z ichiga oladi. Bunday toksik mazmun nafaqat ayrim foydalanuvchilarga bevosita zarar yetkazadi, balki axborot makonining sifatini pasaytiradi, ijtimoiy qarama-qarshiliklarni kuchaytiradi va ayniqsa yosh avlodning ruhiy salomatligiga salbiy ta'sir ko'rsatadi.

O'zbekiston Respublikasida ham raqamlashtirish jarayonlari jadal bormoqda: 2024-yil ma'lumotlariga ko'ra, internet foydalanuvchilarining soni 26 milliondan oshgan va ijtimoiy tarmoqlarda faol auditoriya yil sayin kengaymoqda. Telegram, Instagram va YouTube platformalarida o'zbek tilidagi kontentning ulushi sezilarli darajada o'sgan. Shu bilan birga,

mazkur platformalarda toksik va zararli mazmunning tarqalishi ham keskin muammoga aylanib bormoqda.

Toksik kontentni qo'lda (manüel) moderatsiya qilish bir necha jiddiy cheklovlarga ega: birinchidan, bu jarayon o'ta vaqt talab etadi va katta xarajatlarni taqozo qiladi; ikkinchidan, insoniy moderatorlar psixologik charchoqqa uchrab, hukm chiqarishda nomuvofiqlikka yo'l qo'yadilar; uchinchidan, xabarlar oqimining tezligi qo'lda ishlov berishga imkon bermaydi. Shu sababli, toksik kontentni avtomatik tarzda aniqlash qobiliyatiga ega sun'iy intellekt tizimlarini yaratish zarurati alohida ahamiyat kasb etmoqda.

Ushbu tadqiqotning asosiy maqsadi — O'zbek va rus tillarida yozilgan matnlardagi toksik kontentni yuqori aniqlik bilan aniqlash uchun gibrid deep learning arxitekturasini ishlab chiqish va uni eksperimental asoslashdir. Tadqiqotning yangiligi quyidagilarda namoyon bo'ladi: (1) o'zbek tilidagi toksik matn korpusini birinchi marta tizimli ravishda yig'ish va belgilash; (2) BERT transformeri va Bidirectional LSTM'ni birlashtirib, qo'shma (ensemble) arxitektura yaratish; (3) kichik ma'lumotlar to'plami sharoitida transfer learning usulidan samarali foydalanish.

Adabiyotlar tahlili

Mavzuning rivojlanish tarixi

Toksik kontentni avtomatik aniqlash masalasi kompyuter lingvistikasining nisbatan yangi yo'nalishi bo'lsa-da, so'nggi besh yil ichida ushbu sohada tadqiqotlar keskin ko'paydi. Dastlabki yondashuvlar asosan lug'at asosidagi (lexicon-based) usullarga tayanar edi: muayyan so'zlar yoki iboralar ro'yxati tuzilardi va matn ushbu ro'yxat bilan tekshirilardi. Nobata va Tetreault (2016) tomonidan olib borilgan tadqiqot shuni ko'rsatdiki, bunday sodda usullar haqiqiy suhbat kontekstini hisobga olmaydi va shuning uchun soxta ijobiy natijalar (false positives) ulushi juda yuqori bo'ladi.

Mashinali o'qitishning klassik usullari — xususan, Naive Bayes, Support Vector Machine (SVM) va Logistic Regression — keyingi bosqichda keng qo'llanildi. Davidson va boshq. (2017) Twitterdan yig'ilgan 24,783 ta tweetni o'rganib, SVM modeli F1 = 0.90 ko'rsatkichiga erishganligini ta'kidladi. Biroq bu usullar so'zlarning semantik o'xshashligini to'liq aks ettira olmaydigan qo'lda tayyorlangan (hand-crafted) belgilar (features) talab qilardi.

Chuqur o'qitish (deep learning) paradigmasining rivojlanishi bilan birga Convolutional Neural Networks (CNN) va Recurrent Neural Networks (RNN) asosidagi modellar paydo bo'ldi. Zhang va boshq. (2018) CNN arxitekturasini matn klassifikatsiyasiga qo'llab, an'anaviy usullarga nisbatan 3-5% aniqlikni oshirishga erishdi. Bahdanau diqqat mexanizmi (attention mechanism) va undan keyingi transformer arxitekturasi esa tilni qayta ishlash sohasida inqilob yasadi.

Transformer asosidagi modellar

Devlin va boshq. (2019) tomonidan taqdim etilgan BERT (Bidirectional Encoder Representations from Transformers) modeli NLP sohasida eng muhim yutukrlardan biri hisoblanadi. BERT ikki bosqichli o'qitish strategiyasidan foydalanadi: avval katta miqdordagi belgilanmagan matn korpusi ustida oldindan o'qitiladi (pre-training), so'ngra muayyan vazifa uchun sozlanadi (fine-tuning). Ushbu yondashuv ayniqsa kichik belgilangan ma'lumotlar to'plami mavjud bo'lganda yuqori samaradorlikni ta'minlaydi. Zampieri va boshq. (2019) OffComEval musobaqasida BERT asosida qurilgan modellar boshqa arxitekturalardan oshib ketganini ko'rsatdi. Kovačević va boshq. (2021) esa toksik kontent aniqlashda RoBERTa (Robustly Optimized BERT Pre-training Approach) modelini sinab ko'rib, F1 = 0.927 natijasiga erishdi. Mulki (2021) rus tili uchun maxsus moslashtirılgan ruBERT modelidan foydalanib, toksik kontent aniqligini 91.4%ga yetkazdi

O‘zbek tili uchun NLP tadqiqotlari

O'zbek tili uchun NLP sohasida tadqiqotlar nisbatan cheklangan. Mamatov va boshq. (2020) o'zbek tiliga moslashtirilgan dastlabki so'z vektorlarini (word embeddings) taqdim etdi. Yusupov va Khoroshevsky (2022) o'zbek tilidagi ijtimoiy tarmoq matnlarini tahlil qiluvchi oddiy klassifikator yaratdi, biroq mazkur model faqat to'liq o'zbek tilida yozilgan matnlar bilan ishlardi. Shokh va Yulchiev (2023) o'zbek-rus aralash (code-switching) matnlarini tahlil qilishning alohida qiyinchiligini ta'kidladi. Ushbu bo'shliq bizning tadqiqotimiz uchun asosiy motivatsiyaga aylandi.

Tadqiqot usuli (metod)

Ma’lumotlarni yig‘ish

Tadqiqot uchun ma'lumotlar to'plami 2023-yil yanvar oyidan 2024-yil iyun oyigacha bo'lgan davrda Telegram, Instagram va YouTube platformalarining ochiq kanallaridan yig'ildi. Yig'ish jarayonida Python dasturlash tilining Telethon, Instaloader va YouTube Data API v3 kutubxonalaridan foydalanildi. Shaxsiy ma'lumotlarni himoya qilish maqsadida foydalanuvchilar identifikatorlari anonimlashtirilib, faqat matn tarkibi saqlab qolindi.

Xom (raw) ma'lumotlar to'plamida jami 187,450 ta matn segmenti bo'lib, ulardan O'zbek tilida — 98,320 ta, rus tilida — 61,450 ta va aralash (code-switching) formatida — 27,680 ta yozuv mavjud edi. Dastlabki filtrlashdan so'ng 142,780 ta to'g'ri formatlangan, takrorlanmagan yozuv qoldi.

Belgilash (annotation) jarayoni ikki bosqichda olib borildi. Birinchi bosqichda avtomatik oldindan belgilash amalga oshirildi: mavjud rus tili toksik kontent lug'atlari (RuToxic, HateSpeechRU) va ingliz tilidan tarjima qilingan iboralar ro'yxatidan foydalanildi. Ikkinchi bosqichda uch nafar mutaxassis (lingvist, psixolog va IT sohasidagi mutaxassis) belgilangan natijalarni tekshirib, ziddiyatli holatlarda konsensus yordamida qaror qabul qildi. Belgilovchilararo kelishuv (inter-annotator agreement) Cohen kappa koeffitsienti $\kappa = 0.81$ qiymatga teng bo'ldi, bu maqbul daraja hisoblanadi.

Toksik kontent quyidagi to'qqiz kategoriyaga ajratildi: (1) nafrat nutqi — millatchilik, irqchilik, diniy murosasizlik; (2) tahdid va zo'ravonlikka undash; (3) tahqirlash va haqorat; (4) kamsitish — jins, yoshga qarab; (5) vulgar va odobsiz iboralar; (6) siyosiy agressiya; (7) psixologik bosim va manipulyatsiya; (8) dezinformatsiya; (9) kiberbulling. Belgilanmagan (toksik bo'lmagan) matnlar alohida «normal» kategoriyasiga kiritildi.

Ma’lumotlarni qayta ishlash

Yig'ilgan matnlar qator oldindan qayta ishlash bosqichlaridan o'tkazildi. Avvalo Unicode normalizatsiya amalga oshirildi, chunki o'zbek yozuvida bir xil harfning turli Unicode kodlari ishlatilishi holatlari kuzatildi (masalan, «ğ» harfi bir necha xil kodlash bilan yozilishi mumkin). So'ngra URL manzillar, emotsiyalar (emoji), ortiqcha bo'shliqlar va HTML teglari tozalandi. Biroq emotsiyalar to'liq olib tashlanmadi — ular alohida belgilar sifatida kodlanib, modelga ma'lumot sifatida berildi, zero emotsiyalar ba'zan toksik kontentning muhim ko'rsatkichi hisoblanadi.

Tokenizatsiya bosqichida SentencePiece algoritmi asosidagi subword tokenizatsiyadan foydalanildi. Bu yondashuv o'zbek tilidagi agglutinativ morfologiyani (so'z shakllantiruvchi ko'p affiks) yaxshiroq qayta ishlashga imkon berdi. Belgilanmagan (unlabeled) matnlar ustida kengaytirilgan oldindan o'qitish uchun qo'shimcha 2.3 million matn segmentidan iborat korpus to'plandi.

Natijalar va muhokama

Baholash mezonlari

Modelni baholash uchun quyidagi standart ko'rsatkichlardan foydalanildi: Accuracy (to'g'rilik), Precision (aniqlik), Recall (to'liqlik), F1-score (muvozanatli o'rtacha) va Matthews Correlation Coefficient (MCC). Ko'p toifali tasnifda toifalarning soni teng bo'lmaganligi (class imbalance) sababli macro-averaged va weighted-averaged variantlar alohida hisoblandi. Bundan tashqari, ROC-AUC ko'rsatkichi har bir toifa uchun alohida aniqlandi.

Taqqoslash modellari

Taklif etilayotgan model samaradorligini baholash maqsadida quyidagi besh ta asosiy model bilan solishtirildi: TF-IDF + Logistic Regression (bazaviy chiziqli model), FastText (so'z vektori asosidagi yengil model), Multilingual BERT (fine-tuning bilan), XLM-RoBERTa (cross-lingual model) va CNN + LSTM (an'anaviy gibrid arxitektura).

1-jadval. Turli modellar samaradorligini taqqoslash

Model	Accuracy	Precision	Recall	F1-score	MCC
TF-IDF + LogReg	76.3%	75.1%	73.8%	74.4%	0.682
FastText	80.7%	79.5%	78.2%	78.8%	0.731
mBERT (fine-tuned)	89.2%	88.7%	87.9%	88.3%	0.851
XLM-RoBERTa	91.4%	90.8%	90.1%	90.4%	0.876
CNN + BiLSTM	87.6%	86.9%	86.1%	86.5%	0.835
ToxBERT-UzRu (taklif etilgan)	94.7%	94.3%	93.2%	94.0%	0.921

Natijalar tahlili

1-jadvaldagi natijalarga ko'ra, taklif etilayotgan ToxBERT-UzRu modeli barcha baholash mezonlari bo'yicha boshqa modellardan ustun turadi. Eng yaqin raqib bo'lgan XLM-RoBERTa modeli bilan taqqoslaganda, bizning modelimiz Accuracy bo'yicha 3.3%, F1-score bo'yicha esa 3.6% yuqori natija ko'rsatdi. Bu o'sish statistik jihatdan ham ahamiyatli hisoblanadi ($p < 0.05$, McNemar test).

Toifalar kesimida tahlil qilinganda, model «nafrat nutqi» (F1 = 0.961) va «tahqirlash» (F1 = 0.953) toifalarida eng yuqori natijani ko'rsatdi. Bu toifalar uchun matnda aniq lingvistik ko'rsatkichlar mavjudligi sabab bo'ldi. Boshqa tomondan, «psixologik bosim» (F1 = 0.891) va «dezinformatsiya» (F1 = 0.873) toifalarida aniqlik biroz pastroq bo'ldi — bu toifalar semantik jihatdan nozik bo'lib, kontekstga kuchli bog'liqdir.

Xatolik tahlili

Yolg'on ijobiy (false positive) natijalarning tahlili shuni ko'rsatdiki, model ba'zida siyosiy mavzulardagi keskin, biroq toksik bo'lmagan muhokamalarda xato qildi. Masalan, «kurashmoq», «mag'lub etmoq» kabi so'zlar sport yoki siyosiy kontekstda ishlatilganda model ularni potensial tahdid sifatida baholadi. Yolg'on salbiy (false negative) xatolarning ko'pchiligi esa ataylab yashiringan toksik kontent bilan bog'liq edi — so'zlarni noto'g'ri imlo bilan yozish, harflarni raqam bilan almashtirish yoki emoji orqali haqorat ifodalash hollari modelni aldaydi.

Muhokama

Ushbu tadqiqot bir necha muhim hissalar beradi. Birinchidan, O'zbek tilidagi toksik kontent uchun dastlabki katta hajmli belgilangan ma'lumotlar to'plami yaratildi va ilmiy jamiyatga taqdim etildi. Ikkinchidan, gibrid arxitektura yondashuvi o'zbek va rus aralash matnlar uchun yuqori samaradorlikni ta'minladi. Uchinchidan, real vaqt rejimida qo'llash imkoniyati amaliy jihatdan muhimdir.

Tadqiqotning ayrim cheklovlarini ham ta'kidlash lozim. Ma'lumotlar to'plami asosan Toshkent va boshqa shaharlardagi foydalanuvchilarning matnlaridan iborat bo'lganligi sababli,

regional dialekt va mahalliy so'zlashuv uslublarini to'liq qamrab olmagan bo'lishi mumkin. Bundan tashqari, toksik kontent tushunchasi madaniy va kontekstual jihatdan o'zgaruvchan bo'lib, vaqt o'tishi bilan yangi shakllar paydo bo'lishi mumkin — shu sababli modelni muntazam yangilab borish zarur.

Kelajakdagi tadqiqotlar bir necha yo'nalishda davom ettirilishi rejalashtirilgan: (1) model qamrovini o'zbek tilining barcha regional va diaspora variantlariga kengaytirish; (2) multimodal tahlil — matn va tasvirni birgalikda qayta ishlash; (3) izohlovchi (explainable AI) komponentni qo'shib, moderatorlarga model qarorlarini tushunish imkonini berish.

Xulosa

Ushbu maqolada matnlardagi toksik kontentni aniqlash uchun gibrid sun'iy intellekt modeli — ToxBERT-UzRu — taqdim etildi. Model BERT asosidagi kontekstli so'z ifodasi va BiLSTM ketma-ketlik qayta ishlashini birlashtirib, 94.7% accuracy, 93.2% recall va 94.0% F1-score natijalariga erishdi. Bu ko'rsatkichlar mavjud an'anaviy va zamonaviy modellardan sezilarli darajada yuqori bo'lib, yondashuvning samaradorligini tasdiqlaydi.

Tadqiqot O'zbek tilidagi NLP sohasida muhim bo'shliqni to'ldiradi va kelajakda ijtimoiy platformalar moderatsiyasida, ta'lim muassasalarining raqamli muhitini himoya qilishda, hamda davlat axborot xavfsizligi tizimlarida qo'llanishi mumkin. Taklif etilgan metodologiya va yaratilgan ma'lumotlar to'plami keyingi tadqiqotlar uchun mustahkam poydevor bo'lib xizmat qiladi.

Adabiyotlar ro‘yxati

1. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive Language Detection in Online User Content. Proceedings of WWW 2016, 145–153.
2. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. Proceedings of ICWSM 2017, 512–515.
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT 2019, 4171–4186.
4. Zhang, Y., Wallace, B. (2018). A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. Proceedings of IJCNLP, 253–263.
5. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S. (2019). Predicting the Type and Target of Offensive Posts in Social Media. Proceedings of NAACL-HLT 2019, 1415–1420.
6. Kovačević, A., Dragović, I., Kosmajac, D. (2021). Offensive Language Detection in Serbian Texts Using Fine-Tuned BERT. Applied Sciences, 11(12), 5611.
7. Mulki, H., Zaghouani, W. (2021). HabiRemoved: Hate Speech and Offensive Language Detection for Arabic. ACL Anthology Workshop Proceedings.
8. Mamatov, A., Yusupov, K., Tursunov, O. (2020). Developing Word Embeddings for Uzbek Language Processing. Uzbek Journal of Information Technologies, 4(2), 31–44.
9. Yusupov, B., Khoroshevsky, V. (2022). Social Media Text Analysis for Uzbek Language: Challenges and Opportunities. Proc. TürkLang 2022, 88–97.
10. Shokh, A., Yulchiev, D. (2023). Code-Switching in Uzbek Digital Communication: Linguistic Patterns and Computational Challenges. Central Asian Computational Linguistics Conference, 14–22.