

MATNLARNI TASNIFLASH UCHUN BERT MODEL VA TENSORFLOW KUTUBXONASINING O'RNI

Berdiyev Jahongir Botir o'g'li

berdiyevjahongir94@gmail.com,

ToshDO'TAU 1-bosqich magistranti

Annotatsiya. Tabiiy tilni qayta ishlash hozirgi kun ilm-fanining muhim masalalaridan biriga aylanib bormoqda. Tabiiy tilni qayta ishlashning eng muhim vazifalaridan biri matnlarni tasniflashdir. Uning maqsadi kiritiladigan matn ma'lumotlarini avvaldan belgilab olingan sinflarga ajratishdan iboratdir. Buning uchun esa BERT modeli va TensorFlow kutubxonasidan foydalanish juda yaxshi samara beradi. Mazkur maqolada matnni tasniflash nima ekanligi, uning tabiiy tilni qayta ishlashdagi ahamiyati, matnni tasniflash borasida jahon tajribasi va erishilgan yutuqlar kabi masalalar yoritilgan. Bundan tashqari BERT modelining yaratilish tarixi va unga bo'lgan ehtiyoj, uning boshqa modellardan farqi va imkoniyatlari hamda Tensorflow kutubxonasini o'rnatish va undan foydalanish, TensorFlowning o'ziga xos xususiyatlari, uning imkoniyatlari va boshqa kutubxonalardan farqi, sun'iy intellekt uchun ahamiyati haqida ham ma'lumotlar kiritilgan. Qolaversa, matnlarni tasniflashda BERT modeli va TensorFlow kutubxonasining o'rni hamda ulardan kutiladigan natijalar qay darajada ijobiy ekanligi kabi masalalarga e'tibor qaratilgan.

Abstract. Natural language processing is becoming one of the important issues of modern science. One of the most important tasks of natural language processing is text classification. Its purpose is to classify input text data into predefined classes. For this, the BERT model and the TensorFlow library are very effective. This article covers issues such as what text classification is, its importance in natural language processing, world experience and achievements in text classification. In addition, the history of the creation of the BERT model and the need for it, its differences and capabilities from other models, and the installation and use of the Tensorflow library, the characteristics of TensorFlow, its capabilities and differences from other libraries, and its importance for artificial intelligence data is also included. In addition, the role of the BERT model and the TensorFlow library in text classification and how positive the expected results are focused.

Аннотация. Обработка естественного языка становится одной из важных проблем современной науки. Одной из наиболее важных задач обработки естественного языка является классификация текста. Его цель — классифицировать входные текстовые данные по предопределенным классам. Для этого очень эффективны модель BERT и библиотека TensorFlow. В данной статье рассматриваются такие вопросы, как что такое классификация текста, ее значение в обработке естественного языка, мировой опыт и достижения в классификации текста. Кроме того, история создания модели BERT и

необходимость в ней, ее отличия и возможности от других моделей, а также установка и использование библиотеки TensorFlow, характеристики TensorFlow, ее возможности и отличия от других библиотек, а также также включена его важность для данных искусственного интеллекта. Кроме того, акцентируется внимание на роли модели BERT и библиотеки TensorFlow в классификации текста и на том, насколько позитивны ожидаемые результаты.

Kalit soʻzlar: *Google, BERT, TensorFlow, sunʼiy intellekt (AI), tuygʻular tahlili, matnni tasniflash.*

Kirish

Zamonlar oʻtgani sari insonning ongi ham rivojlanib boradi. Bu esa, oʻz navbatida insonning bilimi, talab va ehtiyojlari ham ortib borishiga, odamlarni oʻz ustida ishlashga, yanada koʻproq va qulayroq natijalarga erishishga sabab boʻladi. Tabiiyki, buning natijasi avvalo, ilm-fan taraqqiyotida oʻz aksini topadi. Oqibatda, ilm-fanning mavjud sohalarigina emas, avval mavjud boʻlmagan va taraqqiyot natijasida yuzaga kelgan yangi sohalar vujudga keladi. Xuddi shunday sohalaridan biri kompyuter texnologiyalari sohasidir. Vaqt oʻtgani sari bu sohaga ehtiyoj tobora ortib bordi va bugungi kunga kelib kompyuter texnologiyalarisiz hozirgi kunni tasavvur qilish qiyin boʻlib qoldi. Uning kirib bormagan biror sohasi qolmadi. Tabiiy tillarni qayta ishlash ham shular jumlasidandir. Tabiiy tillarni qayta ishlash esa, koʻplab masalalarni hal qilishda muhim ahamiyat kasb etadi. Kompyuter texnologiyalariga ehtiyoj tobora ortib borar ekan, turli xil dasturlar, ilovalar, modellar va shu kabi koʻplab vositalar yaratilaveradi.

Adabiyotlar sharhi. Jahonda kompyuter lingvistikasi sohasida BERT modeli va TensorFlow kutubxonasidan foydalanib tabiiy tilni qayta ishlash, jumladan, matnni tasniflash borasida koʻplab tadqiqotlar olib borilgan. N.Baratalipour va boshqalar [Baratalipour va b., 2020], L.V. Motovskix [Мотовских, 2022], I.V.Selevanova [Селеванова, 2020], D.R.Baymurzina [Баймурзина, 2021], N.Kant va boshqalar [Kant va b., 2019], I.F.Putra va A.Purvarianti [Putra, Purvarianti], S.Prabhu va boshqalar [Prabhu va b., 2021], R.Buoy va boshqalar [Buoy va b., 2021], M.Özkan va G.Kar [Özkan, Kar, 2022], D.Samuel va boshqalar [Samuel va b., 2023], X.Fam [Fam, 2023], D.Noza va boshqalar [Noza va b., 2021], M.Baz va boshqalar [Baz va b., 2022] kabi olimlarni ana shunday tadqiqot olib borgan tadqiqotchilar sirasiga kiritish mumkin.

J.Devling va boshqalar tomonidan BERT modelining yaratilishi matnni tasniflashda yangi bosqich boʻldi. Ular ushbu modelning boshqa til modellaridan ustun tomonlari, qulayliklari, tabiiy tilni qayta ishlashdagi ahamiyati va, umuman olganda, model haqida batafsil maʼlumot beradi. Khang Pham esa matnlarni tasniflash haqida batafsil toʻxtalib, matnning tasnifi va BERT modeli haqida sodda va tushunarli fikr, matn tasnifi uchun BERT modeli qanday ishlashi hamda kutubxonalar, ulardan foydalanish haqida maʼlumotlar keltiradi. BERT yordamida matnlarni tasniflash boʻyicha qoʻllanma ham berib, matnni tasniflash uchun ilk

qadamdan to natijagacha bo'lgan jarayonni ko'rsatib bergan. Rus tadqiqotchisi L.V.Motovskix internet yangiliklari matnlarini tasniflar ekan, “Matnni avtomatik tasniflash uchun foydalaniladigan usullardan biri TF-IDF matritsasi bilan foydalanishga asoslangan LSA yashirin semantik tahlil usuli hisoblanadi. Axborot texnologiyalarining rivojlanishi va tasniflash uchun yangi dasturiy yechimlarning paydo bo'lishi LSA yordamida olingan natijalarni, xususan, turli tasniflagichlardan foydalanish orqali sezilarli darajada yaxshilandi” [Мотовских, 2022], deya fikr bildiradi.

Jahonda tabiiy tilni qayta ishlash, jumladan, matnlarni tasniflash ishlari yaxshi o'rganilgan bo'lishiga qaramay, yurtimizda bu borada qilinishi kerak bo'lgan ishlar talaygina ekanligiga ko'zimiz tushadi. Hatto matnlarni tasniflash tizimi bo'yicha olib borilgan ishlarga ko'zimiz tushmaydi. Faqatgina E.Kuriyozov, U.Salayev, S.Matlatipov, G'.Matlatipovlar tomonidan “BERTbek” nomi modelning yaratilganligini ko'rishimiz mumkin. Ushbu tadqiqotchilar tomonidan “O'zbek tili uchun matn tasniflash ma'lumotlar to'plami va tahlili” nomli maqola ham chop etilgan. Mualliflarning fikriga ko'ra, Eng yaxshi samaradorlikka BERTbek modeli erishadi, bu esa transformatorga asoslangan BERT modeli bo'lib, O'zbekiston korpusida o'qitilgan [Kuriyozov va b., 2023]. Ushbu maqolada BERTbek modelining xususiyatlari, imkoniyatlari, erishilgan natijalari, umuman olganda, modelni atroflicha yoritib bergan.

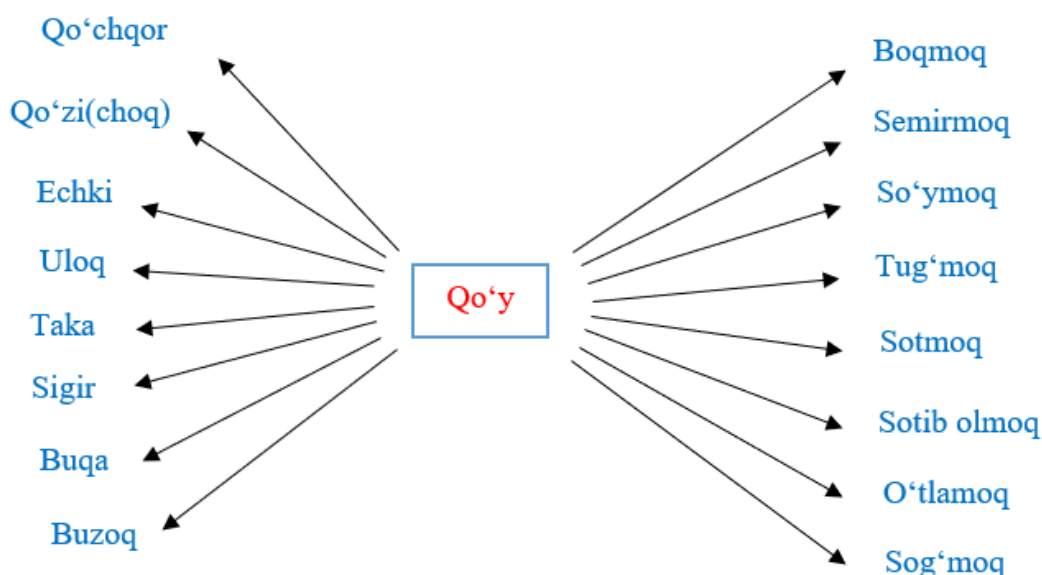
Asosiy qism

BERT modeliga bo'lgan ehtiyoj. Google qidiruv tizimi dunyodagi eng mashhur qidiruv tizimlaridan biri ekanligi barchaga ma'lum. Bu esa o'z foydalanuvchilari uchun Google jamoasi tomonidan yanada ko'proq imkoniyatlar va qulayliklar yaratishni taqozo qiladi. Foydalanuvchilari soni kundan kunga ortib borayotgani hamda axborotga va axborotni qayta ishlashga bo'lgan ehtiyoj borgan sari ortayotganini inobatga olgan holda Google jamoasi ushbu qidiruv tizimining rivojlanib borishi va undan foydalanuvchilarga yanada ko'proq qulaylik va imkoniyat yaratish maqsadida ko'plab yangi loyihalarga ehtiyoj sezdi.

Shunday loyihalardan biri BERT modelining yaratilganligi bo'ldi. BERT modeli 2018-yilda Google tadqiqotchilari jamoasi tomonidan ishlab chiqilgan til modellari oilasi hisoblanadi. BERT (inglizcha Bidirectional Encoder Representations from Transformers – transformatorlardan ikki tomonlama kodlovchi tasvirlar) – bu transformatorlar arxitekturasiga asoslangan til modeli bo'lib, keyinchalik ularni [tabiiy tilni qayta ishlashning](#) keng ko'lamli vazifalarida qo'llash maqsadida til ko'rinishlarini oldindan tayyorlash uchun mo'ljallangan. Masalan, BERT modeli yordamida matnlarni tasniflash, his-tuyg'ularni tahlil qilish, matn yaratish, tillarni aniqlash kabi ko'plab vazifalarini amalga oshirish mumkin. Matnlarni tasniflash tizimini ishlab chiqishda, yuqorida ta'kidlanganidek, eng samarali usul BERT modelidan foydalanishdir. Chunki BERT modelining boshqa til modellaridan farqi shundaki, bu model qidiruvda kontekstdagi so'zlarning butun to'plamidan foydalanadi. Ya'ni “Tilni taqdim etishning so'nggi modellaridan farqli

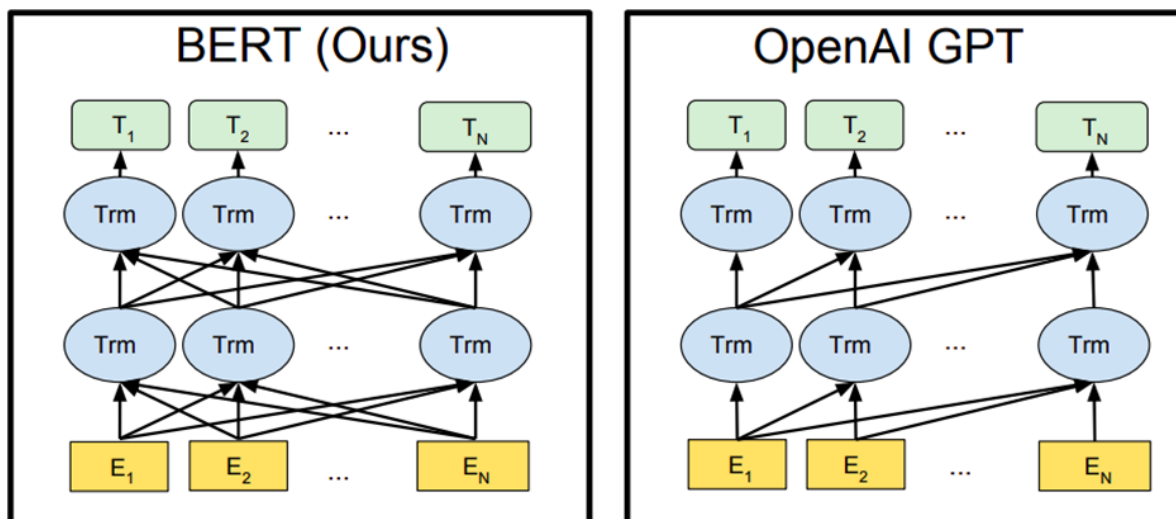
o'laroq, BERT barcha qatlamlarda chap va o'ng kontekstni birgalikda shartlash orqali yorliqsiz matndan chuqur ikki tomonlama tasvirlarni oldindan tayyorlash uchun mo'ljallangan". Shunday ekan, bu model orqali matnni tasniflash, his-tuyg'ularni tahlil qilish, spamni aniqlash, savollarga javob berish kabi vazifalarni yaratish uchun BERT modeli arxitekturasini ba'zi o'zgarishlar bilan osongina sozlab olish mumkin.

BERT modeli orqali tabiiy tilni qayta ishlashda, jumladan, matnni tasniflashda ikki tomonlama transformator kodlovchisidan foydalanadi. Bundan maqsad esa gaplardagi har bir so'zning unga bog'liq bo'lgan va u bilan birga keladigan barcha so'zlarni hisobga oladi. Masalan, BERT [qo'y] so'zi hayvon ekanligini tushunish uchun [qo'y og'zidan cho'p olmagan] jumlasidagi [qo'y] so'zi bilan bog'liq bo'lgan matndagi barcha so'zlarni tekshirib chiqadi. (1-rasmga qarang)



1-rasm. Qo'y so'zi bilan bog'liq bo'lgan so'zlar to'plami

BERTning bunday xususiyati kiritilgan matnni juda yaxshi tushunish va matn bilan ishlashda yuqori darajadagi aniqlikka erishish imkonini beradi. Qiyoslash uchun BERT modeli va GPT modellarini keltirish mumkin. BERT modeli matn bo'ylab yo'nalishsiz harakat qiladi. Bu esa matnni yaxshiroq tushunishga va aniqroq natijani olishga yordam beradi. GPT esa bir tomonlama chapdan o'ngga tomon harakat qiladi. Bu esa kerakli natijani ishlab chiqishda biroz noaniqlikni yuzaga keltirishi mumkin. (2-rasmga qarang) Ammo, shuni ham nazarda tutish kerakki, tabiiy tilni qayta ishlashda eng yaxshi va yuqori natijaga erishish uchun ma'lumotlar to'plamining hajmi qay darajada ko'p ekanligiga bog'liq bo'ladi.



2-rasm. BERT va GPTning ishlash sxemasi

BERT modeli oldindan tayyorlanadigan modeldir. Bu tabiiy tilni qayta ishlash uchun sozlanishdan oldin turli xil katta hajmdagi matnlarning o'qitilishini anglatadi. Uni ma'lum bir tilga o'qitgach, qandaydir aniq vazifa, masalan, his-tuyg'ular tahlili bo'yicha moslashtirish mumkin. Bu esa BERT modeliga unga berilgan vazifaning aniqligini oshirish imkonini beradi.

Matnlarni tasniflash jarayonini, birinchi navbatda, ma'lum bir dasturiy ta'minot kutubxonasini Python ish muhitiga import qilib olishdan boshlash kerak. Bunday dasturiy ta'minot kutubxonalarining bir qancha turlari mavjud. Dasturiy ta'minot kutubxonalarini mashinani o'qitish (ML), neyron tarmoq yaratish kabi muammolarni hal qilish uchun zarur bo'lgan barcha vositalarni taqdim etadi. Shunday kutubxonalardan biri TensorFlowdir (TensorFlow 2015-yilda Google Brain jamoasi tomonidan ishlab chiqilgan). U quyidagicha sozlanadi:

```
Pip install tensorflow
```

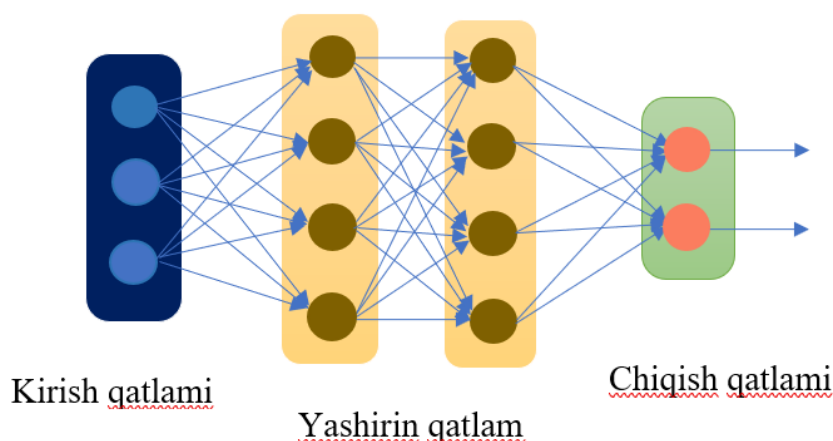
Bu buyruq orqali TensorFlow kutubxonasi o'rnatiladi.

```
Import tensorflow as tf
```

Bu buyruq orqali esa TensorFlow kutubxonasi ishga tushiriladi.

Yuqoridagi amallar bajarib bo'lingandan so'ng esa TensorFlowga tegishli barcha funksiyalarni va modullarni *tf* nomi orqali chaqirish va kerakli amallarni bajarish mumkin.

TensorFlow mashinani o'rganishning kichik to'plami bo'lmish sun'iy intellektga asoslangan chuqur o'rganish (Deep learning)dir. Chuqur o'rganishning asosiy qismlaridan esa biri neyron tarmoqdir. Neyron tarmoq quyidagi sxemaga o'xshaydi:



3-rasm. Neyron tarmoqning ishlash sxemasi

Bu yerda kirish qatlam neyron tarmog'ini qurish uchun katta hajmdagi turli xil ma'lumotlarni qabul qiladi. Bu ma'lumotlar matn, rasm, ovozi fayllar, video fayllar kabi turli xil ma'lumotlar bo'lishi mumkin:

– yashirin qatlamda kiritilgan ma'lumotlar qayta ishlanadi, ma'lumotlarni o'qish, ya'ni tanib olish jarayoni amalga oshadi va raqamlashtiriladi. O'quv jarayoni tugagandan so'ng ma'lumotlar chiqish qatlamiga o'tkaziladi.

– chiqish qatlami tegishli faollashtirish funksiyalarini qo'llash orqali bashorat qilingan chiqishni hosil qiladi. Masalan, matnni tasniflovchi ilovada ma'lum bir matnning qaysi guruhga tegishli ekanligi, tillarni aniqlovchi ilovada esa matnning qaysi tilga tegishli ekanligi haqidagi natijani beradi.(3-rasmga qarang)

Xulosa

Xulosa qilib aytadigan bo'lsak, avvalo, BERT modelining yaratilishi tabiiy tilni qayta ishlashda katta inqilob bo'ldi. Shuningdek, BERT modelining osongina sozlanishi istalgan vazifalarni bajarish uchun imkon beradi. TensorFlow kutubxonasidan foydalanish ham ma'lumotlarni qayta ishlashda eng samarali natijalarni olish imkonini beradi. Jumladan, matnlarni tasniflashda, his-tuyg'ularni tahlil qilishda va boshqa tabiiy tilni qayta ishlash bilan bog'liq jarayonlarda BERT modeli va TensorFlow dasturiy ta'minot kutubxonasidan foydalanish eng samarali natijalarni olishga yordam beradi.



Foydalanilgan adabiyotlar:

1. Baz M., Sami A. , Prottasha 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.
2. Kuriyozov E., Salaev U., Matlatipov S., Matlatipov G'. (2023). Text classification dataset and analysis for Uzbek language. 10.48550/arXiv.2302.14494.
3. Мотовских Л.В. Автоматизированная классификация новостных
4. интернет-текстов (на основе метода латентно-семантического анализа LSA) / Л.В. Мотовских. – Москва: Московский государственный лингвистический университет, 2022.
5. [https://en.wikipedia.org/wiki/BERT_\(language_model\)](https://en.wikipedia.org/wiki/BERT_(language_model))
6. <https://arxiv.org/abs/1810.04805>
7. <https://www.simplilearn.com/tutorials/deep-learning-tutorial/what-is-tensorflow>