DOMEN VA HODISAGA MOSLASHUVCHAN MATN KENGAYTIRISH YONDASHUVLARINING INQIROZ HOLATLARIDAGI TVITLARNI TASNIFLASHDAGI SAMARADORLIGI

Subxoqulov Umidjon
Buxoro davlat uiversiteti

Annotatsiya. Ushbu maqolada domen va hodisaga moslashuvchan matn kengaytirish (text augmentation) yondashuvining favqulodda vaziyatlarga oid tvitlarni avtomatik tasniflashdagi samaradorligi ilmiy-tahliliy jihatdan baholangan. Maqolaning maqsadi matn kengaytirishning WordNet, Word2Vec va ProMinent Words (PM-Words) asosidagi algoritmlarini metodologik va eksperimental jihatdan tahlil qilish, ularning afzalliklari, cheklovlari hamda amaliy qo'llash imkoniyatlarini baholashdan iborat. Tahlil jarayonida ma'lumotlarni oldindan qayta ishlash bosqichlari, algoritmlarning ishlash tamoyillari, hisoblash murakkabligi, tajriba dizayni va baholash mezonlari o'rganildi. Natijalar domen va hodisaga moslashtirilgan matn kengaytirish usullari klassifikatsiya sifat ko'rsatkichlarini sezilarli darajada yaxshilashini ko'rsatadi. Shu bilan birga, ma'lumotlar oqishi (data leakage), baholash protokoli va tajriba dizayni bilan bog'liq ayrim metodologik cheklovlar aniqlandi. Xulosa sifatida, domen va hodisaga mos matn kengaytirish yondashuvi matnli ma'lumotlar yetishmovchiligi kuzatiladigan vaziyatlarda samarali yechim bo'lishi mumkinligi, biroq uning ishonchliligini oshirish uchun yanada qat'iy eksperimental baholash va zamonaviy kontekstual til modellari bilan qiyosiy tadqiqotlar o'tkazish zarurligi ta'kidlanadi.

Kalit so'zlar: matnli ma'lumotlarni kengaytirish; inqiroz tvitlarini tasniflash; domenga moslashuvchan augmentatsiya; WordNet; Word2Vec; ProMinent Words; CrisisLex; LSTM.

1. Kirish. Matnli ma'lumotlarni kengaytirish (text augmentation) masalasi tasvirlarni kengaytirishga nisbatan ancha murakkab hisoblanadi, chunki matnning sintaktik tuzilishi va semantik mazmuni bir vaqtning o'zida saqlanib qolishi talab etiladi. Ijtimoiy tarmoqlardagi qisqa, norasmiy uslubdagi matnlar xususan, Twitter platformasidagi tvitlar bu vazifani yanada murakkablashtiradi, chunki ularda qisqartmalar, xeshteglar va notekis grammatik qurilmalar ko'p uchraydi. Ko'rib chiqilayotgan tadqiqotning markaziy muammosi shundan iboratki, tabiiy ofatlar (zilzila, sel va h.k.) davomida yozilgan «yordam kerak» yoki «yordam taklif etiladi» toifasidagi tvitlar juda kam sonli yorliqlangan namunalarga ega, bu esa chuqur o'rganish modellarining ushbu toifalarni ishonchli tasniflashiga to'sqinlik qiladi. Mualliflar ushbu muammoni klassik sinonim almashtirish yoki tarjima orqali qayta tarjima kabi umumiy usullar bilan emas, balki domen (inqiroz sohasi) va joriy hodisa (masalan, ma'lum bir zilzila) kontekstini hisobga oluvchi maxsus mexanizm orqali yechishni taklif qiladi. Tadqiqot gipotezasi quyidagicha shakllantirilishi mumkin: agar kengaytirilgan matnlar nafaqat leksik jihatdan xilma-xil, balki domen va hodisaga tegishli atamalarni ustuvor tarzda saqlab qolsa, natijada olingan sun'iy namunalar tasniflagich uchun «informativroq» bo'ladi va bu klassifikatsiya sifat ko'rsatkichlarining sezilarli o'sishiga olib keladi. Ushbu gipotezani tekshirish uchun mualliflar uchta mustaqil kengaytirish strategiyasini (WordNet asosida, Word2Vec asosida va PM-Words xususiyati asosida) ishlab chiqib, ularni bitta umumiy quvur liniyasi (pipeline) doirasida birlashtiradi. Mazkur sharh maqolasining maqsadi - asl ishning metodologik yechimlarini tanqidiy tahlil qilish, taklif etilgan algoritmlarning ichki mantiqiy asoslarini, hisoblash murakkabligini va eksperimental dalillar bazasining ishonchlilik darajasini baholashdir. Bunda tarjima yoki referat emas, balki mustaqil ilmiy sharh (critical review) formatida yondashuv qo'llaniladi.

2. Metodlar

Asl tadqiqotda taklif etilgan tizim to’rtta bosqichdan tashkil topgan: (1) tvitlarni oldindan qayta ishlash, (2) so’zlarni vektorlash, (3) PM-Words xususiyatini qurish va (4) uch xil algoritm yordamida kengaytirilgan tvitlarni generatsiya qilish. Quyida har bir bosqich alohida tahlil qilinadi.

2.1. Oldindan qayta ishlash va vektorlash

Tvitlar TweetNLP vositasi yordamida tokenizatsiya qilinib, har bir so’z uchun nutq qismi (POS) belgisi aniqlanadi, so’ngra stop-so’zlar olib tashlanadi. Bu yerda muhim metodologik qaror - POS belgilarini butun jarayon davomida saqlab qolish kengaytirish natijasida olingan matnning grammatik tuzilishini asl tvitnikiga yaqin holatda ushlab turishga xizmat qiladi. Word2Vec modeli o’z korpusi (in-domain corpus) asosida o’qitiladi, bu esa domenga xos so’z birikmalarining semantik yaqinligini umumiy, oldindan tayyorlangan modellarga qaraganda aniqroq aks ettirish imkonini beradi. Ushbu tanlov metodologik jihatdan asoslangan, chunki tashqi, keng domenli embeddinglar inqiroz-spetsifik leksikani yetarlicha qamrab olmasligi mumkin.

2.2. ProMinent Words (PM-Words) xususiyati

PM-Words - mualliflarning asosiy hissasi hisoblanadi. Bu xususiyat CrisisLex lug’ati, xeshtag ro’yxati, Word2Vec bo’yicha eng o’xshash so’zlar ro’yxati va chastotaviy mezon asosida shakllantirilgan nomzod ro’yxatining kesishmasidan iborat. Nomzod ro’yxatiga chastotasi mediandan yuqori bo’lgan so’zlar kiritiladi, bu esa statistik jihatdan oddiy, ammo samarali filtrlash mexanizmi hisoblanadi. PM-Words ro’yxatining uzunligi lug’atdagi noyob so’zlar sonining foizi (tajribada 3%) sifatida belgilanadi bu gipermetr tanlovi keyinchalik natijalarga sezilarli ta’sir ko’rsatishi mumkin bo’lgan joy bo’lib, ammo maqolada ushbu foizning optimalligini asoslovchi tizimli grid-qidiruv yoki sezuvchanlik tahlili keltirilmagan.

2.3. Uchta kengaytirish algoritmi

Birinchi algoritm (WordNet asosida) har bir so’z uchun avval giperonim (Algoritm 4), so’ngra sinonim (Algoritm 5) qidiradi; giperonim CrisisLex atamalari (Olteanu va boshq., 2014) bilan mos kelgan taqdirdagina qo’llaniladi, bu esa domenga tegishlilikni ta’minlaydigan filtr vazifasini bajaradi. Ikkinchi algoritm (PM-Words asosida) xuddi shu giperonim tekshiruvidan so’ng, agar mos so’z topilmasa, sinonimni PM-Words ro’yxati bilan solishtiradi (Algoritm 6). Uchinchi algoritm (Word2Vec asosida) esa POS mosligini talab qilmasdan, faqat embeddinglar fazosidagi eng yaqin qo’shni so’zni tanlaydi (Algoritm 7). Barcha holatlarda xeshteg bilan boshlangan so’zlar o’zgarishsiz qoldiriladi bu qaror hodisaga oid identifikatorlarni (masalan, #NepalEarthquake) yo’qotmaslik uchun mantiqan asoslangan. Hisoblash murakkabligi nuqtai nazaridan, giperonimga asoslangan tekshiruv (Algoritm 4) uch qatlamli ichma-ich sikl (giperonimlar, ularning lemmalari va CrisisLex atamalari bo’yicha) tufayli $O(h \cdot l \cdot c)$ tartibida baholanadi, bu esa amaliyotda h va l qiymatlari kichik bo’lsa-da, kengaytirilishi kerak bo’lgan tvitlar soni ortishi bilan chiziqli bo’lmagan o’sishga sabab bo’lishi mumkin. Word2Vec asosidagi algoritm, aksincha, oldindan hisoblangan «eng o’xshash so’z» funksiyasiga tayangani sababli amortizatsiyalangan $O(1)$ qidiruvga yaqinlashadi, bu uning amaliy tezligini oshiradi, biroq POS mosligi talab qilinmagani semantik jihatdan mos kelmaydigan (masalan, ot o’rniga fe’l) almashtirishlar xavfini oshiradi.

2.4. Tasniflash bosqichi va tajriba dizayni

Kengaytirilgan va asl tvitlar birlashtirilib, LSTM (32 tugun, tanh faollashtirish funksiyasi), shuningdek qo’shimcha tekshiruv sifatida MLP va CNN arxitekturalari bilan ikkilik tasniflash (inqiroz - inqiroz emas) amalga oshiriladi. Salbiy sinf sifatida Olimpiada mavzusidagi tvitlar tanlangani metodologik jihatdan e’tiborga loyiq: bu domenlar orasidagi farq juda katta bo’lgani uchun tasniflash vazifasini nisbatan osonlashtirishi va shu sababli olingan yuqori aniqlik ko’rsatkichlarini haddan tashqari optimistik qilib ko’rsatishi mumkin. Haqiqiy amaliy stsenariylarda «inqiroz emas» sinfi ko’pincha boshqa yaqin mavzudagi (masalan, boshqa ijtimoiy

voqealar) tvitlardan iborat bo'ladi, bu esa model uchun ancha qiyinroq vazifani yaratadi.

3. Natijalar

Tajriba uchun asl tvitlar EMTerms terminologik resursi (Temnikova va boshq., 2015) hamda NARMADA loyihasi (Hiware va boshq., 2020) doirasida yig'ilgan zilzila voqealariga oid ma'lumotlardan olingan. Tajriba natijalari kengaytirishdan oldin va keyingi holatlarni, PM-Words xususiyati mavjud va mavjud bo'lmagan holatlarda solishtiradi.

Quyidagi jadval asl tadqiqotda keltirilgan asosiy sonli ko'rsatkichlarni umumlashtiradi.

Holat	Aniqlik (Accuracy)	Precision	Recall	F-mezon
Kengaytirishdan oldin (PM-Wordssiz)	0.839	0.689	0.981	0.810
Kengaytirishdan oldin (PM-Words bilan)	0.855	0.729	0.915	0.811
Kengaytirishdan keyin (PM-Wordssiz)	0.985	0.982	0.987	0.984
Kengaytirishdan keyin (PM-Words bilan)	0.987	0.985	0.989	0.987

Jadvaldan ko'rinib turibdiki, kengaytirishdan so'ng barcha to'rtta ko'rsatkich bo'yicha keskin sakrash kuzatiladi aniqlik 0.84 dan 0.98 dan yuqori darajaga ko'tariladi. Bunday katta farq (13–29 foiz oralig'ida) ikki xil talqinga ega bo'lishi mumkin: birinchidan, bu haqiqatan ham domen-hodisa mos kengaytirishning samaradorligini ko'rsatishi mumkin; ikkinchidan, bu shuningdek ma'lumotlar to'plamining sun'iy ravishda kattalashishi (2612 tadan 10448 tagacha) va aynan bir xil manba tvitlaridan olingan uchta variantning yuqori korrelyatsiyalanganligi hisobiga model uchun vazifaning osonlashib qolishi natijasi bo'lishi ham mumkin. Asl maqolada train/test bo'linishi tasodifiy (80/20) amalga oshirilgani, biroq kengaytirilgan versiyalar bitta asl tvitdan hosil bo'lgani sababli, agar bitta tvitning uchta varianti turli qismlarga (train va test) tushib qolsa, bu ma'lumotlar oqishi (data leakage) xavfini keltirib chiqarishi mumkin bu masala maqolada aniq muhokama qilinmagan.

PM-Words xususiyatining qo'shimcha ta'siri nisbatan mo''tadil, taxminan 0.2–4 foiz oralig'ida ko'rsatkichlarni yaxshilaydi. Bu shuni ko'rsatadiki, asosiy yutuq ma'lumotlar hajmining kengayishidan kelib chiqadi, PM-Words esa qo'shimcha, nisbatan kichik hissa qo'shuvchi omil sifatida namoyon bo'ladi. Uchta chuqur o'rganish arxitekturasi (MLP, CNN, LSTM) bo'yicha o'tkazilgan qo'shimcha tekshiruv barcha modellarda o'xshash yaxshilanish tendensiyasini tasdiqlaydi (aniqlik taxminan 0.83–0.84 dan 0.98–0.99 gacha), bu natijaning arxitekturaga bog'liq bo'lmagan barqarorligini ko'rsatadi.

So'zlarni almashtirish statistikasi ham qiziqarli naqshni ochib beradi: WordNet asosidagi algoritmda so'zlarning 58 foizi sinonim, atigi 1 foizi giperonim orqali almashtiriladi, qolgani xeshteg yoki o'zgarishsiz so'zlardir. PM-Words asosidagi algoritmda almashtirish ulushi sezilarli kamayadi (taxminan 22 foiz), bu PM-Words filtrining ancha qattiqroq, tanlab olishga yo'naltirilgan tabiatidan dalolat beradi. Word2Vec asosidagi algoritm esa deyarli barcha so'zlarni (92%) almashtiradi, bu esa uning eng yuqori leksik xilma-xillikni, biroq eng past semantik nazoratni

ta’minlashini ko’rsatadi.

4. Muhokama

Taklif etilgan yondashuvning asosiy afzalligi shundaki, u kengaytirish jarayonini butunlay tasodifiy emas, balki domen bilimiga (CrisisLex) va statistik hodisaga xoslikka (PM-Words) yo’naltirilgan holda amalga oshiradi. Bu POS mosligi va giperonim-sinonim filtrlash orqali sintaktik izchillikni saqlab qolish bilan birga, generatsiya qilingan matnning asl tvit mavzusidan chetga chiqib ketmasligini ta’minlaydi. Uchta mustaqil algoritmning parallel qo’llanilishi esa ansambl tabiatidagi xilma-xillikni beradi: WordNet leksik aniqlikni, Word2Vec kontekstual moslikni, PM-Words esa hodisaga xos terminologik barqarorlikni ta’minlaydi. Bu, masalan, kontekstual til modeli asosida so’z bashorat qiluvchi kengaytirish usuli (Kobayashi, 2018) bilan solishtirganda, domenga xos cheklovlarni aniq kiritish imkonini beruvchi ustunlikka ega, ammo o’sha usuldagi kabi chuqur kontekstual bashorat o’rniga sodda chastotaviy va lug’aviy filtrlarga tayanadi. Shu bilan birga, bir qator metodologik cheklovlar mavjud. Birinchidan, faqat «birinchi ma’no» (first sense) sinonimining tanlanishi soddalashtirilgan yechim bo’lib, WordNet’dagi ma’no tartiblanishi chastotaga asoslangan umumiy til uchun mo’ljallangan, ijtimoiy tarmoq kontekstidagi ma’noni har doim ham to’g’ri aks ettiravermaydi. Ikkinchidan, giperonim asosidagi filtrlash faqat CrisisLex ro’yxatidagi 380 ta so’z bilan cheklangan, bu lug’atning cheklangan qamrovi tufayli ko’plab potentsial mos giperonimlarning rad etilishiga olib kelishi mumkin. Uchinchidan, yuqorida qayd etilganidek, bitta asl tvitdan uchta bog’liq versiya yaratilishi statistik mustaqillik talabini buzadi va aniq train/validation/test ajratish protokoli tasvirlanmagani natijalarning takrorlanuvchanligiga shubha tug’diradi. Yana bir muhim jihat baholash metodologiyasining o’zi. Ikkilik tasniflash vazifasida salbiy sinf sifatida mavzu jihatidan mutlaqo boshqa (sport) domenning tanlanishi, klassifikatorga oson ajratiladigan xususiyatlar (masalan, leksik markerlar) taqdim etadi. Bu amaliy foydalanishda, ya’ni haqiqiy vaqtli oqimda turli xil, ko’pincha bir-biriga yaqin mavzudagi tvitlar orasida farqlash zarurati tug’ilganda, ko’rsatilgan yuqori aniqlik ko’rsatkichlarining amalda takrorlanishiga shubha bilan qarashga asos beradi. Shunga qaramay, mualliflarning kengaytirish algoritmi domendan mustaqil, umumlashtiriladigan dizaynga ega ekanligi istalgan domenning atama ro’yxati bilan almashtirilishi mumkinligi yondashuvning amaliy qiymatini oshiradi va uni faqat inqiroz tvitlari bilan cheklanmagan holda qo’llash imkonini beradi.

Amaliy ahamiyat nuqtai nazaridan, ushbu yondashuv favqulodda vaziyatlarga javob berish tizimlari uchun dolzarb, chunki real vaqt rejimida yangi hodisa (masalan, yangi zilzila) yuz berganda, mavjud yorliqlangan ma’lumotlar deyarli har doim yetarli emas. PM-Words kabi tezkor, hisoblash jihatidan yengil mexanizmlar orqali domenga moslashtirilgan kengaytirish, og’ir qayta o’qitish yoki qo’lda annotatsiyalashga qaraganda ancha tezkor yechim bo’lib xizmat qilishi mumkin. Biroq amaliy joriy etishdan oldin, yuqorida qayd etilgan ma’lumotlar oqishi va baholash og’ishi masalalarini bartaraf etuvchi qo’shimcha, qattiqroq nazorat qilingan tajribalar o’tkazish tavsiya etiladi.

5. Xulosa

Ko’rib chiqilgan tadqiqot matnli ma’lumotlarni kengaytirish sohasida domen va hodisaga xoslikni tizimli ravishda hisobga oluvchi konkret, amaliy jihatdan qo’llash mumkin bo’lgan algoritmik yechimni taqdim etadi. WordNet, Word2Vec va PM-Words komponentlarining birlashtirilgan qo’llanilishi orqali generatsiya qilingan tvitlarning sintaktik tuzilishi va domenga tegishliligi saqlanib qolinadi, bu esa klassik tasodifiy sinonim almashtirish usullaridan farqli, kontekstga sezgir yondashuvni ifodalaydi. Shu bilan birga, tajriba natijalarining haddan tashqari yuqoriligi (0.98 dan ortiq aniqlik) va uni izohlovchi metodologik omillar - xususan, kengaytirilgan namunalar orasidagi yuqori korrelyatsiya va soddalashtirilgan ikkilik baholash sxemasi - natijalarni ehtiyotkorlik bilan talqin qilishni talab qiladi. Kelgusi tadqiqotlar uchun quyidagi

yo’nalishlar tavsiya etiladi: (1) asl tvit darajasida qat’iy train/test ajratish protokolini joriy etish, (2) ko’p sinfli va domen jihatidan yaqinroq salbiy namunalar bilan baholashni kengaytirish, (3) PM-Words uzunlik gipermetri (3%) bo’yicha tizimli sezuvchanlik tahlili o’tkazish va (4) transformer asosidagi zamonaviy kontekstual model (masalan, BERT asosidagi maskalangan so’z bashorati) bilan solishtirma tahlil qilish. Ushbu yo’nalishlar amalga oshirilgan taqdirda, taklif etilgan domen-hodisaga moslashuvchan kengaytirish yondashuvi favqulodda vaziyatlarda matnli ma’lumotlar tanqisligini bartaraf etishning yanada ishonchli va umumlashtiriladigan vositasiga aylanishi mumkin.

Foydalanilgan adabiyotlar

1. Ramachandran, D., & Parvathi, R. (2021). A novel domain and event adaptive tweet augmentation approach for enhancing the classification of crisis related tweets. Data & Knowledge Engineering, 135, 101913. https://doi.org/10.1016/j.datak.2021.101913

2. Fellbaum, C. (2012). WordNet. In The Encyclopedia of Applied Linguistics. Wiley-Blackwell.

3. Olteanu, A., Castillo, C., Diaz, F., & Vieweg, S. (2014). CrisisLex: A lexicon for collecting and filtering microblogged communications in crises. Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media.

4. Temnikova, I. P., Castillo, C., & Vieweg, S. (2015). EMTerms 1.0: A terminological resource for crisis tweets. Proceedings of the ISCRAM Conference.

5. Hiware, K., Dutt, R., Sinha, S., Patro, S., Ghosh, K., & Ghosh, S. (2020). NARMADA: Need and available resource managing assistant for disasters and adversities. Proceedings of the ACL Workshop on Natural Language Processing for Social Media (SocialNLP).

6. Kobayashi, S. (2018). Contextual augmentation: Data augmentation by words with paradigmatic relations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2, 452–457.