BERT MODELI YORDAMIDA MATNLARNI TASNIFLASH

Alayev Ruhillo Habibovich,
Texnika fanlari bo‘yicha falfasa doktori PhD, Dotsent
mr.ruhillo@gmail.com
O‘zMU

Berdiyev Jahongir Botir o‘g‘li,
Kompyuter lingvistikasi mutaxassisligi magistranti
berdiyevjahongir94@gmail.com
ToshDO‘TAU

Annotatsiya. Hozirgi kunda tabiiy tilni qayta ishlash (NLP) sohasi sun’iy intellektning eng jadal rivojlanayotgan yo‘nalishlaridan biri hisoblanadi. Bu sohada erishilayotgan yutuqlar turli tildagi matnlarni avtomatik tarzda tahlil qilish, tushunish va tasniflash imkonini bermoqda. Ayniqsa, chuqur o‘rganishga asoslangan modellar, xususan, BERT (Bidirectional Encoder Representations from Transformers) modeli ushbu jarayonlarda samaradorlikni sezilarli darajada oshirdi. Google kompaniyasi tomonidan ishlab chiqilgan ushbu model til birliklarini kontekstual tarzda, ya’ni ikki tomonlama yo‘nalishda tahlil qilish orqali yuqori aniqlikka erishadi. Mazkur maqolada BERT modelidan foydalanilgan holda matnlarni muayyan kategoriyalar bo‘yicha tasniflash masalasi ko‘rib chiqiladi. Tadqiqotda modelni o‘qitish, sinovdan o‘tkazish va baholash bosqichlari amalga oshiriladi. Natijalar asosida BERT modelining matn tasnifi sohasidagi samaradorligi tahlil qilinadi va tegishli xulosalar chiqariladi.

Abstract. Currently, Natural Language Processing (NLP) is one of the most rapidly developing fields within artificial intelligence. Advances in this domain enable automatic analysis, comprehension, and classification of texts in various languages. Deep learning models, particularly BERT (Bidirectional Encoder Representations from Transformers), have demonstrated high effectiveness in this regard. Developed by Google, BERT analyzes linguistic units in a contextual manner, processing information bidirectionally, which contributes to its high accuracy. This research focuses on the task of classifying texts into specific categories using the BERT model. The study involves the implementation of model training, testing, and evaluation stages. Based on the obtained results, the effectiveness of BERT in text classification tasks is analyzed, and relevant conclusions are drawn.

Аннотация. В настоящее время обработка естественного языка (NLP) является одним из наиболее динамично развивающихся направлений искусственного интеллекта. Достижения в этой области позволяют автоматически анализировать, понимать и классифицировать тексты на

различных языках. Особенно эффективно в этом контексте показали себя модели глубокого обучения, в частности, модель BERT (Bidirectional Encoder Representations from Transformers). Разработанная компанией Google, данная модель анализирует языковые единицы в контекстуальном виде, то есть в двух направлениях, что обеспечивает высокую точность. В данной научной работе рассматривается задача классификации текстов по определённым категориям с использованием модели BERT. В ходе исследования реализуются этапы обучения модели, тестирования и оценки. На основе полученных результатов проводится анализ эффективности модели BERT в задачах текстовой классификации и формулируются соответствующие выводы.

Katil so‘zlar. *Tasniflash, tokenizatsiya, matnlarni tozalash, tokenizator, fine tuning.*

Bilamizki, ko‘p resursli tillarda, jumladan, ingliz tilida til modellarida bajariladigan vazifalar samarali ancha samarali natijalar bermoqda. Bu esa bir tomondan tillarning ko‘p resursliligi bo‘lsa, boshqa tomondan ularning morfologik qurilishi bilan bog‘liq. Kam resursli tillarni modellashtirish esa yetarlicha aniq natija bermaydi. Kam resursli til bo‘lishiga qaramay, hozirgi kunga kelib o‘zbek tili ham modellashtirishda yaxshi natijalar bermoqda. Ammo shunga qaramay ma’lumotlarning kamligi hamda o‘zbek tilining morfoligik qurulish jihatdan agglyutinativ til ekanligi o‘zbek tilini modellashtirishda ba’zi muammolarni keltirib chiqaradi. Chunki ingliz tili kabi flektiv tillardan farqli ravishda o‘zbek tili kabi agglyutinativ tillarning so‘z tartibida aniqlikning yo‘qligi, qo‘shimchalarning qo‘shilish tartibi hamda so‘zga qo‘shimcha qo‘shish natijasida asos va qo‘shimchada tovush o‘zgarish hodisalari mavjud, bu esa o‘zbek tilini modellashtirishda qiyinchiliklarni keltirib chiqaradi. O‘zbek tilining boshqa tillardan farqini quyidagicha farqlash mumkin:

1. **Agglutinativlik:** O‘zbek tili so‘z va so‘z shakli yasashda qo‘shimchalar orqali yangi ma’nolar hosil qiladi. Masalan, “kitob” so‘zidan “kitobxon”, “kitoblar” kabi shakllar hosil bo‘ladi. Bu holat esa matnlarni tokenlash jarayonida ayrim muammoli vaziyatni keltirib chiqaradi. Chunki tokenlash jarayonida har bir so‘zni yoki belgini bitta token sifatida belgilash juda katta xotira talab qiladi. Buning oldini olish uchun tokenlarga ajratishda ko‘p qo‘llaniladigan belgilar va so‘zlar tokenlashtiriladi. Shundan kelib chiqib, “kitob”, “kitobxon”, “kitoblar” so‘zlarining har birini emas, balki faqat “kitob” so‘zini token sifatida ajratish samarali hisoblanadi.

2. Lug‘at tarkibi: O‘zbek tilida arabcha, forscha, ruscha va boshqa tillardan olingan so‘zlar ko‘p. Bu esa sof o‘zbek tilidagi so‘zlarni ajratishni qiyinlashtirsa, ikkinchi tomondan gap qurilishi va so‘z yasalishi, ularga qo‘shimcha qo‘shilishdagi

ba’zi istisnolarni keltirib chiqaradi. Bu esa o‘z navbatida o‘zbek tilidagi matnlarni qayta ishlashdagi jarayonini murakkablashtiradi.

3. Sintaksis: Gap tuzilishi ba’zan boshqa tillardan sezilarli darajada farq qiladi.

Agar BERT modeli O‘zbek tiliga moslashtirilsa, u o‘zbek tilidagi matnlarni yaxshiroq tahlil qilishi, tarjima qilish, savol-javob tizimlarini yaxshilash va boshqa ko‘plab sohalarda qo‘llanilishi mumkin bo‘ladi.

BERT modelini O‘zbek tiliga moslashtirish jarayoni bir necha qadamdan iborat. Quyida ular batafsil tushuntiriladi.

Ma’lumotlarni to‘plash. Ma’lumotlarni to‘plash va katta hajmli ma’lumotlar bazasi NLP muammolarini yechishning samarali yo‘lidir. Har qanday til modellari beradigan natijalarining aniqligi ma’lumotlar bazasining hajiga bog‘liq. Jumladan, BERT Modelini o‘zbek tiliga o‘qitish uchun ham katta hajmdagi ma’lumotlar to‘plami kerak. Bu ma’lumotlar O‘zbek tilidagi kitoblar, maqolalar, veb-saytlar, ijtimoiy tarmoqlar postlari va boshqa matn manbalaridan olinadi. Biz ma’lumotlar to‘plami sifatida *daryo.uz* veb saytidagi maqolalarni kategoriyalar bo‘yicha ajratib to‘pladik va ushbu maqolalardan ma’lumotlar to‘plami sifatida foydalandik. Ma’lumotlar to‘plamini tayyorlashda e’tibor qilish kerak bo‘lgan jihati shundaki, matnlarni tasniflash uchun yig‘ilgan ma’lumotlar bazasi tarkibidagi kategoriyalar hajmi jihatidan bir biriga teng bo‘lishi kerak. Agar kategoriyalar orasidagi muvozanat buzilsa, dastur bir tomonga og‘ib ketishi va dastur xato ishlay boshlashi mumkin.

Matnlarni tozalash. Bu ma’lumotlar dasturlar to‘g‘ri va sifatli ishlashi uchun, avvalo, uni tozalab olish kerak. Matnlarni tozalashning bir qancha usullari mavjud. Matnlarni avtomatik tozalash uchun maxsus dasturlardan foydalanish mumkin, yoki inson qo‘l mehnati bilan tozalash mumkin. Qo‘l mehnati bilan tozalash katta hajmli matnlar uchun juda ko‘p vaqt va xarajat talab qiladi. Ammo sifatli tozalash imkonini beradi.

Biz matnlarni tozalash uchun o‘zbek tilidagi matnlarni tozalovchi dasturdan foydalandik. Tozalash jarayonida HTML kodlari, reklama matnlari, turli xil emojilar olib tashlandi va takroriy jumlalarning faqat bittasi olib qolindi.

Tasniflash jarayoni. Keyingi bosqich matnlarni tasniflash jarayonidir. Matnlarni tasniflashning bir qancha turi bor. Matnni tasniflash ijtimoiy media postlarining kayfiyatini tahlil qilishda, ayniqsa, nafrat nutqi kabi salbiy his-tuyg‘ularni aniqlashda ham qimmatlidir. Mashinani o‘rganish modelidan foydalanib, matnni haqoratli til va nafrat nutqi uchun tasniflash va kuzatish mumkin[1]. Biz matnlarni turli kategoriyalarga ajratishni ko‘rib chiqamiz. Buning uchun to‘plangan matnlarni kategoriyar bo‘yicha ajratib olish kerak. *Daryo.uz* veb saytidan to‘plangan

maqolalar quyidagi 15ta kategoriyaga ajratildi va tasniflash uchun tayyorlandi: “Madaniyat”, “Qonunchilik”, “Sport”, “Salomatlik”, “Dunyo”, “O‘zbekiston”, “Jamiyat”, “Foto”, “Ayollar”, “Texnologiya”, “Siyosat”, “Iqtisodiyot”, “Pazandachilik”, “Jinoyat”, “Avto”.

Matnlarni tasniflash uchun ishlatiladigan matnlar to‘plamining hajmi juda katta bo‘ladi. Bu esa matnlarga ishlov berishni qiyinlashtiradi. Bundan tashqari, BERT modeli o‘ta murakkab struktura bo‘lgani uchun katta hisoblash resurslariga muhtoj. Shuning uchun matnlarni tasniflashda GPU (Grafik protsessor)dan foydalanildi. Quyidagi rasmda buni ko‘rish mumkin(2.2.1-rasm):

```
import tensorflow as tf

gpus = tf.config.experimental.list_physical_devices('GPU')
if gpus:
    try:
        for gpu in gpus:
        tf.config.experimental.set_memory_growth(gpu, True)
    print("GPU aniqlangan va ishlatilmoqda.")
 except RuntimeError as e:
     print(e)
else:
    print("GPU topilmadi. CPU ishlatilmoqda.")
```

1-rasm. GPU (grafik protsessor)ning mavjudligini tekshiruvchi va undan foydalanishga ruxsat beruvchi kod

Ma’lumotlarni BERT modeli uchun mos formatga keltirish uchun oldindan qayta ishlash funksiyasini (preprocess_data) aniqlaymiz. Bu funksiya matnlarni tokenlashtiradi, padding qo‘shadi va yorliqlarni kodlaydi.

BERT modeli orqali tokenlash uchun quyidagi kodni ko‘rib chiqish mumkin:

```
from transformers import AutoTokenizer
tokenizer = AutoTokenizer.from_pretrained("bert-base-multilingual-cased")
```

Bu yerda AutoTokenizer BERT modeliga mos keladigan mos keladigan tokenizatorni avtomatik yuklash uchun ishlatiladi. AutoTokenizer transformer modellari uchun maxsus ishlab chiqilgan bo‘lib, u turli modellarga mos ravishda **tokenizatsiya** jarayonini amalga oshiradi. Tokenizatsiya bu matnni kichik qismlarga ajratish jarayonidir. Transformer modellar so‘zlarni emas, balki **tokenlar** bilan ishlaydi, shuning uchun har bir matn modelga tushunarlilik darajasida kodlanishi kerak. AutoTokenizer orqali tokenizatsiya qilingan matn maxsus **token**

identifikatorlari orqali ifodalanadi. U turli xil tokenizatorlarni qo‘llab-quvvatlaydi va ishlayotgan modelga mos ravishda eng samarali usulni tanlaydi.

tokenizer = AutoTokenizer.from_pretrained(“bert-base-multilingual-cased”) – bu kod ko‘p tilli BERT modeli uchun oldindan o‘qitilgan tokenizatorni yuklaydi (“**bert-base-multilingual-cased”** modeli **100dan ortiq tilni,** jumladan, **o‘zbek tilini** ham qo‘llab-quvvatlaydi). Ushbu tokenizator matnlarni qismlarga bo‘lib, IDlarga o‘giradi. “**Cased” versiya** bo‘lgani uchun **“bert-base-multilingual-cased”** katta-kichik harflarni farqlaydi.

Tokenizatsiyada **WordPiece** yoki **SentencePiece tokenizatsiyalaridan foydalaniladi. Lekin bu yerda** bert-base-uncased modeli ishlatilganligi uchun WordPiece tokenizatsiyasi ishlatilgan. Chunki **bert-base-multilingual-cased modeli WordPiece tokenizatsiyasidan foydalanadi. Bundan tashqari, o‘zbek tilidagi matnlarni tasniflashda WordPiece tokenizatsiyasi samarali hisoblanadi.**

O‘zbek tilidagi matnlarni tasniflashda WordPiece tokenizatsiyasi samarali bo‘lishi mumkin, lekin uning natijaviyligi matnning o‘ziga xos tuzilishi va so‘zlarning murakkabligi kabi omillarga bog‘liq bo‘ladi. Ba’zi hollarda SentencePiece ham samarali bo‘lishi mumkin, ayniqsa, o‘zbek tilining agglutinativ xususiyatlarini hisobga olsak.

Shuni ham ta’kidlash kerakki, o‘zbek tili agglutinativ til bo‘lgani uchun so‘zlarni qismlarga ajratish juda muhim. Chunki ishlatilayotgan model har bir so‘z shaklini alohida lug‘at birligi sifatida saqlasa, juda ko‘p xotira talab qiladi. Bu esa nafaqat joy, balki mablag‘ sarfini ham oshiradi. Shu sababli modellar, asosan, lug‘at sifatida so‘zlarning asosiy qismini saqlaydi. Tokenizatsiya jarayonida esa model lug‘atida mavjud bo‘lmagan, ammo unga o‘xshash so‘zlarni alohida tokenlarga ajratib hisoblaydi. Masalan, “kitoblarimizdan” so‘zi “kitob”, “lar”, “imiz” va “dan” qismlariga ajratiladi. Model lug‘atida esa faqat “kitob” so‘zi saqlanadi.

Shundan kelib chiqib, o‘zbek tili uchun WordPiece tokenizatsiyasidan foydalanish samarali ekani ta’kidlanadi. Ushbu tokenizatsiya so‘zlarning asos qismini lug‘at sifatida ajratib, qo‘shimchalarni alohida token sifatida tahlil qiladi.

Odatda, til modellari kirish ketma-ketligini bir yo‘nalishda o‘qiydi: chapdan o‘ngga yoki o‘ngdan chapga. Maqsad keyingi so‘zni bashorat qilish/yaratish bo‘lsa, bunday bir yo‘nalishli trening yaxshi ishlaydi. Ammo til kontekstini chuqurroq anglash uchun BERT ikki tomonlama treningdan foydalanadi. Ba’zan u “yo‘nalishsiz” deb ham ataladi. Shunday qilib, u bir vaqtning o‘zida oldingi va keyingi tokenlarni hisobga oladi. BERT Transformatorning ikki tomonlama treningini tilni modellashtirishda qo‘llaydi, matn tasvirlarini o‘rganadi[2]. BERT modelini maxsus vazifalar uchun moslashtirishda ikkita yondashuv mavjud: modelni noldan oldindan o‘qitish yoki oldindan o‘qitilgan modeldan foydalanib, uni fine-tuning qilish. Modelni noldan oldindan o‘qitish katta hajmdagi ma’lumotlar to‘plami va yuqori hisoblash

quvvatiga ega qurilmalarni talab qiladi. Ushbu jarayon vaqt va resurs jihatidan ancha talabchan bo‘lgani uchun, odatda, maxsus vazifalar, jumladan, matnlarni tasniflash kabi masalalar uchun oldindan o‘qitilgan BERT modellaridan foydalaniladi. Keyin esa ushbu model fine-tuning orqali yangi ma’lumotlarga moslashtiriladi, ya’ni modelning barcha yoki ba’zi qatlamlari qayta o‘qitilib, tasniflash vazifasi uchun optimallashtiriladi.

Maxsus vazifalar uchun moslashtirish(fine-tuning). BERT kirish matnini yuqori o‘lchamli bo‘shliqda aks ettirish uchun ko‘p qatlamli ikki tomonlama transformator kodlovchisidan foydalanadi. Ya’ni, u jumladagi har bir so‘zning butun kontekstini hisobga olishi mumkin, bu esa matnning ma’nosini yaxshiroq tushunishga yordam beradi[3]. Oldindan o‘qitilgan **BERT model**ini o‘zimizga kerakli vazifaga moslashtirish uchun **fine-tuning** amalga oshiriladi. Quyidagi kodda “**bert-base-multilingual-cased**” modeli yuklanib, **ko‘p tegli klassifikatsiya (multi-label classification)** muammosi uchun moslashtiriladi. Model **yorliqlar soniga (num_labels)** qarab sozlanadi va har bir yorliqning **id bilan mosligi (id2label, label2id)** belgilanadi:

```
from transformers import AutoModelForSequenceClassification
model = AutoModelForSequenceClassification.from_pretrained(
    "bert-base-multilingual-cased",
    problem_type="multi_label_classification",
    num_labels=len(labels),
    id2label=id2label,
    label2id=label2id
)
```

Bu jarayon **oldindan o‘qitilgan modelni maxsus vazifaga moslashtirishga** yordam beradi va **modelni fine-tuning qilishga tayyorlaydi.** trainer.train() **kodi orqali fine-tuning qilish boshlanadi va o‘quv ma’lumotlari asosida sozlanadi.**

Maxsus vazifalar uchun moslashtirish (fine-tuning) jarayoni tugagandan so‘ng, modelning natijalari baholanadi. Bunda o‘qitilgan modelning samaradorligini tekshirish uchun **sinov (test) ma’lumotlari** ustida baholash amalga oshiriladi. Baholash jarayonida **F1-score, ROC-AUC va aniqlik (accuracy)** kabi metrikalar qo‘llanilib, modelning tasniflash sifati aniqlanadi. Agar model natijalari kutilgan darajada bo‘lmasa, **giperparametrlarni sozlash (hyperparameter tuning) yoki ma’lumotlarni kengaytirish (data augmentation)** kabi usullar bilan natijalarni yaxshilash mumkin.

XULOSA

Xulosa qilib shuni aytish mumkinki, BERT modelining yaratilishi tabiiy tilni qayta ishlashda katta inqilob bo‘ldi. Shuningdek, BERT modelining oldindan o‘qitilganligi turli vazifalar uchun moslashtirishni osonlashtiradi. Shuningdek, BERT modeli matnlarni tasniflash, his-tuyg‘ularni tahlil qilish, xabarlarning rost yoki yolg‘onligini aniqlash kabi vazifalarda samarali vosita hisoblanadi. Bundan

tashqari, BERT yordamida tokenizatsiya qilish uchun turli tillarning grmmatik qurilishi uchun mos ravishda tokenizatorlar mavjudligi bilan ham bu model qulay hisoblanadi.

Foydalanilgan adabiyotlar:

1. Khang Pham (2023). Text classification with BERT
2. Shre Varsheni (2024). Why and how to use BERT for NLP Text Classification? *Analytics Vidhya*
3. https://medium.com/@khang.pham.exxact/text-classification-with-bert-7afaacc5e49b