EasyChair Preprint

№ 11050

Fake News Classification Using Morphological Tag and N-Grams

Botir Elov, Nizomaddin Khudayberganov and Zilola Khusainova

October 9, 2023

MORFOLOGIK TEG VA N-GRAMMLAR VOSITASIDA SOXTA YANGILIKLARNI TASNIFLASH

Botir Elov Boltayevich[1]
texnika fanlari falsafa doktori, dotsent.
Kompyuter lingvistikasi va raqamli texnologiyalar
kafedrasi kafedra mudiri, ToshDO‘TAU,
Nizomiddin Xudayberganov Uktambay o‘g‘li[2]
Kompyuter lingvistikasi va raqamli texnologiyalar
kafedrasi o‘qituvchisi, ToshDO‘TAU,
Zilola Xusainova Yuldashevna[3]
ToshDO‘TAU tayanch doktoranti

Annotatsiya. Bugungi kunda ijtimoiy tarmoqlardagi soxta yangiliklarni samarali aniqlash usullarini o‘rganish juda muhim va dolzarb vazifa hisoblanadi. Ushbu usullar ko‘plab tadqiqot sohalarida, jumladan morfologik tahlilda o‘rganiladi. Ba’zi NLP tadqiqotchilarning ta’kidlashicha, oddiy kontent bilan bog‘liq n-gramlar va POS teglash orqali soxta yangiliklarni tasniflash uchun etarli emas. Biroq, ular so‘nggi o‘n yillikda bu bayonotlarni eksperimental ravishda tasdiqlashi mumkin bo‘lgan hech qanday empirik tadqiqot natijalarini olmaganlar. Ushbu qarama-qarshilikni hisobga olgan holda, maqolaning asosiy maqsadi soxta va haqiqiy yangiliklarni to‘g‘ri tasniflash uchun **n-gramlar** va **POS teglashdan** umumiy foydalanish imkoniyatlarini eksperimental baholashdan iborat. Korpus matnlarini POS teglarning n-grammlari aniqlandi va keyinchalik tahlil qilindi. Soxta yangiliklarni aniqlashning dastlabki ishlov berish bosqichida n-grammlarning turli guruhlariga POS teglash asoslangan uchta usul taklif qilindi va qo‘llanildi. Shu maqsadda n-gramm o‘lchami birinchi bo‘lib tekshirildi. Aniqlangan n-grammlar asosida yetarli darajada umumlashtirish uchun qaror daraxtlarining eng mos chuqurligi aniqlandi. Nihoyat, tavsiya etilgan usullarga asoslangan modellarning ishlash ko‘rsatkichlari standartlashtirilgan TF-IDF qiymatlari bilan taqqoslandi. **Aniqlik (precision)**, **recall** va **f1-score** kabi modelning samaradorlik ko‘rsatkichlari bir necha marta tekshiridi. Shunigdek, TF-IDF usulini POS teglash yordamida yaxshilash mumkinmi degan savol batafsil o‘rganildi. Tadqiqot natijalari shuni

---

[1] *Elov Botir Boltayevich* – texnika fanlari falsafa doktori, dotsent. Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti.
E-pochta: elov@navoiy-uni.uz
ORCID: 0000-0001-5032-6648
[2] *Xudayberganov Nizomaddin Uktambay o‘g‘li* – o‘qituvchi. Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti.
E-pochta: nizomaddin@navoiy-uni.uz
[3] *Xusainova Zilola Yuldashevna* – Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti doktoranti.
E-pochta: xusainovazilola@navoiy-uni.uz
ORCID: 0000-0003-4357-7515

ko‘rsatdiki, yangi taklif qilingan metod an’anaviy TF-IDF texnikasi bilan solishtirilganda aniqroq ko‘rsatkichlarni qayd etdi. Xulosa sifatida morfologik tahlilning asosiy TF-IDF metodini yaxshilashi mumkinligini aytish mumkin.

Kalit so‘zlar: *Soxta yangiliklarni aniqlash, matnni intellektual tahlil qilish, tabiiy tilni qayta ishlash, POS teglash, morfologik analiz.*

Kirish

Soxta yangiliklar hozirda rivojlangan dunyoning eng katta muammolaridan biri hisoblanadi [1;107]. Shaxsiy yoki siyosiy manfaatlar uchun yolg‘on ma’lumot yoki yolg‘on xabarlarni tarqatish, albatta, yangilik bo‘lmasa-da, ijtimoiy media kabi hozirgi tendentsiyalar har bir shaxsga har qachongidan ham oson yolg‘on ma’lumot yaratish imkonini beradi [2;213]. Maqolada morfologik tahlildan foydalangan holda o‘zbek tilidagi soxta va haqiqiy yangiliklarni tasniflash uchun to‘rtta taklif qilingan modelni baholash haqida so‘z boradi.

Morfologik analiz tabiiy tilni qayta ishlash tadqiqotining asosiy vositalaridan biridir. U kontekstdagi so‘zning morfologik xususiyatlari sifatida POS teglari bilan bog‘liq bo‘lib, ilmiy tadqiqotlarda **uslubga asoslangan usul** sifatida ta’riflangan [3; 3207]. Lingvistik funksiyalar orqali matn tarkibidan *belgilar, so‘zlar, gaplar* va *hujjatlar* kabi turli darajadagi strukturlangan ma’lumotlar aniqlanadi. Gap darajasidagi funksiyalar gaplar miqyosiga asoslangan barcha muhim atributlarni aniqlaydi. Bu turdagi funksiyalarga *POS teglar, gapning o‘rtacha uzunligi, tvit/postning o‘rtacha uzunligi, tinish belgilarining chastotasi, gapdagi manoga ega so‘z birikmalari* va *iboralar, gapning o‘rtacha qutbliligi (ijobiy, neytral yoki salbiy)*, yoki *gapning murakkabligini aniqlash* kabilarni misol sifatida keltirish mumkin [4;6].

Mavjud ilmiy tadqiqot ishlarida asosan, noto‘g‘ri (yolg‘on) ma’lumotlarning ichki xususiyatlarini aniqlash maqsadida standart lingvistik xususiyatlarni, jumladan **leksik, sintaktik, semantik** va **diskurs** xususiyatlarini o‘rganadi. Sintaktik xususiyatlarni *POS teglar, tinish belgilar* va *chuqur sintaktik chastotalar* kabi guruhlarga ajratish mumkin [5;172]. Ushbu maqolada n-grammga asoslangan POS teglash orqali o‘zbek tilidagi matnni morfologik analiz qilish orqali soxta yangiliklarni tasniflash masalai ko‘rib chiqiladi.

N-gramm – N ta token (so‘z)lardan iborat ketma-ketligidir. Matndagi N-grammlar *ko‘p so‘zli iboralar* yoki *leksik birliklar* sifatida aniqlanadi. Quyidagi so‘z birikmalari mos ravishda 2- va 3-grammni ifodalaydi: *"Amir Temur"* , *"Katta Buxoro kanali"*. Ko‘p hollarda matndagi alohida so‘zlarni (tokenlarni) tahlil qilishdan ko‘ra **N-grammlarni tahlil qilish** samarali natijalarni qaytaradi. Ba’zi ilmiy tadqiqot ishlarida oddiy kontent bilan bog‘liq n-gramlar va POS teglash usulining tasniflash vazifasi uchun yetarli emasligi isbotlangan [6;2,7;1783]. Biroq, bu asosan mualliflarning fikrini aks ettiradi xolos. Chunki ular so‘nggi o‘n yillikda ushbu bayonotlarni tasdiqlovchi hech qanday empirik tadqiqot natijalarini tushunmagan yoki nashr etmagan.

Ushbu qarama-qarshilikni hisobga olgan holda, maqolaning asosiy maqsadi soxta va haqiqiy yangiliklarni to‘g‘ri tasniflash uchun n-gramlar va POS teglashdan

foydalanish imkoniyatlarini eksperimental baholashdan iborat. Shu sababli, POS teglashning (n-gramm) berilgan namunasidan **n** ta elementning uzluksiz ketma-ketligi tahlil qilindi. Ushbu maqsadga erishish uchun POS teglariga asoslangan usullar taklif qilingan va ishlatilgan. Keyingi qadamlarda, ushbu usullar matn xususiyatlarini baholash uchun standartlashtirilgan mos yozuvlar TF-IDF metodi bilan taqqoslandi. Shuningdek, TF-IDF metodi natijasi samaradorligini va aniqligini POS teglash usuli yordamida yaxshilash mumkinligi batafsil o‘rganiladi [7;1784,8;12,9;32]. Maqolada keltiriladigan barcha usullarni matnga dastlabki ishlov berish bosqichida qo‘llash mumkin. Olingan natijalar to‘plami qarorlar daraxti tasniflagichlari yordamida tahlil qilinadi.

Maqola tanlangan klassifikatorning kirish vektorlarini oldindan qayta ishlash uchun tavsiya etilgan usullarni taqdim etish va baholashga qaratilgan. Ushbu usullar matn POS teglaridan foydalangan holda n-grammlarni shakllantirishga asoslangan. Barcha tavsiya etilgan usullar n-grammning turli darajalariga qo‘llanilgan bo‘lib, ushbu usullarning natijalari qarorlar daraxti tasniflagichining kirish vektorlari sifatida ishlatilgan. POS teglashning n-grammlarga asoslangan taklif qilingan yondashuvning muvofiqligini baholash uchun quyidagi metodologiya qo‘llanilgan:

- *Tahlil qilingan ma’lumotlar to‘plamida POS teglarni aniqlash;*
- *POS teglaridan foydalanib N-gramm (1 gramm, 2 gramm, 3 gramm, 4 gramm) larni aniqlash. N-gramm POS teglar ketma-ketligini ifodalaydi.*
- *Hujjatlarda n-grammlar chastotasini hisoblash. Ya’ni, tekshirilgan soxta va haqiqiy yangiliklarda n-grammning nisbiy chastotasi hisoblanadi.*
- *POS teglash va boshqariladigan TF-IDF metodi uchun tavsiya etilgan uchta usuldan foydalangan holda tasniflagichlarning kirish vektorlarini aniqlash;*
- *Qarorlar daraxti tasniflagichlarini qo‘llash. N-grammlarning turli chuqurliklari va uzunligi bo‘yicha parametrlarni sozlash;*
- *Qaror daraxtlarining xususiyatlarini, asosan daraxtlarning aniqligi, chuqurligi va vaqt ko‘rsatkichlarini aniqlash va taqqoslash.*

POS teglar

Ma’lumotlar to‘plamidagi yangiliklardagi barcha so‘zlariga TreeTagger nomli vosita yordamida morfologik teglar tayinlangan. Shmid 1994 yilda English Penn Treebank deb nomlangan teglar to‘plamini ishlab chiqqan. Ingliz tilidagi to‘liqroq Penn Treebank teglar to‘plami 35 ta morfologik tegni o‘z ichiga oladi [10;8,11;56]. Biroq, tadqiqot maqsadini hisobga olgan holda, ba’zi teglar paydo bo‘lishining past chastotasi yoki nomuvofiqligi sababli keyingi tahlillarga kiritilmagan. Shuning uchun tahlilda foydalanilgan morfologik teglarning yakuniy ro‘yxati quyidagi 1-jadvalda keltirilgan [11;53,12;43,13;58]:

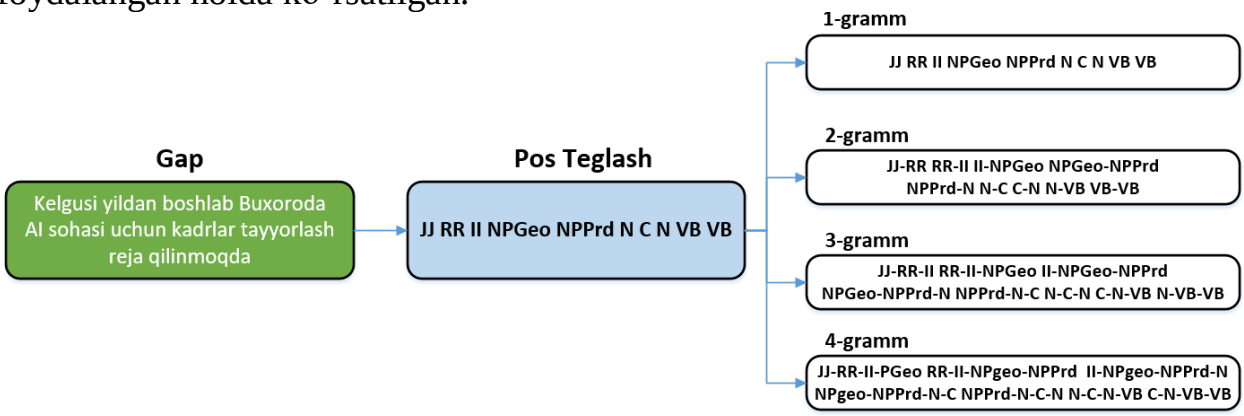
1-jadval. *Yangiliklarni tasniflashda foydalaniladigan morfologik teglar*

GTAG	Pos teglar
C guruh	C (bog‘lovchi), NUM (sanoq son)
E guruh	EX (biror narsani tasdiqlash uchun ishlatiladi)
F guruh	NP (Neologizm)
I guruh	II (ko‘makchi so‘zlar)

J guruh	JJ (sifat), JJR (sifat, Qiyosiy daraja), JJT (sifat, orttirma daraja)
M guruh	MD (modal)
N guruh	N (ot, birlik), N (ko‘plikdagi ot), NP (atoqli ot, nomlar), NP (atoqli ot, ko‘plik)
P guruh	PInd (belgilash olmoshi), PossP (egalik kategoriyasi), PP (kishilik olmoshi)
R guruh	RR (ravish), RRR (ravish, qiyosiy), RRT (ravish, orttirma), Prt (yuklama)
T guruh	(fe’lning lug‘at shakli “moq”)
U guruh	UH (undov so‘zlar)
V guruh	VB (fe’l), TPast (fe’l, o‘tgan zamon), VBS (fe’l, sifatdosh) Prs3s (fe’l, birlik, hozirgi zamon), VBZ (fe’l, 3-shaxs birlik. Hozirgi zamon)
W guruh	PQues (aniqlovchi so‘zlar, so‘roq olmoshi)

POS teglardan N-grammni aniqlash

Ushbu bosqichda ma’lumotlarni boshlang‘ich qayta ishlash POS teglaridan N-grammlar olingan. Natijada, POS teglarining berilgan namunasidan n-gramm ketma-ketligi shakllantirildi. Quyidagi 1-rasmda bu jarayon 2019-yilda Facebookda baham ko‘rilgan eng ko‘p ko‘rilgan o‘ninchi soxta yangiliklardan olingan gaplardan foydalangan holda ko‘rsatilgan.



1-rasm. Soxta yangilikni POS teglash va N-grammlarga ajratish.

Yuqorida keltirilgan *“Kelgusi yildan boshlab Buxoroda AI sohasi uchun kadrlar tayyorlash reja qilinmoqda”* gapiga mos aniqlangan POS teglar quyidagicha:

- **JJ** (Sifat),
- **RR** (Ravish),
- **II** (Ko‘makchi),
- **NPGeo** (Geografik nom),
- **NPPrd** (Mahsulot nomi),
- **N** (Ot),
- **C** (Bog‘lovchi),
- **VB** (Fe’l).

1-gramm va aniqlangan POS teglar bir xil bo‘lganligi sababli, keyingi tadqiqotlarda ishlatiladigan 1-grammli kirish fayli aniqlangan POS teglari bilan bir xil bo‘ladi. TF-

IDF usuli uchun n-grammlar xuddi shu tarzda shakllantirilgan. Ammo shuni ta’kidlash kerakki, ushbu usulda so‘zning **lemmalari** yoki **stemlarini** ifodalovchi terminlar ishlatilgan.

Kirish vektorlarini boshlang‘ich qayta ishlash usullari

Tanlangan klassifikator uchun kirish vektorlarini boshlang‘ich qayta ishlash uchun quyidagi to‘rtta usul qo‘llanilgan.

Term frequency - inverse document frequency (TF-IDF) usulidan tokenlarning korpus hujjatlaridagi ahamiyatini baholash uchun foydalaniladi [7]. TF-IDF yondashuvi odatda *“shovqin”* sifatida identifikatsiya qilinadigan ma’lum bir domen bilan yuqori darajada bog‘liq bo‘lgan ko‘p ishlatiladigan terminlarani aniqlash uchun ishlatiladi. An’anaviy TF-IDF usuli katta hajmdagi ma’lumot (yangilik)larni qayta ishlash uchun qo‘llanilmaydi. Odatda, TF-IDF og‘irligi ikki elementdan iborat: birinchisi terminning normalangan chastotasi (Term Frequency, TF), ikkinchisi esa teskari hujjat chastotasi (Inverse Document FrequencyIDF). Quyidagi belgilanishlarni aniqlab olamiz:

- **t** – temin/so‘z;
- **d** – hujjat;
- **w** – hujjatdagi istalgan termin.

d hujjatdagi **t** termin/so‘zning chastotasi quyidagi formula orqali hisoblanadi:

$$\mathbf{tf}(\mathbf{t},\mathbf{d}) = \frac{\boldsymbol{f}(\boldsymbol{t},\boldsymbol{d})}{\boldsymbol{f}(\boldsymbol{w},\boldsymbol{d})}$$

Bu yerda $\boldsymbol{f}(\boldsymbol{t},\boldsymbol{d})$ – **d** hujjatdagi termin/so‘zlar soni va $\boldsymbol{f}(\boldsymbol{w},\boldsymbol{d})$ – hujjatidagi barcha ternminlar soni. Shuningdek, TF-IDF qiymatni hisoblashda ma’lum bir termin/so‘z sodir bo‘lgan barcha hujjatlar soni ham hisobga olinadi. Bu qiymat $\boldsymbol{idf}(\boldsymbol{t},\boldsymbol{D})$ kabi belgilanadi va uning qiymati quyidagi formula orqali aniqlanadi:

$$\mathbf{idf}(\mathbf{t},\mathbf{D}) = \mathbf{ln}\frac{\boldsymbol{N}}{\sum(\boldsymbol{d}\in\boldsymbol{D}:\boldsymbol{t}\in\boldsymbol{d})+\mathbf{1}}$$

Bu yerda, **D** – hujjatlar korpusi va **N** – korpusdagi hujjatlar soni.

Tf-Idf qiymati quyidagi formula orqali aniqlanadi:

$$\mathbf{Tf}-\mathbf{Idf}(\mathbf{t},\mathbf{d},\mathbf{D}) = \mathbf{tf}(\mathbf{t},\mathbf{d})\cdot\mathbf{idf}(\mathbf{t},\mathbf{D})$$

Tf formulasi turli xil variantlarga ega bo‘lib, masalan

$$\mathbf{log}(\mathbf{tf}(\mathbf{t},\mathbf{d})) \text{ yoki } \boldsymbol{log}\big(\mathbf{tf}(\mathbf{t},\mathbf{d})\big)+\mathbf{1}$$

Xuddi shunday, **idf** qiymatni hisoblash mumkin bo‘lgan bir nechta variantlari mavjud. Yuqorida keltirilgan formulalarga mos **Tf-Idf** qiymatni hisoblash uchun Pythondagi **scikit-learn**[4] kutubxonasidan foydalanish mumkin. Tajribada qo‘llaniladigan TF-IDF usuli orqali tavsiflangan tanlangan xususiyatlarini taqqoslash uchun ishlatiladi. Xuddi shu ma’lumotlar to‘plami keyingi qadamlar uchun kirish sifatida ishlatilgan. Biroq, bu holatda nomuhim so‘zlari olib tashlangan.

[4] https://scikit-learn.org

POS chastotasi (PosF) usuli

Ushbu usul Term Frequency usuliga o‘xshashidir. Biroq, u POS teglar chastotasi bilan hisoblab chiqadi. Quyidagi belgilanishlarni aniqlab olamiz:

- **pos** – identifikatsiyalangan POS teg;
- **d** – hujjat;
- **w** – hujjatda aniqlangan har qanday POS teg.

Bu holda **d** hujjatidagi **POS** teglar chastotasini quyidagicha hisoblash mumkin:

$$\mathbf{PosF}(\boldsymbol{pos}, \mathbf{d}) = \frac{\boldsymbol{f}(\boldsymbol{pos}, \boldsymbol{d})}{\boldsymbol{f}(\boldsymbol{w}, \boldsymbol{d})} \ (\mathbf{3})$$

Bu yerda $\boldsymbol{f}(\boldsymbol{pos}, \boldsymbol{d})$ – **d** hujjatdagi POS teglar soni va $\boldsymbol{f}(\boldsymbol{w}, \boldsymbol{d})$ – hujjatidagi barcha POS teglar soni. Yuqorida keltirilgan 3-formula orqali hujjatda aniqlangan POS teglarining tahlil qilingan ro‘yxati doirasidagi har bir POS tegining nisbiy chastotasini ifodalaydi.

PosF-IDF usuli

Ushbu usul TF-IDF usulining analogidir. Yuqorida keltirilgan PosF usuliga o‘xshab, u alohida so‘zlar va gaplar asosida tahlil qilingan ma’lumotlar to‘plamidagi har bir hujjatda aniqlangan POS teglarini ko‘rib chiqadi. Faqat identifikatsiyalangan POS teglardan iborat hujjatlar PosF-IDFni hisoblash uchun kerakli ma’lumotlar hisoblanadi. Hujjatdagi POS teglarining nisbiy chastotasidan tashqari, ma’lum bir POS tegi aniqlangan barcha hujjatlar soni ham hisobga olinadi.

TF-IDF va PosF usullarini birlashtirish

Ushbu usul an’anaviy TF-IDF usulini POS teglash yordamida yaxshilash mumkinmi yoki yo‘qligini tasdiqlash uchun ishlab chiqilgan. Shu maqsadda har bir hujjat uchun quyidagi vektorlar shakllantirilgan:

- *Tf-Idf vektori;*
- *Hujjatdagi POS teglarining nisbiy chastotasini ifodalovchi PosF vektori.*

Yuqorida keltirilgan $\overrightarrow{\boldsymbol{Tf-Idf}(\boldsymbol{d})}$ va $\overrightarrow{\boldsymbol{PosF}(\boldsymbol{d})}$ vektorlarni birlashtirilgan natijasida yangi $\overrightarrow{\boldsymbol{merge}(\boldsymbol{d})}$vektori hosil qilinadi **($\boldsymbol{m} \leq \boldsymbol{n}$).**

$$\overrightarrow{\boldsymbol{Tf-Idf}(\boldsymbol{d})} = (t_1, t_2, \dots, t_n),$$
$$\overrightarrow{\boldsymbol{PosF}(\boldsymbol{d})} = (p_1, p_2, \dots, p_m),$$
$$\overrightarrow{\boldsymbol{merge}(\boldsymbol{d})} = (t_1, t_2, \dots, t_n, p_1, p_2, \dots, p_m),$$

Ma’lumotlarni tasniflashdagi kirish vektorlarini boshlang‘ich qayta ishlash uchun bir qator usullar ishlab chiqildi. Ushbu usullarni avvalgi TF-IDF usulining modifikatsiyasi deb hisoblash mumkin bo‘lib, bunda POS teglari asl terminlarga qo‘shimcha ravishda hisobga olinadi. Natijada, yuqorida tavsiflangan to‘rtta usul tipik o‘zgarishlarni ifodalash orqali terminlar va POS teglar asosida hosil qilingan gibrid usul orqali asosiy xususiyatlarini taqqoslash va tahlil qilish imkonini beradi.

Qaror daraxtlari vositasida modellashtirish

Bugungi kunda ma’lumotlarni tasniflash uchun quyidagi tasniflagichlardan foydalanish mumkin:

- *qarorlar daraxti tasniflagichlari (tree classifiers);*
- *Bayes klassifikatorlari (Bayesian classifiers);*
- *eng yaqin k-qo‘shni tasniflagichlari (k-nearest-neighbour classifiers)*;
- *holatlarga asoslangan fikrlash (case-based reasoning)*;
- *genetik algoritmlar (genetic algorithms)*;
- *qo‘pol to‘plamlar* (*rough sets)*;
- *oshkormas mantiq usullari (fuzzy logic techniques)*.

Kirish vektorlarini hisoblash uchun tavsiya etilgan usullarning mosligini baholash va ularning xususiyatlarini tahlil qilish uchun *qaror daraxtlari (decision trees)* usulini ko‘rib chiqamiz. Qaror daraxtlari nafaqat ishlarni oddiy tasniflash imkonini beradi, balki ular bir vaqtning o‘zida oson izohlanadigan va tushunarli tasniflash qoidalarini yaratadi. Xuddi shu yondashuv Kapusta, Benko va Munk tadqiqot ishlarida qisman qo‘llanilgan.

Qarorlar daraxti yaratilayotganda qo‘llaniladigan *ma’lumot olish*, *Jini indeksi* kabi atributlarni tanlash ko‘rsatkichlari muhim omili hisoblanadi. Qaror daraxtini ishlab chiqishning har bir bosqichida eng yaxshi funksiya har doim tanlanadi. Ushbu funksiya kirish atributlari soniga bog‘liq emas. Bu esa tanlangan klassifikatorning kiritilishida ko‘proq atributlar (kirish vektorining elementlari) berilgan bo‘lsa ham, aniqlik o‘zgarmasligini anglatadi.

K-o‘lchovli tekshiruv

Amalga oshirilgan tajribada shakllantirilgan qaror daraxtlarini taqqoslash uchun qaror daraxtlarining *tugunlari* yoki *barglari* soni kabi muhim xususiyatlardan foydalaniladi. Bu xususiyatlar daraxtning o‘lchamini anglatib, ularni mos ravishda kamaytirish kerak.

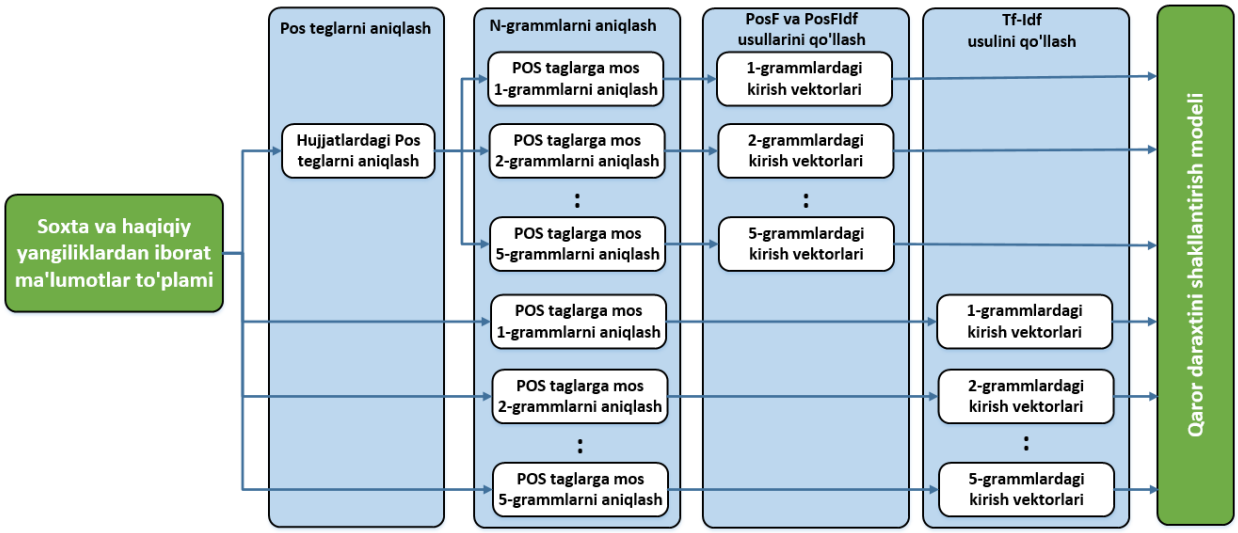
Bir vaqtning o‘zida *precision*, *recall* va *f1-score* kabi modelning ishlash ko‘rsatkichlari bir necha (10) marotaba tekshirish orqali sinovdan o‘tkaziladi. Modellarni baholash uchun *K-o‘lchovli tekshiruv*dan foydalanilgan. Bu, odatda, boshqa usullarga nisbatan kamroq noxolis modelga olib keladi, chunki u asl ma’lumotlar to‘plamidagi har bir kuzatuv mashg‘ulot va test to‘plamida paydo bo‘lish imkoniyatiga ega bo‘lishini ta’minlaydi.

n-gramm uzunligini o‘rnatish

Yuqorida keltirilgan kirish vektorlarini boshlang‘ich qayta ishlash usullarida umumiy shartlar talab etilgan. Shu sababli birinchi qadam sifatida n-grammdagi eng yuqori qiymatlar aniqlanadi. Ko‘pgina NLP vazifalarida odatda **n={1,2,3}** oraliqdagi qiymatlardan foydalaniladi. **n** ning yuqori qiymati (4 gramm, 5 gramm va boshqalar) apparat va dasturiy ta’minotga, hisoblash vaqtiga va umumiy ishlashga sezilarli murakkabliklarni yuzaga keltiradi. Boshqa tomondan, yaratilgan modellarning aniqligini oshirishda yuqori n-grammlarning potentsial hissasi cheklangan. Yuqoridagi fikrlarni tasdiqlash uchun bir nechta qarorlar daraxti modellari ishlab

chiqildi. *Tokenlar/so‘zlar* va *POS teglar* uchun N-grammlar (1 gramm, 2 gramm, ..., 5 gramm) tayyorlandi. Keyinchalik, TF-IDF usuli n-gramm tokenlar/so‘zlar uchun qo‘llanildi. Shu bilan birga, n-gramm POS teglarida **PosF** va **PosfIdf** usuli qo‘llanildi.

Natijada, kirish vektorlaridan iborat **15** ta fayl yaratildi (*1-5 gramm × 3 usul*). Quyidagi 2-rasmda ushbu jarayonning individual bosqichlari ko‘rsatilgan.



2-rasm. Kirish vektorlari tajriba jarayoni

Har bir boshlang‘ich qayta ishlangan 15 ta fayl ustida 10 marta o‘tkazilgan sinov natijasida o‘nta qaror daraxti modeli shakllantirildi. Barcha holatlarda **aniqlik darajasi** sifatida modelning samaradorlik o‘lchovi hisoblandi.

Amalga oshirilgan sinov natijalari shuni ko‘rsatadiki, aniqlik darajasi n-gramm uzunligi bilan, asosan, TF-IDF usulini qo‘llashda pasayadi. Cheklangan vaqt va hisoblash jarayoning murakkabligi tufayli kattaroq o‘lchamli n-grammlarni (6 gramm, 7 gramm va boshqalar) qayta ishlash imkoni mavjud emas. Shu sababli tadqiqotda **n=5** maksimal chegara sifatida qabul qilingan.

Qarorlar daraxti modelini ishlab chiqish jarayonida n-grammlarni bitta kirish fayliga birlashtirish eng yuqori aniqlikni ta’minlaganini qayd etish lozim. Natijada, keyingi qadamdagi barcha tajribalar 1-gramm, 2-gramm, 3-gramm va 4-gramm (1,4) dan iborat fayl ustida amallar bajariladi.

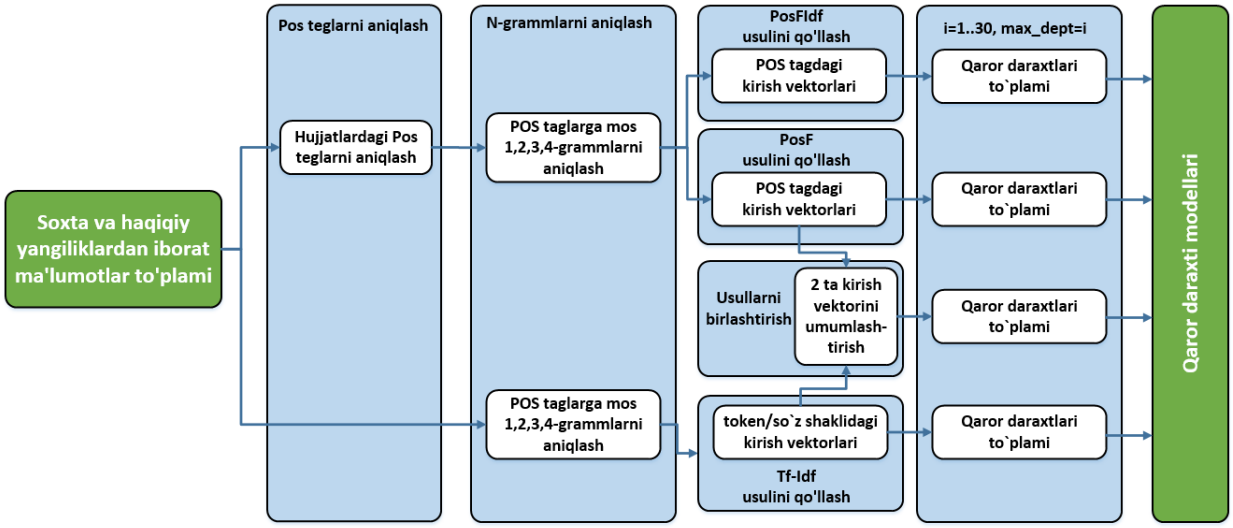
Amalga oshirilgan natijalariga ko‘ra ijtimoiy tarmoqlardagi soxta yangiliklarni samarali aniqlash uchun quyidagi ketma-ketlikda amallani bajarish lozim:

1. *Ma’lumotlar to‘plamidagi POS teglarni aniqlash.*
2. *Aniqlangan POS teglar uchun PosF va PosfIdf kirish vektorlarini qo‘llash.*
3. *Kirish vektorlarini aniqlash uchun TF-IDF usulini qo‘llash. Ushbu usul orqali stemming algoritmi tomonidan o‘zak so‘zlar aniqlanadi. Shuningdek, ushbu qadamda nomuhim so‘zlari olib tashlanadi.*
4. *PosF va TF-IDF metodlari orqlai vektorlarni birlashtirish.*
5. *Quyida keltirilgan amallarni maksimal chuqurlikning turli qiymatlari bilan takrorlash (1…n):*

– *PosF, PosfIdf, TfIdf va Merge metidlarining kirish vektorlarini 10 martalik o‘zaro tekshirish talablariga muvofiq o‘qitish va to‘plam ostilariga tasodifiy taqsimlash.*
– *Berilgan maksimal chuqurlikka ega bo‘lgan har bir mashg‘ulot to‘plami uchun qarorlar daraxtini hisoblash.*
– *Sinov kichik hajmli ma’lumotlar to‘plamida model bashoratlarining sifatini sinab ko‘rish. Quyidagi xarakteristikalar o‘rnatiladi:*
  - *prec_fake (soxta guruh uchun aniqlik);*
  - *prec_real (haqiqiy guruh uchun aniqlik);*
  - *rec_fake (soxta guruh yozuvi);*
  - *rec_real (haqiqiy guruh yozuvi);*
  - *f1-ball/baho;*
  - *har bir iteratsiyaga sarflangan vaqt.*
– *Natijalarni tahlil qilish (modellarni baholash).*

1 – 4 bosqichlar natijasi yuqorida aytib o‘tilgan to‘rtta kirish vektoridir. Taklif etilayotgan metodologiyaning beshinchi bosqichi ushbu to‘rtta tekshirilgan qadamni/natijani baholashga qaratilgan.

Taklif etilayotgan metodologiyani 10 marta o‘zaro tekshirish bilan qo‘llash natijasida 1200 ta turli xil qarorlar daraxtlari yaratildi. Quyidagi 3-rasmda tajriba metodologiyasining alohida bosqichlari tasvirlangan.



3-rasm. Tavsiya etilgan usul bosqichlari

Xulosa

Ushbu maqolada n-grammalar orqali til korpusidan yaratilgan POS teglashga asoslangan ijtimoiy tarmoqlardagi soxta yangiliklarni ishonchli tasniflash usullari tahlil qilindi. Maqolada POS teglashga asoslangan ikkita usul taklif qilindi va TF-IDF usuli asosida o‘zaro solishtirildi. Olingan natijalar PosF va TF-IDF usuli o‘rtasidagi statistik jihatdan ahamiyatsiz farqlarni ko‘rsatdi. Bu farqlar barcha kuzatilgan ishlash ko‘rsatkichlarida, jumladan, *accuracy*, *precision*, *recall* va *f1-score* asosida solishtirildi. Shu sababli, morfologik tahlilni soxta yangiliklar tasnifiga qo‘llash mumkin degan xulosaga kelish mumkin. Shuningdek, tavsiflovchi statistik jadvallar TF-IDF usuli statistik jihatdan ahamiyatsiz bo‘lsa-da, yaxshiroq

natijalarga erishishini ko‘rsatdi. Morfologik tahlilga asoslangan usullar zamonaviy ma’lumotlar to‘plamida, shu jumladan o‘zbek tili korpusidagi 1100 ta haqiqiy va yolg‘on yangiliklarda sinovdan o‘tkazildi va samarali natija qaytardi.

Foydalanilgan adabiyotlar ro‘yxati

1. Jang, S. M., Geng, T., Queenie Li, J. Y., Xia, R., Huang, C. T., Kim, H., & Tang, J. (2018). A computational approach for examining the roots and spreading patterns of fake news: Evolution tree analysis. *Computers in Human Behavior*, *84*. https://doi.org/10.1016/j.chb.2018.02.032
2. Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. In *Journal of Economic Perspectives* (Vol. 31, Issue 2). https://doi.org/10.1257/jep.31.2.211
3. Zafarani et al. (2019) Zafarani R, Zhou X, Shu K, Liu H. Fake news research: theories, detection strategies, and open problems. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD ‘19; New York, NY, USA: Association for Computing Machinery; 2019. pp. 3207–3208.*
4. Khan, J. Y., Khondaker, Md. T. I., Afroz, S., Uddin, G., & Iqbal, A. (2021). A benchmark study of machine learning models for online fake news detection. *Machine Learning with Applications*, *4*. https://doi.org/10.1016/j.mlwa.2021.100032
5. Feng, S., Banerjee, R., & Choi, Y. (2012). Syntactic stylometry for deception detection. *50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference*, *2*.
6. Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, *52*(1). https://doi.org/10.1002/pra2.2015.145052010082
7. B.Elov, Z.Xusainova, N.Xudayberganov. O‘zbek tili korpusi matnlari uchun TF-IDF statistik ko‘rsatkichni hisoblash. *SCIENCE AND INNOVATION INTERNATIONAL SCIENTIFIC JOURNAL VOLUME 1 ISSUE 8 UIF-2022: 8.2 ISSN: 2181-3337* *https://www.academia.edu/105829396/OZBEK_TILI_KORPUSI_MATNLARI_UCHUN_TF_IDF_STATISTIK_KORSATKICHNI_HISOBLASH*
8. B.ELov, Sh.Khamroeva, Z.Xusainova (2023). The pipeline processing of NLP. *E3S Web of Conferences 413, 03011, INTERAGROMASH 2023.* https://doi.org/10.1051/e3sconf/202341303011
9. Boltayevich, E. B., Mirdjonovna, H. S., & Ilxomovna, A. X. (2023). Methods for Creating a Morphological Analyzer. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 13741 LNCS*. https://doi.org/10.1007/978-3-031-27199-1_4
10. Kaing, H., Ding, C., Utiyama, M., Sumita, E., Sam, S., Seng, S., Sudoh, K., & Nakamura, S. (2021). Towards Tokenization and Part-of-Speech Tagging for Khmer: Data and Discussion. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *20*(6). https://doi.org/10.1145/3464378

11. B.Elov, Sh.Hamroyeva, O.Abdullayeva, M.Uzoqova. O‘zbek tilida pos tegging masalasi: muammo va takliflar. *O‘zbekiston: til va madaniyat (Amaliy filologiya), 2022, 5(4).*
12. B.Elov, Sh.Hamroyeva, O.Abdullayeva, Z.Xusainova, N.Xudayberganov. O‘zbek, turk va uyg‘ur tillarida pos teglash va stemming. *O‘zbekiston: til va madaniyat (Kompyuter lingvistikasi), 2023, 1(6).*
13. B.Elov, E.Adalı, Sh.Khamroeva, O.Abdullayeva, Z.Xusainova, N.Xudayberganov. The Problem of Pos Tagging and Stemming for Agglutinative Languages. *8 th International Conference on Computer Science and Engineering UBMK 2023, Mehmet Akif Ersoy University, Burdur – Turkey*