

BERT MODELINING QO‘LLANILISHI VA VAZIFALARI

Berdiyev Jahongir Botir o‘g‘li

Alisher Navoiy nomidagi

Toshkent davlat o‘zbek tili va adabiyoti universiteti,

kompyuter lingvistikasi yo‘nalishi magistranti.

Annotatsiya: Til modellari mashinaga tabiiy tillarni yaxshiroq tushunishga yordam beradigan vosita hisoblanadi. Hozirgi kungacha bir qancha til modellari yaratilgan va ularning barchasi ma‘lum bir vazifani bajarish uchun mo‘ljallangan. Shunday modellardan biri BERT modelidir. Bu model boshqa modellardan o‘zining afzalligi, yutuqlari, imkoniyatlari bilan ajralib turadi. Ushbu til modeli matnlar ustida ko‘plab amallar bajarish va juda aniq darajadagi natijalarni taqdim eta olishi bilan mashhurdir. Misol tariqasida matnlardagi va mediadagi hissiyotlarni aniqlash, matnlarni va tasvirlarni tasniflash, mashina tarjimasi, tillarni va nutqni aniqlash, spamni aniqlash, savollarga javob berish, matn yaratish kabi ko‘plab amaliyotlarni bajarishda muhim vosita ekanligini ta‘kidlab o‘tish joiz.

Mazkur maqolada BERT modelining yaratilish tarixi, bu modelning yaratilishiga bo‘lgan ehtiyoj, uning imkoniyatlari haqida fikr yuritiladi. Bundan tashqari, BERT modeli qanday ishlaydi, BERT qanday o‘qitiladi, BERT modeli o‘zi nima kabi savollarga ham javob berilgan. Aytish mumkinki, BERT modelining qo‘llanilishi tabiiy tilga ishlov berishda samarali natijalarni beradi.

Kalit so‘zlar: BERT, tasniflash, mashina tarjimasi, mashinali o‘qitish

Abstract: Language models are tools that assist machines in comprehending natural languages. Several language models have been created to date, each designed to perform a specific task. One such model is the BERT model, which is distinguished from others by its advantages, achievements, and capabilities. This language model is renowned for its ability to perform various operations on texts and provide highly accurate results. As an example, it is important to note that BERT is a valuable tool for various practices, including emotion detection in texts and media, text and image classification, machine translation, language and speech recognition, spam detection, question answering, and text creation. This article discusses the history of BERT’s creation, the need for its development, and its potential applications. Additionally, this text answers questions about the functionality, training, and purpose of the BERT model. It can be concluded that implementing the BERT model yields effective results in natural language processing.

Keywords: BERT, classification, machine translation, machine learning

Аннотация: Языковые модели — это инструменты, которые помогают машинам лучше понимать естественные языки. На сегодняшний день создано несколько языковых моделей, и все они предназначены для выполнения конкретной задачи. Одной из таких моделей является модель BERT. Данная модель отличается от других моделей своими преимуществами, достижениями и возможностями. Эта языковая модель известна своей способностью выполнять множество операций над текстами и предоставлять очень точные результаты. В качестве примера следует отметить, что это важный инструмент во многих

практиках, таких как обнаружение эмоций в текстах и медиа, классификация текстов и изображений, машинный перевод, распознавание языка и речи, обнаружение спама, ответы на вопросы, создание текста.

В данной статье рассматривается история создания модели BERT, необходимость создания этой модели и ее возможности. Кроме того, даются ответы на такие вопросы, как, как работает модель BERT, как обучается BERT и что такое модель BERT. Можно сказать, что применение модели BERT дает эффективные результаты при обработке естественного языка.

Ключевые слова: BERT, классификация, машинный перевод, машинное обучение

Inson mavjudki, u doim rivojlanib, yangilikka intilib yashaydi. Inson qiyinchilik ko'rgani sari o'zi uchun yengillik yaratishga, yaxshi yashash uchun o'ziga qulay imkoniyatlar yaratishga intilaveradi. Bu esa inson oldiga ko'plab vazifalar, maqsadlar qo'yadi. Har bir inson o'z imkoniyati va qobiliyatidan kelib chiqib qaysidir sohani rivojlantirishga intiladi. Kimdir tabiatshunoslik bilan shug'ullansa, yana kimdir sanoatni rivojlantirishga hissa qo'shadi. Xuddi shunday sohalardan biri kompyuter texnologiyalari sohasidir. Kundan kunga ushbu sohaga bo'lgan ehtiyoj tobora ortib bormoqda. Bu esa ko'plab sohalarda bo'lgani kabi tilshunoslik sohasiga ham o'z ta'sirini o'tkazmoqda. Buning natijasi o'laroq tabiiy tilni qayta ishlashda ham ko'plab izlanishlar, tadqiqotlar yuzaga keldi. Jahon tajribasida bu bo'yicha samarali ishlar amalga oshirilgan bo'lsa-da, yurtimizda bu sohada yetarlicha tadqiqotlar amalga oshirilmagan. Shunday ekan, tabiiy tilni qayta ishlash uchun eng mashhur til modellaridan biri BERT modeli orqali tabiiy tilga ishlov berish eng samarali natijalarni beradi.

Jahonda til modellarini yaratish va ulardan foydalanish bo'yicha ko'plab tadqiqotlar olib borilgan. Jumladan, BERT modelining yaratilishi tabiiy tillarni qayta ishlashda yangi bosqich bo'ldi. Ushbu model Google kompaniyasining tadqiqotchilari, xususan, uning Brain jamoasi (Y.Devling, M.V. Chang, K.Li, K.Toutanova) tomonidan 2017-yilda ishlab chiqilgan va 2018-yilda ochiq manbaali transformator sifatida e'lon qilinadi.¹ Shuningdek, jahonda BERT modelidan foydalanib matnlar bilan bog'liq turli xil ishlar amalga oshirildi. Jumladan, D.Noza va boshqalar², M.Baz va boshqalar³, N.Baratalipour va boshqalar⁴, Y.SH. Chuang⁵, P.Jain va boshqalar⁶, M.Ashraf va boshqalar⁷, D.Kiela va boshqalar⁸, A.S.Talaat⁹, Z.Guo va M.L.Nguyen¹⁰,

¹ Devling J., Chang M.W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J.Devling, M.W.Chang, K.Lee, K.Toutanova - Proceedings of NAACL-HLT 2019, pp. 4171-4186

² <https://towardsdatascience.com/sentiment-analysis-and-emotion-recognition-in-italian-using-bert-92f5c8fe8a2>

³ Baz M., Sami A., Prtash N., Kowsher Md, Murad S., Bairagi A., Masud M. (2022). Transfer Learning for Sentiment Analysis Using BERT Based Supervised Fine-Tuning. Sensors. 22. 10.3390/s22114157.

⁴ Baratalipour N., Suen Y., Ormandjieva O. (2020). Abusive Language Detection Using BERT Pre-trained Embedding. 10.1007/978-3-030-59830-3_60.

⁵ <https://arxiv.org/abs/2307.15331>

⁶ Jain P., Taneja K., Taneja H. (2024). Advertisement Image Classification Using Deep Learning with BERT: A Novel Approach Exploiting Textual Features. 10.1007/978-981-99-6547-2_34.

⁷ Ashraf M., Zhou X., Vivone G., Chen L., Chen R., Majdard R. (2024). Spatial-Spectral BERT for Hyperspectral Image Classification. Remote Sensing. 16. 539. 10.3390/rs16030539.

⁸ <https://arxiv.org/abs/1909.02950>

⁹ <https://rdcu.be/dzKUK>

¹⁰ Guo Z., Nguyen M.L. / Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10 th International Joint Conference on Natural Language Processing: Student Research Workshop, pages 101-107 December 4 - 7, 2020.

T.Patzak¹¹, I.V.Selivanova¹², R.Buoy va boshqalar¹³ kabi tadqiqotchilar BERT modeli yordamida hissiyotlar tahlili, tasvirlar va matn tasniflari, tarjima jarayonlari, kiritilgan matnning qaysi tilga tegishli ekanligini aniqlovchi va shu kabi matnlar ustida turli xil amallar bajaruvchi bir qancha dasturlar yaratgan.

O'zbek tadqiqotchilari tomonidan ham BERT modeli asosida olib borilgan bir qancha ishlarni ko'rishimiz mumkin. Masalan, matnlarni tasniflash uchun E.Kuriyozov va boshqalar¹⁴ tomonidan yaratilgan BERTbek modelini ko'rishimiz mumkin.

Tabiiy tillarni qayta ishlash sohasidagi ilk qadamlar XX asrning 50-yillaridan boshlangan va bu hozirgi kunigacha bir qancha bosqichlarni bosib o'tgan. Tabiiy tilni qayta ishlash uchun harakatlarning boshlanganiga yarim asrdan oshdi va bu davr mobaydida tabiiy tilni qayta ishlash uchun ko'plab til modellari yaratildi. Bu til modellarining barchasi o'z vaqtida ma'lum bir yutuqlarga erishgan bo'lsa-da, shunga yarasha kamchiliklardan ham holi emas edi. BERT modelining yaratilishi esa tabiiy tilni qayta ishlash borasidagi ko'plab muammolarni hal qildi hamda tabiiy tilni qayta ishlashda boshqa modellarga nisbatan ancha yaxshi natijalarni bera boshladi.

BERT tabiiy tilni qayta ishlovchi til modeli bo'lib, u 2017-yilda Google tadqiqotchilari jamoasi tomonidan ishlab chiqilgan va 2018-yilda ochiq manbaa sifatida ommaga e'lon qilingan. Uning kengaytmasi Bidirectional Encoder Representations from Transformers. Tarjimasi transformatorlardan ikki tomonlama kodlovchi tasvirlar deyiladi. Ikki tomonlama kodlovchi deganda model kiritilgan ma'lumotlarni chapdan o'ngga va o'ngdan chapga harakatlangan holda o'qiydi. Katta hajmdagi oldindan o'qitilgan matnlar GPT-2, BERT, RoBERTA, ALBERT kabi keng tarqalgan ko'plab til modellari yordamida tabiiy tillarni tushunish vazifalarida qo'llaniladi. Ular orasida BERT eng samarali hisoblanadi[Guo Z. va boshqalar, 2020: 2]. Misol tariqasida BERT modelini GPT bilan solishtirishimiz mumkin. GPT BERTdan farqli o'laroq ma'lumotlarni bir tomonlama, ya'ni chapdan o'ngga harakatlanib o'qiydi, ya'ni so'zning faqat chap tomonini hisobga oladi. BERT modeli esa kontekstdagi butun so'zlar to'plamidan foydalanadi(1-rasmga qarang). Shuning uchun BERTni yo'nalishsiz harakatlanuvchi model ham deyishadi. Buni quyidagi rasmda ko'rishimiz mumkin:

BERT o'zgaruvchan neyron tarmoq modellari asosida yaratilgan katta miqdordagi axborotni qayta ishlaydigan modellardan biridir. "Bu model orqali matnni tasniflash, his-tuyg'ularni tahlil qilish, spamni aniqlash, savollarga javob berish kabi vazifalarni yaratish uchun BERT modeli arxitekturasiga faqatgina ba'zi o'zgarishlarni kiritish kifoya qiladi[Berdiyev J., 2024: 176]." BERT zamonaviy neyron tarmoq modeli bo'lib, ma'lumotni tekshirish va belgilashni bir qadamda bajaradi. Bu esa ma'lumotlar ustida ko'plab amallar bajarishga imkon beradi. Jumladan, BERT modeli hissiyotlarni tahlil qilish, savollarga javob berish, tasvirlarni va matnlarni tasnif qilish, mashina tarjimasini va shu kabi ko'plab amallarni bajara oladi. Misol uchun, biz BERT

¹¹ Patzak T. Automatische Klassifikation von Alltagsszenarien mit einem BERT-basierten Siamese-Netzwerk. T.Patzak - Humboldt-Universität zu Berlin, 2021.

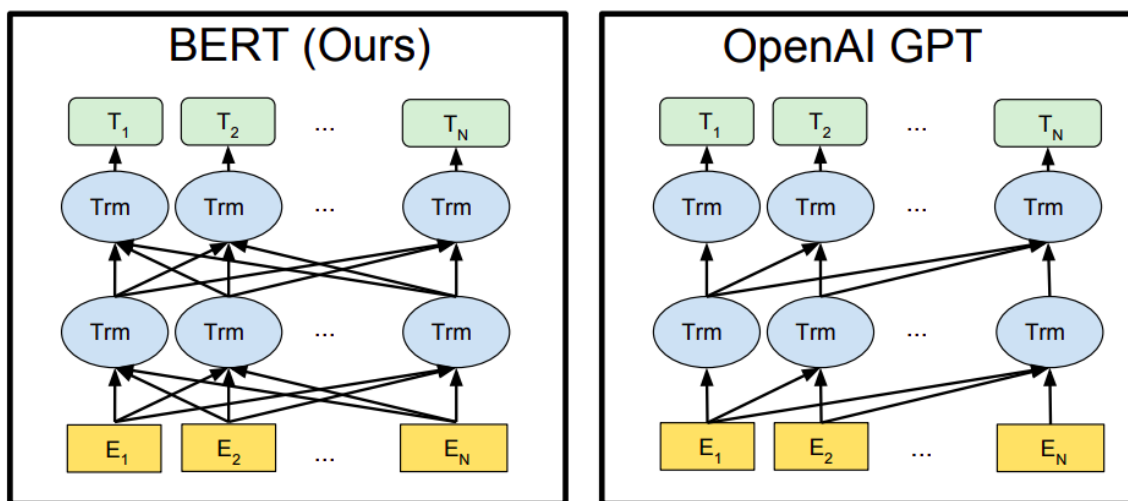
¹² Селиванова И.В. Методы тематической классификации научных текстов на основе теоретико-информационного подхода

¹³ <https://arxiv.org/abs/2112.06748>

¹⁴

<https://www.researchgate.net/publication/368877295> Text classification dataset and analysis for Uzbek language

modelini birinchi navbatda Vikipediya (maskali LM vazifasi) kabi korpusda o'rgatamiz va keyin bir nechta qo'shimcha qatlamlarni qo'shish orqali sharhlarni salbiy, ijobiy yoki neytral deb tasniflash kabi vazifasini bajarish uchun modelni o'z ma'lumotlarimizga moslashtiramiz. Amalda, biz faqat tasniflash vazifasi uchun [CLS] tokenidan foydalanamiz[1]. Bundan tashqari BERT modeli orqali mashina tarjimai yaratish ham juda yaxshi samara beradi va tarjima aniqligini oshiradi. Yuqorida ta'kidlaganimizdek, BERT modeli oldindan o'qitilgach, boshqa vazifalar singari mashina tarjimai uchun ham BERT modelini o'rgatish uchun bir nechta tillardagi katta hajmli matnlarda oldindan tayyorlanadi. Oldindan tayyorgarlik jarayonida BERT modeli jumlar ichida yetishmayotgan so'zlarni bashorat qilishni o'rganadi, turli tillarning grammatikasi, sintaksisi va semantikasini samarali o'rganadi. Bundan tashqari, nozik sozlash usullari oldindan tayyorlangan BERT modelini aniq tarjima vazifalariga moslashtiradi va uning ishlashini aniq va ravon tarjimalar uchun optimallashtiradi.



1-rasm. BERT va GPT modellarning ishlash sxemasi

BERT modelining asosiy xususiyati matn ustida amallarni bajarishda kontekstdagi so'zlarning butun to'plamidan foydalanishidir. Boshqa til modellari esa matnning chap yoki o'ng tomonidagi so'zlardan foydalangan holda natijani taqdim etadi. BERT modelining boshqa til modellaridan yana bitta farqi shuki, bu model bir vaqtning o'zida ikkita vazifa bo'yicha o'qitiladi, ya'ni keying jumlani bashorat qilish va yetishmayotgan tokenni yaratish. Buni D.Samuel va boshqalar shunday tushuntiradi: "Asl BERT modeli ikkita o'z-o'zini boshqaradigan o'quv maqsadi - niqoblangan tilni modellashtirish (MLM) va keyingi jumlani bashorat qilish (NSP) orqali o'qitiladi"[Samuel D. Va boshqalar, 2023: 1958]. Buni Googledan foydalangan paytimizda qidiruv uchun jumlani yozish jarayonida tizim biz uchun taxminiy tavsiya berishida ko'rishimiz mumkin.

🌐 tensorflow.org/guide/basics

🌐 TensorFlow asoslari | TensorFlow yadrosi - tensorflow.org/guide/basics

🔍 tens - Поиск Google

🕒 tensorflow qo'llanmasi

🕒 tensorflowni o'rnatish

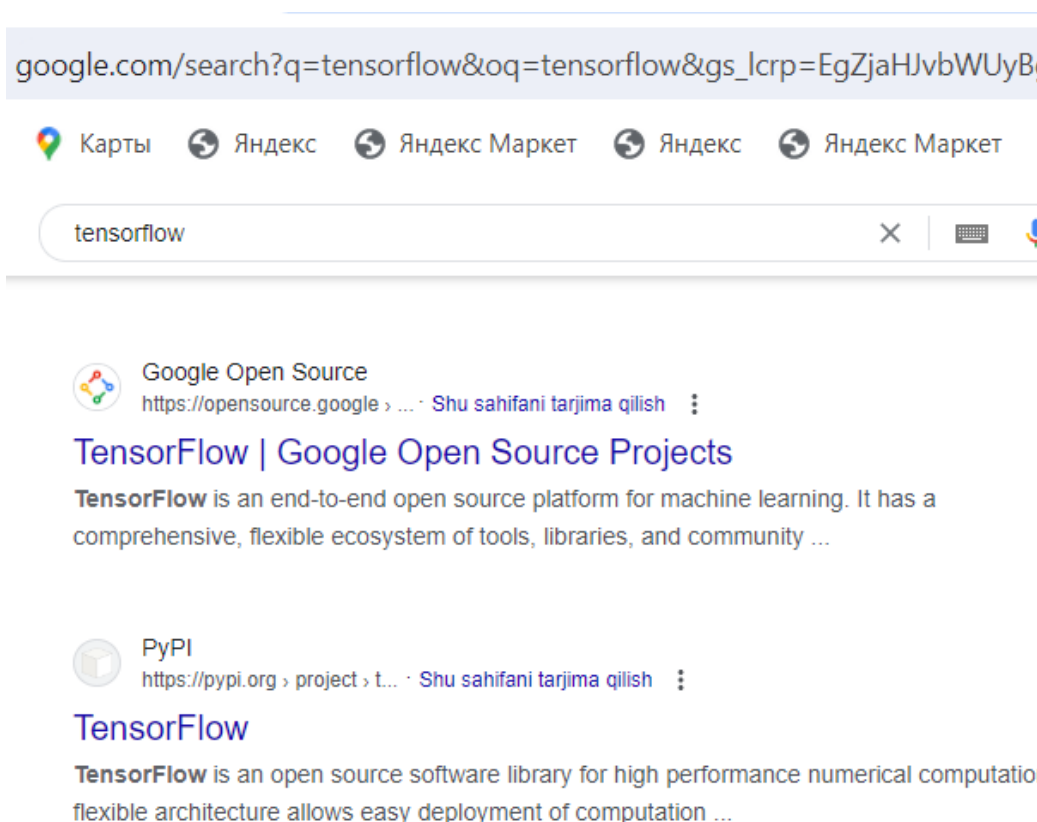
🌐 BERT-for-Text-Classification-with-TensorFlow/Fine_Tune_BERT_for_Text_Classification_with_Tens... - gi

🌐 BERT | bilan matnni tasniflash Matn | TensorFlow - tensorflow.org/text/tutorials/classify_text_with_ber

🌐 TensorFlowni qanday o'rnataman? - Zamonaviy texnologiyalar 2024 - uz.answers-technology.com/13

2-rasm. Googlening qidiruvda yetishmayotgan jumlaning taxminiy tavsiyasi

Bu yerda Googlega *ten* degan bo'g'inni kiritishimiz bilan bizga so'zning taxminiy davomini va biz qidiruvda ko'p foydalanadigan so'zlarga tegishli taxminiy natijalarni tavsiya qilib beryapti(2-rasmga qarang). Agar biz *tensorflow* so'zini qidiruvga bersak, quyidagicha natija chiqarib beradi:



google.com/search?q=tensorflow&oq=tensorflow&gs_lcrp=EgZjaHJvbWUyB...

📍 Карты 🌐 Яндекс 🌐 Яндекс Маркет 🌐 Яндекс 🌐 Яндекс Маркет

tensorflow

🌐 Google Open Source
<https://opensource.google> > ... · Shu sahifani tarjima qilish

TensorFlow | Google Open Source Projects

TensorFlow is an end-to-end open source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries, and community ...

📦 PyPI
<https://pypi.org> > project > t... · Shu sahifani tarjima qilish

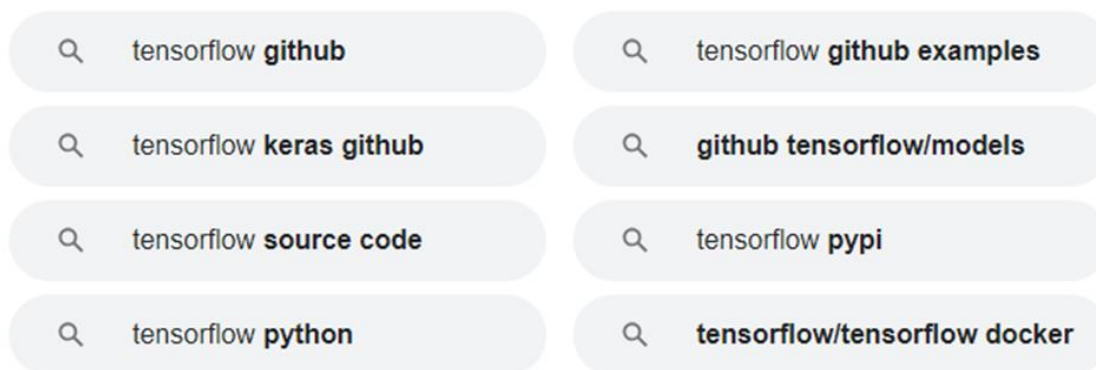
TensorFlow

TensorFlow is an open source software library for high performance numerical computation flexible architecture allows easy deployment of computation ...

3-rasm. Qidiruv natijasida foydalanuvchiga berilgan tavsiyalar

Ko‘rinib turganidek, Google *tensorflow* so‘ziga tegishli bir qancha natijalarni chiqarib berdi(3-rasmga qarang), bundan tashqari, aloqador qidiruvlar qismida qidirilayotgan so‘zga aloqador, foydalanuvchiga kerak bo‘lishi mumkin bo‘lgan yoki qiziq bo‘lgan bir qancha tavsiyalarni ham berdi(4-rasmga qarang).

Aloqador qidiruvlar :



4-rasm. Foydalanuvchiga foydasi tegishi mumkin bo‘lgan tavsiyalar

BERT modeli matnlar bilan bog‘liq bir qancha vazifalarni bajaradi. Jumladan, matnlarni tasniflash, hissiyotlarni tahlil qilish, tarjima jarayoni, tilni aniqlash (matnning tili qaysi tilga tegishli ekanligi) kabi amaliyotlarni misol keltirishimiz mumkin. Yuqorida ta’kidlaganimizdagi, BERT modeli qanchalik yuqori aniqlikda ishlamasin, baribir u ham kamchiliklardan mustasno emas. Misol uchun ushbu model yordamida yaratilgan tillarni aniqlash va tarjima qiluvchi dasturlarga e’tibor beraylik. Tillarni aniqlovchi dastur deyarli xatosiz ishlaydi, chunki ushbu dastur bir qancha tillar bilan tanishtiriladi, ammo ularning ma’nosi ahamiyatsiz qolaveradi. Tarjima qiluvchi dastur esa bir tildagi matnning tarjimasini boshqa tilga tarjima qilishda so‘zlarning ma’nosiga diqqat qiladi. Mashina fikrday olmaganligi uchun hatto model qanchalik katta korpusda o‘qitilmasin, ba’zan xato ishlay boshlaydi. Bu, ayniqsa, badiiy tarjimada yaqqol ko‘rinadi. Chunki badiiy asarlarda so‘zlarning ma’nosi butunlay boshqa ma’noda kelishi kuzatiladi, hatto ikki asarda kelgan aynan bir so‘z bir-biriga o‘xshamagan boshqa-boshqa ma’nalarda kelishi kuzatiladi. Bu esa mashinaning yaxshi natija bermasligiga olib kelishi mumkin.

BERT modeli yorliqsiz (oldindan tayyorlangan) katta ma’lumotlar to‘plamlarida oldindan o‘qitiladi va keyinchalik murakkabroq vazifalarga (nozik sozlash) moslashtirilishi mumkin. Ya’ni bu jarayonlarning barchasi, avvalo, matnlarni qayta ishlashdan so‘ng bajariladi, matn birinchi navbatda raqamli token identifikatorlariga aylantiriladi va tensorlarga ajratiladi[Patzak T., 2021: 3]. Matnlarni BERT modeliga kiritishdan oldin ularni raqamli token identifikatorlariga aylantirish va bir nechta Tensorlarga ajratish kerak. Buning uchun tanlangan matn TensorFlow yoki PyTorch kutubxonalaridan biriga yuklanadi. Yuklangan matn esa ushbu kutubxonalarda qayta ishlanadi, ya’ni kiritilgan matn tokenizatsiya qilinadi, so‘ngra zichlangan vektor ko‘rinishga aylantiriladi. TensorFlow matn ma’lumotlari uchun bir qator oldindan ishlov berish vositalarini, jumladan, normallashtirish (kichik harflar, tinish belgilarini olib tashlash), bir xil uzunlikdagi to‘ldirish ketma-ketligi va lug‘atdan tashqari (OOV) tokenlar bilan ishlashni taklif etadi. TensorFlow matn tasniflash modellari va ketma-ketlik modellarini (masalan, takroriy neyron tarmoqlar (RNN), uzoq

qisqa muddatli xotira tarmoqlari (LSTM) va transformatorlar) yaratish uchun Keras kabi yuqori darajadagi APIlarni taqdim etadi. TensorFlow transformator modellarini, shu jumladan BERT, GPT va T5-ni, shuningdek, maxsus ma'lumotlar to'plamlarida uzatishni o'rganish yoki nozik sozlash uchun oldindan o'qitilgan nazorat nuqtalarini ta'minlaydi. Ayniqsa, BERT kabi transformator arxitekturalari tomonidan ommalashgan e'tibor mexanizmlari modellarga kirish ketma-ketligining tegishli qismlariga e'tibor qaratishga imkon beradi va turli NLP vazifalarida ishlashni yaxshilaydi.

Xulosa sifatida shuni aytish mumkinki, BERT modeli yaratilgunga qadar bir qancha til modellari yaratilgan va ularning kamchiliklarini tuzatish maqsadida ushbu modelga ehtiyoj sezilgan va yaratilgan. BERT modelining tabiiy tilni tushunish va qayta ishlashdagi o'rni muhim ahamiyat kasb etadi. Uning kontekstni tushunishi, oldindan tayyorlangan taqdimotlar va o'rganish qobiliyati turli vazifalar va ilovalar bo'yicha NLP modellarining ishlashini yaxshilashga xizmat qiladi.

BERT modelining turli xil vazifalar bo'yicha o'qitilishi tabiiy tilni qayta ishlashda inson mehnatini sezilarli darajada kamaytiradi va osonlashtiradi. Bundan tashqari, ba'zan inson ilg'amay qoladigan xato va kamchiliklarni tuzatishi mumkin. Ammo BERT modelining o'zi ham kamchiliklardan holi emasligini unutmaslik lozim. Bu modelning qay darajada aniq ishlashi uning bajaradigan vazifasiga ham bog'liq bo'ladi. Ayniqsa, uning juda aniq ishlashi va ba'zi kamchiliklar bilan ishlashini tillarni aniqlash vazifasi va mashina tarjimasida yaqqol ko'rish mumkin. Ammo shunday bo'lsa-da, tabiiy tillarni qayta ishlash bo'yicha tadqiqotlar va ishlanmalar davom etar ekan, BERT modeli inson tilini tushunishda yana-da ko'proq aniqlik bilan ishlash imkoniyatiga ega bo'la boradi.

Foydalanilgan adabiyotlar:

1. Guo Z., Nguyen M.L. / Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10 th International Joint Conference on Natural Language Processing: Student Research Workshop, December 4 - 7, 2020.- B.101–107
2. Berdiyev J.B. BERT va TensorFlowning matnlarni tasniflashdagi ahamiyati / J.B.Berdiyev. - Educational Research in Universal Sciences (ERUS) journal, 2024. – B.176.
3. David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal.2023. Trained on 100 million words and still in shape: BERT meets British National Corpus. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1954–1974, Dubrovnik, Croatia. Association for Computational Linguistics.
4. Patzak T. Automatische Klassifikation von Alltagsszenarien mit einem BERT-basierten Siamese-Netzwerk / T.Patzak. - Humboldt-Universität zu Berlin, Institut für Informatik, 2021.
5. <https://dzone.com/articles/bert-transformers-how-do-they-work>