BADIIY MATNDA MUALLIF IDIOSTILINI ANIQLASHNING SUN'IY INTELLEKTGA ASOSLANGAN MODELI: ROALD DAHLNING "THE WITCHES" ASARI MISOLIDA

Saidov Akmal Azimovich
Buxoro davlat universiteti
saidakmal9596@gmail.com

Annotatsiya. *Maqolada "The Witches" asaridagi nutq qatlamlarini ajratish orqali idiostilni o'rganish modeli sinovdan o'tkaziladi. 244 fragmentda belgili n-grammalar va logistik regressiya qo'llanib, macro-F1 0,762 ga erishildi. Grafonlarni saqlashning ahamiyati va modelning talqin chegaralari ko'rsatildi.*

Kalit so'zlar: *idiostil, korpus, grafon, narratsiya, tasniflash, modellashtirish.*

Аннотация. *Представлена модель изучения идиостиля через разграничение речевых слоёв произведения «Ведьмы». На 244 фрагментах символьные n-граммы и логистическая регрессия обеспечили macro-F1 0,762. Показаны значение сохранения графонов и ограничения интерпретации результатов классификации.*

Ключевые слова: *идиостиль, корпус, графон, нарратия, классификация, моделирование.*

Abstract. *This study models idiostyle through distinguishing narrative and character speech in The Witches. Character n-grams and logistic regression achieved a macro-F1 of 0.762 on 244 passages. The analysis demonstrates the importance of preserving graphons and specifies interpretive limitations.*

Keywords: *idiostyle, corpus, graphon, narration, classification, modelling.*

Kirish

Muallif idiostilini kompyuter yordamida o'rganishda ikki vazifani farqlash zarur: matnning kimga mansubligini aniqlash va ma'lum muallifning badiiy ifoda usullarini tushuntirish. Burrowsning stilistik masofaga oid tadqiqoti chastotalarning qiyosiy imkoniyatini ko'rsatsa [2], Stamatatos matnni ifodalash va tasniflash usullarini mualliflik atributsiyasining alohida tarkibiy qismlari sifatida yoritadi [8]. Biroq tasniflashdagi yuqori natija hali badiiy uslubning to'liq izohi emas.

"The Witches" asarida bolaning birinchi shaxs bayoni, buvining izohlovchi nutqi va Bosh jodugarning keskin individuallashgan replikalari yonma-yon keladi [3]. Ularni bitta chastotaviy ro'yxatda aralashtirish personajga xos belgini butun

muallif uslubiga tenglashtirish xavfini tug‘diradi. Korpus stilistikasida miqdoriy kuzatishni kontekst va xarakter yaratish vazifasi bilan bog‘lash ushbu muammoni yumshatadi [6].

Maqolaning maqsadi — asar ichidagi nutq qatlamlarini mashinaviy o‘rganish yordamida farqlashni idiostil tahlilining tekshiriladigan boshlang‘ich bosqichi sifatida modellashtirish. Tadqiqot savoli quyidagicha: grafonlarni saqlaydigan belgili model nutq qatlamlarini qay darajada ajratadi va grafik cho‘zilish olib tashlanganda natija qanday o‘zgaradi? Ish yangi tasniflagich algoritmini emas, mavjud algoritmni matnshunoslik nazorati va badiiy talqin bilan birlashtiruvchi amaliy modelni taklif etadi.

Material va metodlar

Asosiy manba — “The Witches” asarining taqdim etilgan 80 sahifali elektron nusxasi; uning titul qismida 1983-yilgi mualliflik huquqi qaydi bor, biroq nashriyot va aniq nashr rekvizitlari ko‘rsatilmagan. Asar 22 sarlavhali bobga ajratildi. Keyingi sahifa havolalari bosma nashrga emas, aynan shu PDF sahifalariga tegishli [3]. Tadqiqot materialiga Dahlning boshqa asarlari va ularning avvalgi statistik ko‘rsatkichlari qo‘shilmadi.

PyMuPDF 1.26.7 orqali matn bloklari olindi. Sarlavhalar va xizmat yozuvlari chiqarildi; bo‘shliqlar birxillashtirilib, yumshoq ko‘chirish belgisi olib tashlandi. Tahlil uchun 20–120 so‘z-tokenli fragmentlar tanlandi. Hisobda harflar ketma-ketligi, uning ichidagi apostrof yoki defis bilan birikkan qismlar bitta token deb olindi; raqamlar va tinish belgilari token sanog‘iga kiritilmadi. Bu so‘z-tokenlar neyron til modellaridagi subtokenlarga teng emas.

N sinfiga qo‘shtirnoqsiz, bosh harf bilan boshlanib, nuqta, so‘roq yoki undov belgisi bilan tugagan bayon bloklari, shu jumladan hikoyachining ichki mulohazalari kiritildi. G va W sinflari uchun qo‘shtirnoq bilan boshlanadigan, qo‘shtirnoqlari juft hamda tashqi izohida mos ravishda buvi yoki Bosh jodugar aniq ko‘rsatilgan bloklar olindi. Bir blokdagi ayni so‘zlovchining replikalari birlashtirildi, “my grandmother said” kabi atributsiya izohlari esa model kirishidan olib tashlandi. Kontekst tekshiruvida boshqa so‘zlovchiga tegishli ikki blok va ichma-ich ko‘chirma nutqli bir blok chiqarildi. Yakuniy tanlanma 244 fragmentdan iborat bo‘ldi.

1-jadval. Tajriba tanlanmasining tarkibi

Nutq qatlami	Fragmentlar	So‘z-tokenlar	Boblar
Hikoyachi bayoni (N)	153	7 921	21
Buvi nutqi (G)	58	2 565	12
Bosh jodugar nutqi (W)	33	1 311	5
Jami	**244**	**11 797**	**22**

Izoh: boblar soni qatlamlar bo‘yicha kesishadi; jami 22 ta alohida bob qamrab olingan.

Nusxadagi noaniq qo‘shilib qolgan so‘zlar va oddiy defisli shakllar taxmin asosida ommaviy tuzatilmadi. Ayniqsa, “rrree-moof”, “vigs”, “vill” kabi personaj nutqini ifodalovchi yozuvlar saqlandi. Grafon bu yerda talaffuz yoki nutq xususiyatini ko‘rsatish uchun ataylab o‘zgartirilgan grafik shaklni anglatadi. Texnik xato bilan uslubiy chekinishni ajratish zarurati korpusdagi shovqinning natijaga ta’siri haqidagi tadqiqotlar bilan hamohangdir [4].

Taklif etilayotgan model to‘rt bo‘g‘indan iborat: manba va nutq qatlamini tekshirish; fragmentni sonli belgilar bilan ifodalash; nutq qatlamini tasniflash; farqlovchi belgini kontekstga qaytarib, idiostilga oid talqin yaratish. Mashinaviy qism TF–IDF vektorlash va ko‘p sinfli logistik regressiyadan tuzildi. TF–IDF birlikning fragmentdagi takrorlanishini uning o‘quv tanlanmasi bo‘ylab tarqalishi bilan muvozanatlashtiradi. Bu tajribada tayyor katta til modeli yoki generativ neyron tarmoq ishlatilmadi; sun’iy intellekt komponenti nazoratli mashinaviy o‘rganishdir.

$$P(c \mid x) = \exp(w_c^T x + b_c) / \Sigma_j \exp(w_j^T x + b_j), \quad c \in \{N, G, W\}.$$

Bu yerda x — fragmentning TF–IDF vektori, c — nutq qatlami, w va b — o‘rgatiladigan parametrlar. Modelning chiqishi muallifning shaxsi emas, nutq qatlami haqidagi bashoratdir. Idiostilga oid xulosa esa ushbu natijadan keyingi kontekstual talqin bosqichida shakllantiriladi.

Tajriba

Hisoblash Python 3.13.5 va scikit-learn 1.8.0 muhitida bajarildi [7]. To‘rtta variant solishtirildi: doimo o‘quv tanlanmasidagi eng katta sinfni tanlaydigan bazaviy model; 1–2 so‘zli n-grammalar modeli; 3–5 belgili n-grammalar modeli; shu belgili modelning r harfi cho‘zilishi qisqartirilgan varianti. So‘z modeli leksik birikmalarni, belgili model esa harf, bo‘shliq va punktuatsiya ketma-ketliklarini o‘rganadi. Funksional so‘zlarning uslubiy axborot tashishi ularni faqat mazmuniy leksika bilan cheklamaslikni asoslaydi [5].

Sinov besh bo‘lakli GroupKFold yordamida boblar bo‘yicha tashkil etildi: bir bobning fragmentlari ayni siklda ham o‘qitishga, ham testga tushmadi. Test boblari ketma-ketligi: 2, 5, 12, 17, 20; 3, 6, 13, 18, 19; 7, 10, 14, 16; 4, 8, 11, 21; 1, 9, 15, 22. Dastlabki ikki test bo‘lagida W sinfi yo‘q, lekin barcha o‘quv bo‘laklarida uchala sinf mavjud. Yakuniy baho bo‘laklarning oddiy o‘rtachasi emas, har bir fragment uchun bir marta olingan test bashoratlarining umumiy natijasidir.

Har siklda TF–IDF lug‘ati faqat o‘quv bo‘lagida tuzildi: min_df = 2, max_features = 10 000, sublinear_tf = True, L2 normallashtirish va kichik harfga o‘tkazish qo‘llandi. Logistik regressiyada C = 1, class_weight = balanced, solver = lbfgs, max_iter = 2000 va random_state = 42 belgilandi. Giperparametrlar test natijasiga qarab tanlanmadi. Baholashda to‘g‘ri tasniflash ulushi va uch sinfga teng og‘irlik beruvchi macro-F1 ishlatildi.

$$Macro\text{-}F1 = (F1_N + F1_G + F1_W) / 3.$$

Ablatsiya — bitta belgi qatlamining ta’sirini tekshirish sinovi — faqat ketmaket uch va undan ortiq r harfini bitta r ga almashtirishdan iborat bo‘ldi. Ushbu o‘zgartirish o‘quv va test matnlariga bir xil qo‘llandi; v shakllari va boshqa xususiyatlar saqlandi. Bu asarni imloviy tahrirlash emas, grafonning tasniflashdagi hissasini tekshiruvchi sun’iy nazorat variantidir.

Natijalar

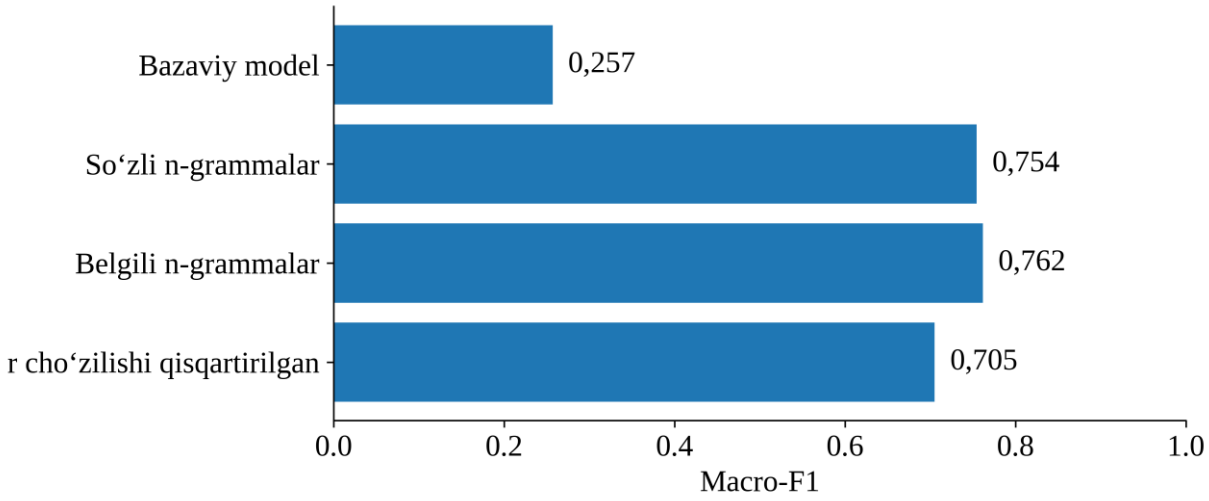
Belgili model 244 fragmentning 196 tasini to‘g‘ri tasnifladi: aniqlik 80,33 foiz, macro-F1 esa 0,762 bo‘ldi (1-rasm). So‘zli model 194 fragmentni to‘g‘ri ajratdi va macro-F1 0,754 natijani berdi. Demak, mazkur sinovda belgili modelning ustunligi kichik; uni statistik jihatdan isbotlangan umumiy ustunlik sifatida talqin qilib bo‘lmaydi. Bazaviy modelning aniqligi 62,70 foiz bo‘lsa-da, macro-F1 0,257 bo‘lishi sinflar hajmi noteng bo‘lganda bitta ko‘rsatkich bilan cheklanmaslik zarurligini ko‘rsatadi.

2-jadval. Boblar bo‘yicha ajratilgan test natijalari

Variant	Aniqlik, %	Macro-F1
Bazaviy model	62,70	0,257
So‘zli n-grammalar (1–2)	79,51	0,754
Belgili n-grammalar (3–5)	80,33	0,762
Belgili model: r cho‘zilishi qisqartirilgan	77,46	0,705

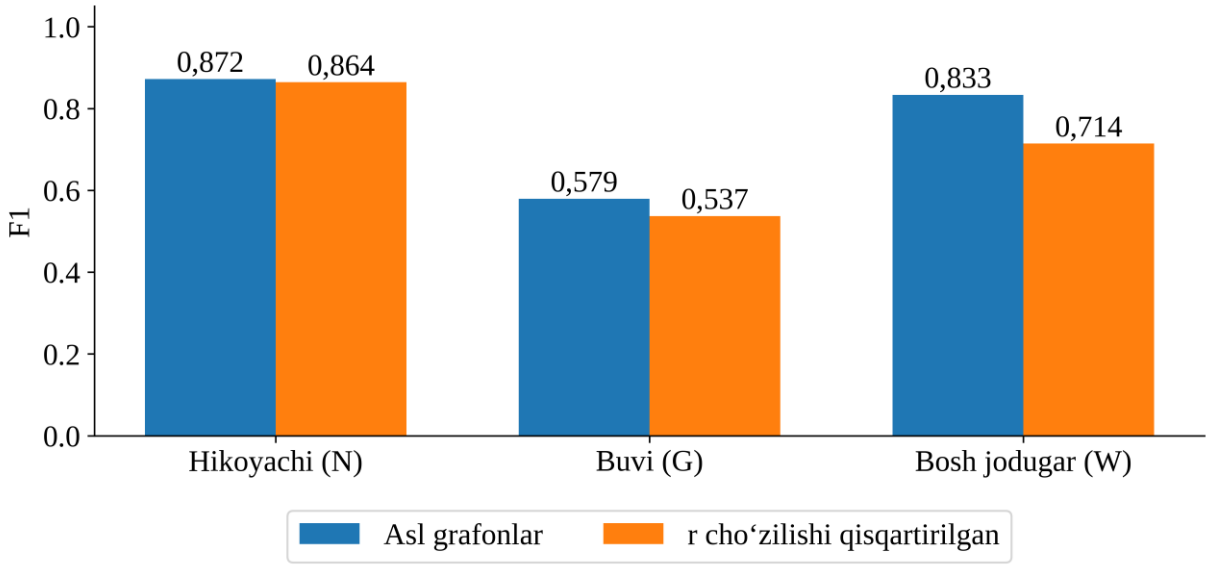
Manba: ushbu maqola uchun bajarilgan hisoblashlar; barcha variantlarda bir xil test bo‘laklari qo‘llangan.

Belgili model N sinfidagi 153 fragmentdan 140 tasini, G sinfidagi 58 fragmentdan 31 tasini va W sinfidagi 33 fragmentdan 25 tasini to‘g‘ri ajratdi. Buvining 25 fragmenti hikoyachi bayoni deb tasniflandi. Shu sababli eng muammoli chegara G–N o‘rtasida kuzatildi. G sinfi bo‘yicha so‘zli modelning F1 qiymati 0,607, belgili modelniki esa 0,579 bo‘ldi: umumiy natijadagi kichik yutuq barcha nutq qatlamlari uchun bir xil emas.



1-rasm. To‘rtta variantning macro-F1 ko‘rsatkichlari.

Faqat r cho‘zilishi qisqartirilganda belgili modelning macro-F1 qiymati 0,762 dan 0,705 ga tushdi; farq yaxlitlanmagan qiymatlar bo‘yicha 0,0565 ni tashkil etdi. W sinfining F1 ko‘rsatkichi 0,833 dan 0,714 ga pasaydi (2-rasm). Tanlangan W fragmentlarida r harfining uch yoki undan ortiq takroridan tuzilgan 80 ta ketma-ketlik bor, N va G fragmentlarida esa bunday ketma-ketlik topilmadi. Bu asarning to‘liq chastotasi emas, faqat tajriba tanlanmasidagi hisobdir.



2-rasm. r cho‘zilishini qisqartirishning sinflar F1 ko‘rsatkichiga ta’siri.

Test baholari olingach, belgilarni sharhlash uchun butun tanlanmada yakuniy model o‘rgatildi. Uning eng katta musbat koeffitsiyentli belgilarida W uchun “rrr”, “vill”; G uchun “you”, “it’s”; N uchun “was”, “were” ketma-ketliklari ko‘rindi. Bu koeffitsiyentlar model ichidagi bog‘lanishni bildiradi, badiiy ta’sirning sababini yoki belgining Dahlga mutlaq xosligini isbotlamaydi.

Muhokama

Natijalarni adabiy kontekstsiz umumlashtirish noto‘g‘ri bo‘ladi. “My Grandmother” bobida “I myself had two separate encounters with witches” jumlasi

bolaning o‘z boshidan kechirgan voqeani hikoya qilishini ko‘rsatadi [3, PDF 3-b.]. Demak, ushbu asarda birinchi shaxs bayoni mavjud; muallif idiostilini avtomatik ravishda tashqi, barcha voqeani biluvchi narratorga tenglashtirib bo‘lmaydi. Modeldagi N sinfi real muallifning bevosita ovozi emas, matnda yaratilgan hikoyachi bayonidir.

Bosh jodugarning “You may rrree-moof your vigs!” replikasidan so‘ng matnning o‘zi uning w va r talaffuzidagi xususiyatni izohlaydi [3, PDF 26-b.]. Shuning uchun bu grafik shakllarni oddiy texnik xato deb yo‘qotish personaj ovozini xiralashtiradi. Ablatsiyadagi pasayish aynan shu belgi tanlangan fragmentlarda foydali bo‘lganini ko‘rsatadi. Biroq rrr belgisi butun muallifning universal belgisi emas, avvalo bitta personaj nutqiga bog‘langan xususiyatdir.

Buvining tushuntirish va ogohlantirishlari hamda hikoyachining jodugarlar haqidagi izohlari bir xil mavzuviy leksikaga yaqinlashishi mumkin. G–N xatolari ushbu izohga mos keladi, ammo uning sababi faqat mavzu ekanini isbotlamaydi. “Witchophile”, “Mouse-Maker”, “crabcruncher”, “blabbersnitch” kabi nomlar esa fantastik dunyoni nomlashning turli usullarini ko‘rsatadi [3, PDF 15–16, 32, 38-b.]. Ularning kam uchrashi yoki modelga katta og‘irlik bilan kirmasligi badiiy ahamiyati pastligini anglatmaydi.

Tadqiqotning chegaralari aniq: material bitta asar va bitta elektron nusxa bilan cheklangan; tanlanma barcha replikalarni qamramaydi; qisqa javoblar, sahifa chegarasida bo‘lingan va so‘zlovchisi faqat olmosh orqali ko‘rsatilgan ayrim parchalar olinmagan. Mustaqil ikkinchi annotator qatnashmagani uchun annotatorlararo muvofiqlik hisoblanmadi; bunday tekshiruv keyingi bosqichda zarur [1]. Nashr artefaktlari, bob mavzusi, nutq vazifasi va sinflar nisbatining ta’siri ham to‘liq ajratilmagan [4; 8].

Shu bois olingan natija “Dahl muallifligini 80,33 foiz aniqlikda topish” emas. U tanlangan fragmentlarning nutq qatlamini tasniflash natijasidir. Idiostilning asarlararo barqarorligini tekshirish uchun Dahlning boshqa asarlari, boshqa yozuvchilarning qiyoslanadigan matnlari va kitoblar bo‘yicha ajratilgan mustaqil test kerak. Hozirgi modelning amaliy qiymati esa umumiy matn ichidagi turli ovozlarni ajratib, tadqiqotchiga aniq kontekstlarga qaytish imkonini berishidadir.

Xulosa

“The Witches” materiali asosida nutq qatlamini tekshirish, mashinaviy tasniflash va kontekstual sharhlashni birlashtiruvchi idiostil tahlili modeli taklif etildi. Uning tasniflash komponenti 244 fragmentda amalda sinovdan o‘tkazildi. Belgili n-grammalar modeli macro-F1 0,762 ga erishdi; r cho‘zilishi qisqartirilganda natijaning 0,705 ga tushishi grafonlarni ehtiyotkorlik bilan saqlash zarurligini ko‘rsatdi.

Badiiy xulosa shundan iboratki, bu asardagi individual uslubni personajlarning farqli nutqi, bolaning bayoni va fantastik nomlashning o‘zaro munosabatida o‘rganish lozim. Sun’iy intellekt ana shu farqlarni izlashni qo‘llab-quvvatlaydi, ammo statistik belgini mustaqil adabiy hukmga aylantirmaydi. Taklif etilgan yondashuv mualliflik korpuslarida nutq qatlamlarini ajratish va keyingi qiyosiy uslub tadqiqotlari uchun sinovdan o‘tgan boshlang‘ich komponent beradi.

Foydalanilgan adabiyotlar

1. Artstein R., Poesio M. Inter-coder agreement for computational linguistics // Computational Linguistics. – 2008. – Vol. 34, No. 4. – P. 555–596. – https://doi.org/10.1162/coli.07-034-R2.

2. Burrows J. “Delta”: A measure of stylistic difference and a guide to likely authorship // Literary and Linguistic Computing. – 2002. – Vol. 17, No. 3. – P. 267–287. – https://doi.org/10.1093/llc/17.3.267.

3. Dahl R. The witches. – 1983. [Tadqiqot uchun taqdim etilgan elektron nusxa].

4. Eder M. Mind your corpus: Systematic errors in authorship attribution // Literary and Linguistic Computing. – 2013. – Vol. 28, No. 4. – P. 603–614. – https://doi.org/10.1093/llc/fqt039.

5. Kestemont M. Function words in authorship attribution: From black magic to theory? // Feldman A., Kazantseva A., Szpakowicz S. (eds.). Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL). – Association for Computational Linguistics, 2014. – P. 59–66. – https://doi.org/10.3115/v1/W14-0908.

6. Mahlberg M. Corpus stylistics and Dickens’s fiction. – Routledge, 2013. – https://doi.org/10.4324/9780203076088