

BERT VA TENSORFLOWNING MATNLARNI TASNIFLASHDAGI AHAMIYATI

Berdiyev Jahongir Botir o'g'li

Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti
universiteti 1-bosqich magistranti

E-mail: berdiyevjahongir94@gmail.com

ANNOTATSIYA

Matnlarni tasniflash (klassifikatsiya qilish) tabiiy tilni qayta ishlash (NLP)da muhim vazifalardan biridir. 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, BERT modelining yaratilish tarixi va unga bo'lgan ehtiyoj, uning boshqa modellardan farqi, imkoniyatlari, sun'iy intellekt uchun ahamiyati hamda Tensorflow kutubxonasini o'rnatish va undan foydalanish, TensorFlowning o'ziga xos xususiyatlari, uning imkoniyatlari va boshqa kutubxonalardan farqi, qolaversa, matnlarni tasniflashda BERT modeli va TensorFlow kutubxonasining o'rni hamda ulardan kutiladigan natijalar qay darajada ijobiy ekanligi kabi masalalarga e'tibor qaratilgan.

Kalit so'zlar: Google, BERT, TensorFlow, sun'iy intellekt(AI), tuyg'ular tahlili, matnni tasniflash.

ABSTRACT

Classification of texts is a crucial task in natural language processing (NLP). The aim is to categorise input text data into predetermined classes. To achieve this, the BERT model and TensorFlow library are highly effective. This article discusses text classification, its significance in natural language processing, global advancements and accomplishments in text classification, the history of the BERT model, its unique features, capabilities, and importance for artificial intelligence, as well as its use in the Tensorflow library. This text discusses the installation and usage of TensorFlow, its features, capabilities, and differences from other libraries. It also covers the role of the BERT model and TensorFlow library in text classification and the potential positive results.

Keywords: Google, BERT, TensorFlow, artificial intelligence (AI), sentiment analysis, text classification.

KIRISH

Zamonlar o'tgani sari insonning ongi ham rivojlanib boradi. Bu esa, o'z navbatida insonning bilimi, talab va ehtiyojlari ham ortib borishiga, odamlarni o'z ustida ishlashga, yanada ko'proq va qulayroq natijalarga erishishga sabab bo'ladi. Tabiiyki, buning natijasi, avvalo, ilm-fan taraqqiyotida o'z aksini topadi. Oqibatda, ilm-fanning mavjud sohalarigina emas, avval mavjud bo'lmagan va taraqqiyot natijasida yuzaga kelgan yangi sohalar ham vujudga keladi. Xuddi shunday sohalardan biri kompyuter texnologiyalari sohasidir. Vaqt o'tgani sari bu sohaga ehtiyoj tobora ortib bormoqda. Mana shu ehtiyoj tufayli turli xil dasturlar, ilovalar, modellar va shu kabi ko'plab vositalar ishlab chiqilmoqda. Shulardan biri hozirgi unda inson aqli bilan "raqobat" qilayotgan sun'iy intellekt uchun muhim bo'lgan BERT modelidir.

BERT modeliga bo'lgan ehtiyoj. Google qidiruv tizimi dunyodagi eng katta 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 amaliy ishlarning natijasidan biri BERT modelining yaratilganligi bo'ldi. BERT modeli 2018-yilda Google tadqiqotchilari jamoasi tomonidan ishlab chiqilgan til modellari oilasi hisoblanadi.

BERT (inglizcha Bidirectional Encoder Rpresentations 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 foydalanish maqsadida til ko'rinishlarini oldindan tayyorlash uchun mo'ljallangan.¹

Muammoning o'rganilganlik darajasi. Jahonda kompyuter lingvistikasi sohasida matnni tasniflash borasida ko'plab tadqiqotlar olib borilgan. L.V. Motovskix², I.V. Selevanova³, D.R. Baymurzina⁴, N. Kant va b.⁵, I.F. Putra va

¹ [https://en.wikipedia.org/wiki/BERT_\(language_model\)](https://en.wikipedia.org/wiki/BERT_(language_model))

² Мотовских Л.В. Автоматизированная классификация новостных интернет-текстов (на основе метода латентно-семантического анализа lsa) / Л.В. Мотовских. – Москва: Московский государственный лингвистический университет, 2022.

³ Селеванова И.В. методы тематической классификации научных текстов на основе теоретико-информационного подхода / И.В. Селеванова – Новосибирск: Государственная публичная научно-техническая библиотека Сибирского отделения Российской академии наук, 2020.

⁴ Баймурзина Д.Р. Диалоговая система для разговоров о нейросетевых моделях и общих тема, / Д.Р. Баймурзина – Москва: Московский физико-технический институт, 2021.

⁵ Kant N. va boshqalar Practical Text Classification With Large Pre-Trained Language Models / N. Kant va boshqalar – Berkeley: University of California, 2019.

A.Purvarianti¹, S.Prabhu va boshqalar², R.Buoy va boshqalar³, M.Özkan va G.Kar⁴, D.Samuel va boshqalar⁵, X.Fam⁶ kabi olimlarni ana shunday tadqiqot olib borgan tadqiqotchilar sirasiga kiritish mumkin.

J.Devling va boshqalar tomonidan BERT modelining yaratilishi matnni tasniflashda yangi bosqich bo'ldi. Ular ushbu modelning boshqa til modellaridan ustun tomonlari, qulayliklari, tabiiy tilni qayta ishlashdagi ahamiyati va, umuman olganda, model haqida batafsil ma'lumot beradi. Khang Pham esa matnlarni tasniflash haqida batafsil to'xtalib, matnning tasnifi va BERT modeli haqida sodda va tushunarli fikr, matn tasnifi uchun BERT modeli qanday ishlashi hamda kutubxonalar, ulardan foydalanish haqida ma'lumotlar beradi. BERT yordamida matnlarni tasniflash bo'yicha qo'llanma 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 echimlarning paydo bo'lishi LSA yordamida olingan natijalarni, xususan, turli tasniflagichlardan foydalanish orqali sezilarli darajada yaxshilandi"⁷, deya fikr bildiradi.

Jahonda matnlarni tasniflash ishlari yaxshi o'rganilgan bo'lishiga qaramay, O'zbekistonda matnlarni tasniflash tizimi bo'yicha olib borilgan ishlarga ko'zimiz tushmaydi. Faqatgina E.Kuriyozov, U.Salayev, S.Matlatipov, G'.Matlatipovlar tomonidan "BERTbek" nomli modelning yaratilganligini ko'rishimiz mumkin. Ushbu tadqiqotchilar tomonidan "O'zbek tili uchun matn tasniflash ma'lumotlar to'plami va tahlili"⁸ nomli maqola ham chop etilgan bo'lib, bu maqola BERTbek modelining xususiyatlari, imkoniyatlari, erishilgan natijalari, umuman olganda, modelni atroflicha tavsiflab bergan.

ASOSIY QISM

Matnlarni tasniflash tizimini ishlab chiqishda, yuqorida takidlanganidek, eng samarali usul BERT modelidan foydalanishdir. Chunki BERT modelining boshqa til

¹ Putra I.F., Purvarianti A. Improving Indonesian Text Classification Using Multilingual Language Model / I.F.Putra, A.Purvarianti – Indonesia: School of Electrical Engineering and Informatics Institut Teknologi Bandung, 2020.

² <https://arxiv.org/abs/2104.14289>, 2021.

³ <https://arxiv.org/abs/2112.06748>

⁴ Özkan M., Kar G. Türkçe dilinde yazılan bilimsel metinlerin derin öğrenme tekniği uygulanarak çoklu sınıflandırılması / M.Özkan, G.Kar – İstanbul: Mühendislik Bilimleri ve Tasarım Dergisi, 2022.

⁵ <https://arxiv.org/abs/2303.09859>

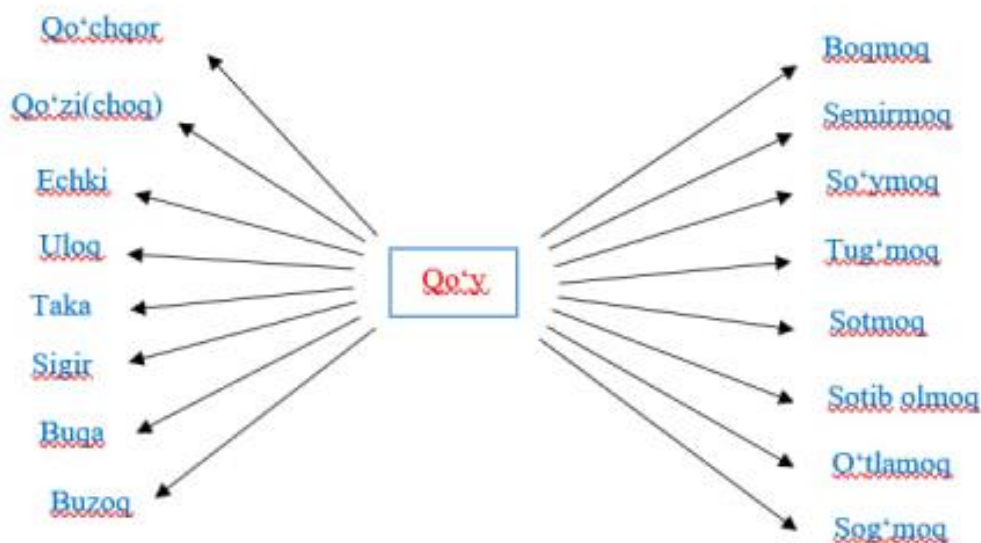
⁶ <https://medium.com/@khang.pham.exact/text-classification-with-bert7afaacc5e49b#:~:text=How%20does%20the%20BERT%20model,the%20meaning%20of%20the%20text>

⁷ Мотовских Л.В. Автоматизированная классификация новостных интернет-текстов (на основе метода латентно-семантического анализа lsa) / Л.В. Мотовских. – Москва: Московский государственный лингвистический университет, 2022.

⁸ https://www.researchgate.net/publication/368877295_Text_classification_dataset_and_analysis_for_Uzbek_language

modellaridan farqi “avvalo, dasturlash bilan bog‘liq bo‘lib, bu tizim to‘g‘ridan to‘g‘ri sun‘iy intellekt bilan ishlaydi. Shunday ekan, bu tizimni ishlab chiqish uchun sun‘iy intellekt tili deb ataladigan Python dasturlash tilidan foydalaniladi. “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.”¹ Ya‘ni bu model orqali matn 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. Chunki ushbu model ma‘lum bir tilga oldindan o‘qitilgan bo‘ladi.

BERT modeli orqali tabiiy tilni qayta ishlashda, jumladan, matn 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-rasm. Qo‘y so‘zi bilan bog‘liq bo‘lgan so‘zlar to‘plami.

BERTning bunday xususiyati kiritilgan matn juda yaxshi tushunish va matn bilan ishlashda yuqori darajadagi aniqlikka erishish imkonini beradi. Ammo, shuni ham nazarda tutish kerakki, tabiiy tilni qayta ishlashda eng yaxshi va yuqori natijaga

¹ <https://arxiv.org/abs/1810.04805>

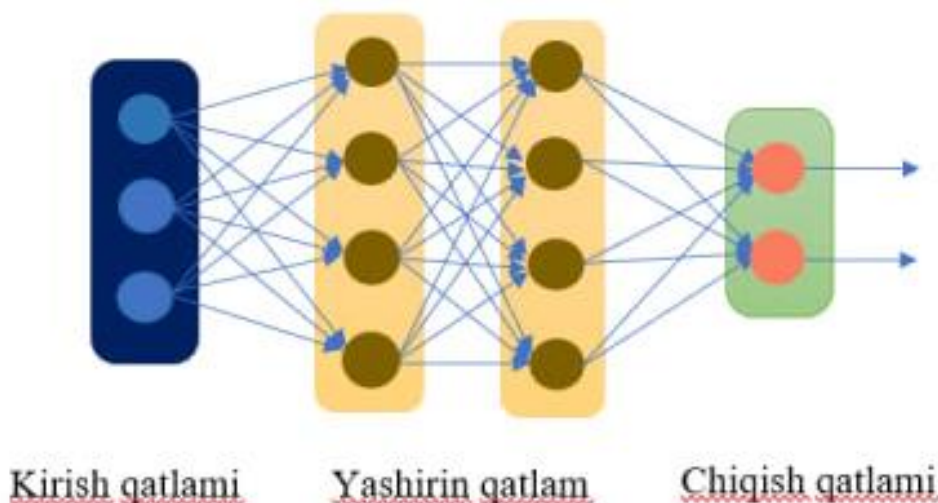
erishish uchun ma'lumotlar to'plami hajmining qay darajada ko'p ekanligiga bog'liq bo'ladi.

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 BERTga 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 --upgrade pip  
Pip install tensorflow  
Import tensorflow as tf
```

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



2-rasm. Neyron tarmoqning ishlash sxemasi

Bu yerda kirish qatlam neyron tarmog'ini qurish uchun katta hajmdagi turli xil ma'lumotlarni qabul qiladi.

Yashirin qatlamda kiritilgan ma'lumotlar qayta ishlanadi, uning o'ziga xos xususiyatlarini ajratib oladi. O'quv jarayoni tugagandan so'ng ma'lumotlar chiqish qatlamiga o'tkaziladi.

Chiqish qatlami tegishli faollashtirish funksiyalarini qo'llash orqali bashorat qilingan chiqishni hosil qiladi¹. Masalan, matnni tasniflovchi ilovada ma'lum bir matnning qaysi guruhga tegishli ekanligi haqida aytadi.

Xulosa qilib aytadigan bo'lsak, avvalo, BERT modelining yaratilishi tabiiy tilni qayta ishlashda katta inqilob bo'ldi. 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 RO'YXATI: (REFERENCES)

1. Мотовских Л.В. Автоматизированная классификация новостных интернет-текстов (на основе метода латентно-семантического анализа lsa), автореф. канд. фил. н. – Москва: Московский государственный лингвистический университет, 2022.

INTERNET MA'LUMOTLARI:

2. [https://en.wikipedia.org/wiki/BERT_\(language_model\)](https://en.wikipedia.org/wiki/BERT_(language_model))

3. https://www.researchgate.net/publication/368877295_Text_classification_dataset_and_analysis_for_Uzbek_language

4. <https://arxiv.org/abs/1810.04805>

5. <https://www.simplilearn.com/tutorials/deep-learning-tutorial/what-is-tensorflow>

¹ <https://www.simplilearn.com/tutorials/deep-learning-tutorial/what-is-tensorflow>