MATNLARNI TO‘LDIRISH ALGORITMLARIDA QO‘LLANILADIGAN METODLARI

Bakayev Ilhom Izatovich
Buxoro davlat universiteti
Abdunasimov Shavkat Shuhratovich
Buxoro davlat universiteti

Annotatsiya. Mazkur maqolada tabiiy tilni qayta ishlash va mashinali o‘qitish tizimlarida ma’lumotlar yetishmasligi muammosini kamaytirishga yo‘naltirilgan matnlarni to‘ldirish algoritmlari tahlil qilingan. Ishning asosiy maqsadi matnni to‘ldirishga asoslangan metodlarni ilmiy-nazariy jihatdan tizimlashtirish hamda ularning qo‘llanish imkoniyatlarini baholashdan iborat. Maqolada sinonimlar asosida almashtirish, qisqartmalar bilan ishlash, so‘zlarni tasviriy va kontekstli fazoda almashtirish, teskari mashinaviy tarjima, shovqin kiritish hamda krossover yondashuvlarining ishlash tamoyillari ko‘rib chiqilgan. Metodlarning afzalliklari va cheklovlari semantik yaxlitlikni saqlash nuqtayi nazaridan tahlil qilingan. O‘tkazilgan tahlil natijalari matnlarni to‘ldirish algoritmlaridan foydalanish tasniflash modellarining barqarorligini oshirish, umumlashuv qobiliyatini yaxshilash hamda o‘quv ma’lumotlari xilma-xilligini ta’minlashga xizmat qilishini ko‘rsatadi. Olingan xulosalar tabiiy tilni qayta ishlashga oid amaliy tizimlarni ishlab chiqish va takomillashtirish jarayonlarida qo‘llanishi mumkin.

Kalit so‘zlar: matnlarni to‘ldirish algoritmlari, tabiiy tilni qayta ishlash, mashinali o‘qitish, ma’lumotlarni kengaytirish, sinonimik almashtirish, kontekstli modellar, teskari mashinaviy tarjima, shovqinli ma’lumotlar.

Аннотация. В статье рассматриваются алгоритмы дополнения текстов, направленные на снижение проблемы недостаточности данных в задачах машинного обучения и обработки естественного языка. Основной целью работы является теоретическая систематизация методов текстового дополнения и оценка возможностей их применения. Проанализированы методы замены слов синонимами, работы с сокращениями, замены слов в векторном и контекстном пространстве, обратного машинного перевода, внесения шумов и кроссовер-подходы. Рассмотрены их преимущества и ограничения с точки зрения сохранения семантической целостности текста. Результаты анализа показывают, что использование алгоритмов дополнения текстов способствует повышению устойчивости моделей классификации, улучшению их обобщающей способности и увеличению разнообразия обучающих данных. Полученные выводы могут быть использованы при разработке и совершенствовании прикладных систем обработки естественного языка.

ISSN 3030-3907

Ключевые слова: дополнение текстов, обработка естественного языка, машинное обучение, расширение данных, синонимическая замена, контекстные модели, обратный машинный перевод, шумовые данные.

Abstract. The article examines text augmentation algorithms aimed at mitigating the data scarcity problem in machine learning and natural language processing tasks. The primary objective is to provide a theoretical systematization of text augmentation methods and to assess their applicability. The study analyzes synonym-based replacement, abbreviation processing, word substitution in vector and contextual spaces, back-translation, noise injection, and crossover-based approaches. The advantages and limitations of these methods are discussed with respect to preserving semantic integrity. The analysis demonstrates that text augmentation algorithms enhance the robustness and generalization capability of classification models and increase the diversity of training data. The obtained results can be applied in the development and improvement of practical natural language processing systems.

Keywords: text augmentation, natural language processing, machine learning, data augmentation, synonym replacement, contextual models, back-translation, noisy data.

KIRISH

Hozirgi kunda ma’lumotlarni tasniflash masalalari dolzarb yo‘nalishlardan biri bo‘lib, u bozor munosabatlarida mahsulotlar to‘g‘risidagi axborotlarni tahlil qilish, jamoatchilik fikrini o‘rganish hamda axborot xavfsizligi tizimlarida elektron xabarlarni spam sifatida aniqlash kabi vazifalarda keng qo‘llanilmoqda. Biroq ma’lumotlarga dastlabki ishlov berish jarayonida yuzaga keladigan turli muammolar sababli an’anaviy tasniflash algoritmlari har doim ham yuqori aniqlikdagi natijalarni ta’minlay olmaydi. Ushbu muammolar qatoriga tabiiy tillar uchun leksik resurslarning cheklanganligi, turli sheva va dialektlarda yozilgan matnlar, imloviy farqlar hamda jargon birliklardan foydalanishni kiritish mumkin. Mazkur kamchiliklarni bartaraf etish maqsadida matnni to‘ldirishga asoslangan yondashuvlar taklif etilgan. Matnni to‘ldirish mashinali o‘qitish vazifalarida mavjud ma’lumotlar to‘plamining tavsifini o‘zgartirish orqali ularni boyitishga qaratilgan usullar majmuasini anglatadi. Ushbu yondashuv doirasida so‘zlarni sinonimlari bilan almashtirish, vektor tasvirlari o‘xshash bo‘lgan birliklarni qo‘llash, kontekstga mos holda so‘zlarni o‘zgartirish, teskari mashinaviy tarjimadan foydalanish, qisqartmalarni ochish yoki qisqartma shaklga keltirish hamda matnni segmentlarga ajratish kabi usullar qo‘llaniladi. Keyingi bo‘limlarda ushbu metodlarning ishlash mexanizmlari batafsil yoritiladi.

Mavjud metodlar

ISSN 3030-3907

Sinonimlarni o‘zgartirish. Eng oddiy metodlardan biri so‘zlarni sinonimlar bilan almashtirish bo‘lib, odatda lug‘at orqali so‘zning sinonimlari yordamida amalga oshiriladi. Masalan [1] ishda Ingliz tilidagi ot so‘z turkumiga tegishli so‘zning sinonimi sifat so‘z turkumiga tegishli so‘z bilan almashtirilishi(mustaqil ma’noga ega bo‘lmagan so‘zlar bundan mustasno) mumkin. Lekin bu har doim ham to‘g‘ri natija bermasligi mualliflar tomonidan ta’kidlab o‘tilgan. O‘zbek tilidada esa so‘z qaysi so‘z turkumida bo‘lsa uning sinonimi ham shu so‘z turkumida bo‘ladi. Ba’zi holatlarda esa so‘zlar bir-biriga sinonim bo‘lmasa-da, ular matn kontekstida sinonimlik vazifasini bajarishi mumkin 1-jadval.

Vatan! Sen qanchalar **azizsan** bizga.

Ona yurt! Sen qanchalar **qadrlisan** bizga

1-jadval.

Sinonim so‘zlar

Vatan	Aziz
Ona yurt	Qadrli
Ona diyor	Qimmatli
Yurtim	qadrdon

O‘zbek tilida bir ma’no bir necha so‘z bilan ifodalanishi mumkin. Masalan, yuz, ya’ni inson bosh qismining old tomoni bet, aft, bashara, nusxa turq, oraz, uzor, ruxsor, chehra so‘zlari orqali ifodalanadi. Ushbu so‘zlar o‘zaro leksik, ya’ni lug‘aviy sinonim hisoblanadi. Bundan tashqari o‘zbek tilida tilchi – tilshunos, o‘rinsiz – noo‘rin kabi qo‘shimchalar yordamida bir ma’noni 2 xil shaklda yuzaga chiqish imkoni ham mavjud 2-jadval.

2-jadval.

Sinonim qo‘shimchalar

Gapingiz**ga** tushunmadim, mehmon.	Gapingiz**ni** tushunmadim, mehmon.
Tanlov ishtirokchilari**ning** biri samarqand**dan** ekan.	Tanlov ishtirokchilari**dan** biri samarqand**lik** ekan.

Nutqiy jarayonda so‘zlovchi maqsadga muvofiq ravishda sinonimlardan birini tanlaydi. So‘zdan o‘rinli foydalanish so‘zlovchining maqsadini aniq, ravon va ta’sirchan ifodalashda katta ahamiyatga ega. O‘zbek tilida sinoimlar qo‘shimcha ma’no anglatishi – anglatmasligiga ko‘ra 2 ga bo‘linadi: to‘liq sinonimlar va ma’noviy sinonimlar. Agar asos ma’no va qo‘shimcha ma’no bir-biriga to‘g‘ri kelsa, to‘liq sinonim bo‘ladi 3-jadval.

3-jadval.

To‘liq sinonimlar.

lingvistika	tilshunoslik
Fazogir	kosmonaft

ISSN 3030-3907

huqushunos	Yurist
ayg‘oqchi	Agent Josus Shpion
anketa	so‘rovnoma
barg	Yaproq
farzand	bola zuriyod

Asos ma’nosi bir xil bo‘lib, lekin qo‘shimcha ma’nosi bir-biridan farq qilsa, ma’noviy sinonim hisoblanadi 4-jadval.

4-jadval.

Ma’noviy sinonimlar

kerak	lozim shart darkor
katta	ulug‘ buyuk ulkan

O‘zbek tilida ma’nodoshlik barcha so‘z turkumlari doirasida uchraydi. Bundan tashqari tilimizda ayrim so‘zlar ko‘chma ma’noda qo‘llanib, boshqa so‘zga vaqtincha ma’nodosh bo‘la oladi. Masalan, qattiq so‘z ko‘chma ma’noda baxil, xasis, ziqna so‘zlari bilan ma’nodosh bo‘la oladi.

Juda **qattiq** odam ekan	Juda **xasis** odam ekan

Qisqartma metodi. Ushbu metod so‘zlar va jumlalarni qisqartma holatga keltirish yoki qisqartmani ochiq holga keltirish. Masalan, ingliz tilida ko‘p so‘zlarning qisqartmalari mavjud va bu metodning ingliz tili uchun realizatsiyasi GitHub veb servisida joylashtirilgan.

O‘zbek tilida tejash til birligini qisqartirish, tushirib qoldirish orqali ketadigan vaqtni kamaytirish, tildagi ortiqchalikni bartaraf etish ehtiyojini kelib chiqadi [2].

Buyam Opke ober	Bu ham Olib kel Olib ber

Olib ber so‘zining ober, sizning so‘zini sizing kabi tejalishi nutqiy qisqaruv hisoblanadi. Tejash barqaror yoki vaqtincha bo‘lishi mumkin. Masalan, o‘zbek tilida jo‘nalish kelishigi dastlab -garu ko‘rinishida bo‘lgan, bugungi kunda -ga shaklida kelib qisqarib qolgan. Bu barqaror tejashdir. Olib ber so‘zining ober

ISSN 3030-3907

shaklida kelishi esa vaqtincha tejash bo‘lib, nutqiy aloqani osonlashtirishga xizmat qiladi.

So‘zlarni tasvirlash metodi. Ushbu metod tasodifiy so‘zlarni ularning tasvirlash fazosidagi (Word Embeddings) eng yaqin so‘zlar bilan almashtiradi. Bu metod har doim ham so‘zning sinonimi emas balki, matnlardagi o‘xshash konteksdagi so‘zlar bilan almashtiriladi. Ba’zi hollarda keraksiz almashtirishlarni oldini olish maqsadida bir so‘z turkumiga tegishli so‘zlar bilan almashtiriladi. Bu kabi matnni to‘ldirishi metodi [3] ilmiy tadqiqot ishida qo‘llanilgan 5-jadval.

Ko‘rgazmalar O‘zbekistonlik **rassomlar** ham ishtirok etishdi.
Ko‘rgazmalar O‘zbekistonlik **mo‘yqalam sohiblari** ham ishtirok etishdi.
Askarlarimiz yurtimiz sarhadlarini ko‘z qorachig‘idek asrashadi.
Vatan posbonlarimiz yurtimiz sarhadlarini ko‘z qorachig‘idek asrashadi.

5-Jadval

So‘zning tasviriy ifoda bilan almashtirishlishi

So‘z	Tasviriy ifoda
rassom	mo‘yqalam sohiblari
Askar	Vatan posbonlari

Konteksli tasvirlash metodi. Word2Vec, fasttext, Glove neyron tarmoqlardan tashqari so‘zning konteksini inobatga oladi. Bir qator [4] [5] ilmiy tadqiqot ishlarida bu kabi tasvirlashlarni qurish uchun ikki yo‘nalishli til modellaridan foydalanilgan. Til modellari ma’lum bir so‘zga emas, balki so‘zlar to‘plamining taqsimlanishini inobatga oladi. Shuning uchun yuqori ehtimollikka mos keladigan har qanday so‘z almashtirish sifatida ishlatilishi mumkin. Bundan tashqari, transformer arxitekturasiga asoslangan modellar ham ishlatilmoqda. Bu modellar katta ma’lumotlar to‘plami orqali o‘qitilib, maskirovkali tokenlar to‘ldiriladi va o‘tkazilib yuborilgan so‘zlar tiklanadi. Shuning uchun ushbu modellar ba’zi bir so‘zlarni maskalarga almashtiradi va transformerlarga uzatadi bu o‘z navbatida maskani so‘zga almashtiradi. Bu model gapdagi ixtiyoriy so‘zlar ketma-ketligi o‘rtasida maskani joylashtiradi, model esa mos so‘z bilan almashtiradi[6]. Quyidagi misolda keltirilgan.

Kecha futbol bahsida O‘zbekistonning qo‘li baland keldi.
Kecha MASK(bo‘lib o‘tgan, o‘tkazilgan) futbol bahsida O‘zbekistonning qo‘li baland keldi.
Kecha futbol bahsida O‘zbekistonning MASK(jamoasi,futbolchilari,komandasi) qo‘li baland keldi.

Muhim va muhim bo‘lmagan so‘zlarni almashtirish va o‘chirish. Ushbu metoddagi yondashuvlar turlicha bo‘lib, olimlar tomonidan turli xil yondashuvlar qo‘llaniladi. Ilmiy tadqiqot [7] ishida TF-IDF qiymati kichik bo‘lgan muhim

ISSN 3030-3907

bo‘lmagan so‘zlar ishda teskarisi ya’ni TF-IDF qiymati katta bo‘lgan muhim bo‘lgan so‘zlar almashtirish taklif qilingan. Ba’zi holatlarda kontekstga nisbatan ham tahlil qilinadi[8].

Vatanim, **men** uchun yagonasan! Vatanim, **biz** uchun yagonasan!

O‘zbek tilida matndagi qiymati kichik bo‘lgan so‘zlarni matnning ma’no mazmuniga biroz ta’sir qilsa-da, almashtirish mumkin.

Teskari mashinaviy tarjima. Avtomatik mashinaviy tarjima matn boshqa tilga o‘giriladi keyin esa teskari joriy tilda o‘giriladi. Bunda matndagi ba’zi bir jumlalar boshqa ko‘rinishda ifodalanadi[7].

Ona yurtim, men uchun azizsan.	Ona yurtim, men uchun azizsan.
My motherland, you are dear to me	Родина моя, ты мне дорога
Vatanim, sen menga azizsan.	Vatanim, sen menga azizsan.

Shovqin metodi. Bu metodda iloji boricha turli xil usullarda matnni buzish tushuniladi. Matnni buzish deganda unga xato harf, tinish belgisini qo‘shish, registr alamashtirish tushuniladi. Xatolarni qo‘shganda, ularni yozish paytida qilingan xatolarga o‘xshash qilib qo‘yishga harakat qilishdir. Bundan tashqari so‘zlarni o‘chirish, almashtirish, tasodifiy joylashtirish shuningdek to‘liq gapni almashtirish. Yana bir ilmiy tadqiqqot ishida [9] gapdagi tasodifiy so‘zni sinonimi almashtirish va tasodifiy joylashtirish.

O‘quvchila**rn**ing ijodiy ishlari tahlil qilindi. O‘quvchila**m**ing ijodiy ishlari tahlil qilindi.

Krossover. Bu metodda [10] matn 2 qismga bo‘linadi: A va B. Bunda har matnni mos A=A1+A2, B=B1+B2 ko‘rinishga keltiramiz. “+” belgisi konkatenatsiyani bildiradi. Shundan so‘ng A1+A2 va B1+B2 o‘qitiladi. Misol:

<table>
<tr><td>Orzularning bo‘lmadi so‘nggi nuqtasi,
Fikrning muntazam aniq qirg‘og‘i.</td><td>A1</td><td rowspan="2">A1+B2
Orzularning bo‘lmadi so‘nggi nuqtasi,
Fikrning muntazam aniq qirg‘og‘i.
Hamisha ufqlar sizni chaqirar,
Misoli dengizda ochilgan yelkan.</td></tr>
<tr><td>Tafakkur-orzuning chin omuxtasi,
Istiqbol mulkining nurli chirog‘i.</td><td>A2</td></tr>
<tr><td>Hey, shirin orzular, teran fikrlar,
Qanchalar go‘zalsiz, qanchalar ulkan!</td><td>B1</td><td>B1+A2
Hey, shirin orzular, teran fikrlar,
Qanchalar go‘zalsiz, qanchalar ulkan!</td></tr>
</table>

ISSN 3030-3907

Hamisha ufqlar sizni chaqirar, Misoli dengizda ochilgan yelkan.	**B2**	Tafakkur-orzuning chin omuxtasi, Istiqbol mulkining nurli chirog‘i.

Bu metod natijasida matn mazmunida o‘zgarishlar paydo bo‘ladi. Ushu misolda she’riy matn tanlanganligi sababli mazmunda noaniqlik paydo bo‘ldi. Katta hajmli matnlarda qo‘llanilganda ushbu metod ijobiy natija berishi mumkin.

XULOSA

Ko‘rib chiqilgan matnlarni to‘ldirish metodlarining aksariyati Python dasturlash tilida ishlab chiqilgan tabiiy tilni qayta ishlashga oid kutubxonalarda, asosan, ingliz tilidagi matnlar uchun mo‘ljallangan algoritmlar sifatida amalga oshirilgan. Shunga qaramasdan, maqolada keltirilgan nazariy tahlillar va misollar ushbu yondashuvlarning ayrimlarini boshqa tillar uchun ham moslashtirish mumkinligini ko‘rsatadi. Xususan, sinonimlar asosida almashtirish, kontekstga mos so‘zlarni tanlash, teskari mashinaviy tarjima hamda shovqin kiritishga asoslangan metodlar lingvistik xususiyatlar hisobga olingan holda o‘zbek tilidagi matnlar bilan ishlash jarayonida ham samarali qo‘llanilishi mumkin. Bunda tilning grammatik tuzilishi, morfologik boyligi va semantik bog‘lanishlarini inobatga olish muhim ahamiyat kasb etadi. Umuman olganda, matnlarni to‘ldirish metodlarini mahalliy til resurslariga moslashtirish mashinali o‘qitish modellarining barqarorligini oshirish, ma’lumotlar tanqisligi sharoitida tasniflash aniqligini yaxshilash hamda tabiiy tilni qayta ishlashga asoslangan amaliy tizimlarni rivojlantirish uchun istiqbolli yo‘nalishlardan biri hisoblanadi.

FOYDALANILGAN ADABIYOTLAR

1. X. Zhang, J. J. Zhao, and Y. LeCun, “Character-level Convolutional Networks for Text Classification,” in *NIPS*, 2015, pp. 1–9. [Online]. Available: https://api.semanticscholar.org/CorpusID:368182

2. B. Mengliyev, O. Xoliyorov, and N. Abdurahmanova, *O‘zbek tilidan universal qo‘llanma*. Toshkent: Akademiknashr, 2019.

3. W. Y. Wang and D. Yang, “That’s So Annoying!!!: A Lexical and Frame-Semantic Embedding Based Data Augmentation Approach to Automatic Categorization of Annoying Behaviors using {\#}petpeeve Tweets,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal: Association for Computational Linguistics, 2015, pp. 2557–2563. doi: 10.18653/v1/D15-1306.

4. M. Fadaee, A. Bisazza, and C. Monz, “Data Augmentation for Low-Resource Neural Machine Translation,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 567–573. doi:

ISSN 3030-3907

10.18653/v1/P17-2090.
5. S. Kobayashi, “Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations,” in *Proceedings of the 2018 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 452–457. doi: 10.18653/v1/N18-2072.
6. S. Garg and G. Ramakrishnan, “BAE: BERT-based Adversarial Examples for Text Classification,” *ArXiv*, vol. abs/2004.01970, 2020, [Online]. Available: https://api.semanticscholar.org/CorpusID:214802269
7. Q. Xie, Z. Dai, E. H. Hovy, M.-T. Luong, and Q. V Le, “Unsupervised Data Augmentation for Consistency Training,” *arXiv Learn.*, 2019, [Online]. Available: https://api.semanticscholar.org/CorpusID:195873898
8. H. Chen and Y. Ji, “Improving the Explainability of Neural Sentiment Classifiers via Data Augmentation,” in *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019, pp. 1–11. [Online]. Available: https://api.semanticscholar.org/CorpusID:203951496
9. J. Wei and K. Zou, “{EDA}: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, 2019, pp. 6382–6388. doi: 10.18653/v1/D19-1670.
10. F. M. Luque, “Atalaya at TASS 2019: Data Augmentation and Robust Embeddings for Sentiment Analysis,” in *IberLEF@SEPLN*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:199448337

ISSN 3030-3907