

TOSHKENT KIMYO XALQARO UNIVERSITETI NAMANGAN FILIALI

Artificial intelligence yo'nalishi talabasi Boqiboyev Nozimjon Iminjon

Pochta: nozimjon0324@gmail.com

Tel: +998972517200

MATNLARNI TAHLIL QILISH ALGORITMI

Annotatsiya. Ushbu tezisda matnlarni avtomatik tahlil qilishning asosiy bosqichlari, ishlatiladigan algoritmlar va ularning amaliy qo'llanilishi ko'rib chiqiladi. Matnni dastlabki qayta ishlash, xususiyatlarni ajratib olish, TF-IDF usuli, mashinaviy o'qitish asosidagi tasniflash hamda zamonaviy neyron tarmoq modellarining afzalliklari va kamchiliklari yoritiladi. Shuningdek, o'zbek tili matnlarini tahlil qilishdagi o'ziga xos qiyinchiliklar tahlil qilinadi.

Kalit so'zlar: matnni tahlil qilish, tabiiy tilni qayta ishlash (NLP), tokenizatsiya, TF-IDF, tasniflash, o'zbek tili.

1. Kirish

Axborot texnologiyalari rivojlanishi bilan har kuni yaratilayotgan matnli ma'lumotlar hajmi keskin oshib bormoqda. Shu jumladan ijtimoiy tarmoqlar, yangiliklar saytlari, elektron hujjatlar, foydalanuvchi sharhlari va elektron pochta xabarlari tarkibidagi ma'lumotlarning katta qismi tuzilmasiz matn ko'rinishida saqlanadi. Bunday hajmdagi matnni inson tomonidan qo'lda o'qib chiqish va tahlil qilish samarasiz va ko'p vaqt talab qiladi. Shu sababli matnlarni avtomatik tahlil qilish algoritmlarini yaratish va takomillashtirish dolzarb vazifalardan biri hisoblanadi.

Matnni tahlil qilish (Text Mining) tabiiy tilni qayta ishlash (Natural Language Processing, NLP) sohasining muhim yo'nalishi bo'lib, uning maqsadi matndan foydali ma'lumot, qonuniyat va ma'noni avtomatik ravishda ajratib olishdan iborat. Bu yo'nalish qidiruv tizimlari, spamga qarshi filtrlar, hissiyotni aniqlash, mashina tarjimai va hujjatlarni avtomatik guruhlash kabi sohalarda keng qo'llaniladi.

Tezisning maqsadi matnlarni tahlil qilishning umumiy algoritmini bosqichma-bosqich ko'rib chiqish, asosiy usullarni solishtirish va o'zbek tili uchun amaliy xususiyatlarni aniqlashdan iborat.

2. Matnni tahlil qilishning umumiy algoritmi

Matnni tahlil qilish jarayoni odatda quyidagi ketma-ketlikdan iborat bo'ladi:

1. Ma'lumotni yig'ish va yuklash.
2. Matnni dastlabki qayta ishlash (preprocessing).
3. Xususiyatlarni ajratib olish va vektorlashtirish.
4. Tahlil modelini qo'llash (tasniflash, klasterlash, ma'no ajratish).
5. Natijani baholash va sharhlash.

Har bir bosqichning sifati yakuniy natijaga bevosita ta'sir qiladi. Ayniqsa, dastlabki qayta ishlash bosqichi xatosiz bajarilmasa, eng murakkab model ham yaxshi natija bermaydi.

2.1. Matnni dastlabki qayta ishlash

Xom matn odatda ortiqcha belgilar, turli registrdagi harflar, tinish belgilari va ahamiyatsiz so'zlarni o'z ichiga oladi. Dastlabki qayta ishlash quyidagi amallardan tashkil topadi:

- *Tokenizatsiya* — matnni alohida so'z yoki gaplarga (tokenlarga) ajratish.
- *Registrni bir xillashtirish* — barcha harflarni kichik harfga o'tkazish.
- *Tozalash* — tinish belgilari, raqamlar, HTML teglar va ortiqcha bo'shliqlarni olib tashlash.
- *Stop-so'zlarni olib tashlash* — "va", "lekin", "uchun", "bilan" kabi ma'no yukiga ega bo'lmagan so'zlarni chiqarib tashlash.
- *Stemming va lemmatizatsiya* — so'zni o'zak yoki lug'aviy shakliga keltirish.

Bu amallar natijasida matn hajmi kamayadi va tahlil uchun muhim bo'lgan ma'lumot ajratib olinadi.

2.2. Xususiyatlarni ajratib olish

Kompyuter matnni to'g'ridan-to'g'ri tushuna olmaydi, shuning uchun matn sonli vektorlarga aylantiriladi. Eng keng tarqalgan usullar:

Bag-of-Words (BoW) modeli. Matn so'zlar to'plami sifatida qaraladi, so'zlar tartibi e'tiborga olinmaydi. Har bir hujjat lug'atdagi so'zlarning uchrash soni bilan ifodalanadi. Usul sodda va tez ishlaydi, ammo so'zlarning ma'no va kontekstini yo'qotadi.

TF-IDF usuli. Bu usul so'zning hujjat uchun qanchalik muhimligini baholaydi. U ikki ko'rsatkichning ko'paytmasi sifatida hisoblanadi:

$$TF-IDF(t, d) = TF(t, d) \times IDF(t)$$

bu yerda $TF(t, d)$ - t so'zining d hujjatda uchrash chastotasi, $IDF(t) = \log(N / df(t))$ - teskari hujjat chastotasi, N - hujjatlarning umumiy soni, $df(t)$ - t so'zi uchraydigan hujjatlar soni. Ko'p hujjatlarda uchraydigan so'zlar past vazn oladi, kam uchraydigan, lekin ma'lum hujjat uchun xos so'zlar yuqori vazn oladi.

N-gramm modellari. Ketma-ket kelgan n ta so'zdan iborat birikmalar xususiyat sifatida olinadi. Bu so'zlar tartibi haqidagi ma'lumotni qisman saqlaydi.

Soʻz embeddingʼlari (Word2Vec, GloVe, FastText). Soʻzlar zich vektor fazosida ifodalanadi va maʼnosi yaqin soʻzlar bir-biriga yaqin joylashadi. Bu usul kontekstni yaxshiroq hisobga oladi.

3. Tahlil algoritmlari

3.1. Matnlarni tasniflash

Tasniflash - matnni oldindan belgilangan sinflardan biriga kiritish vazifasi (masalan, "spam" yoki "spam emas", "ijobiy" yoki "salbiy" fikr). Quyidagi algoritmlar keng qoʻllaniladi:

- *Naive Bayes - Bayes teoremasiga asoslangan ehtimoliy algoritm. Sodda, tez va kichik maʼlumotlar toʻplamida ham yaxshi natija beradi.*
- *Support Vector Machine (SVM)- sinflarni ajratuvchi optimal gipertekislikni topadi. Yuqori oʻlchamli matn maʼlumotlarida samarali ishlaydi.*
- *Logistik regressiya - sinfga tegishlilik ehtimolini hisoblaydi, natijalarni sharhlash oson.*
- *Qaror daraxtlari va Random Forest - bir nechta daraxtlar natijasini birlashtirib, aniqlikni oshiradi.*

3.2. Klasterlash

Klasterlash oʻqituvchisiz oʻrganish usuli boʻlib, oldindan belgilangan sinflarsiz oʻxshash matnlarni guruhlaydi. Eng koʻp ishlatiladigan algoritm K-means. U hujjatlarni vektor fazosidagi masofa boʻyicha K ta klasterga ajratadi. Mavzularni aniqlash uchun LDA (Latent Dirichlet Allocation) modeli ham qoʻllaniladi.

3.3. Zamonaviy neyron tarmoq modellari

Soʻnggi yillarda chuqur oʻrganish usullari matn tahlilida yetakchi oʻrinni egalladi. Rekurrent neyron tarmoqlar (RNN, LSTM) matndagi ketma-ketlikni hisobga oladi. Transformer arxitekturasiga asoslangan BERT va GPT turidagi modellar esa soʻzning maʼnosini butun gap kontekstida tushunadi va koʻpchilik vazifalarda eng yuqori aniqlikni koʻrsatadi. Ularning kamchiligi — katta hisoblash resurslari va katta hajmdagi oʻqitish maʼlumotlariga ehtiyoj.

4. Algoritmning soddalashtirilgan namunasi

Quyida TF-IDF va Naive Bayes yordamida matnni tasniflashning Python tilidagi namunasi keltirilgan:

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.pipeline import make_pipeline
matnlar = ["Bu mahsulot juda yaxshi", "Xizmat sifati past",
"Ajoyib natija"]
belgilar = ["ijobiy", "salbiy", "ijobiy"]
```

```
model = make_pipeline(TfidfVectorizer(lowercase=True),  
MultinomialNB())  
model.fit(matnlar, belgilar)  
print(model.predict(["Juda yaxshi xizmat"]))
```

Bu yerda matn avval TF-IDF vektoriga aylantiriladi, so'ngra Naive Bayes modeli yordamida sinf aniqlanadi.

5. O'zbek tili matnlarini tahlil qilishning xususiyatlari

O'zbek tili agglyutinativ til bo'lib, so'zlar o'zakka qo'shimchalar (affikslar) ketma-ket qo'shilishi orqali yasaladi va o'zgaradi. Masalan, "kitob" so'zidan "kitoblarimizdan" shakli hosil bo'ladi. Bu holat tahlilda bir qator qiyinchiliklarni keltirib chiqaradi:

- *Morfologik murakkablik. Bir o'zakning ko'plab shakllari mavjud, shuning uchun stemming va lemmatizatsiya uchun maxsus qoidalar zarur.*
- *Yozuv ko'pligi. O'zbek tilida kirill va lotin alifbolari birga qo'llanadi, shuningdek apostrof belgilari (o', g', tutuq belgisi) turlicha yoziladi. Bu transliteratsiya va belgilarni bir xillashtirish bosqichini talab qiladi.*
- *Resurslarning cheklanganligi. Inglizcha matnlarga nisbatan o'zbek tili uchun tayyor korpuslar, lug'atlar va o'qitilgan modellar kam.*
- *Aralash matnlar. Ijtimoiy tarmoqlarda o'zbekcha, ruscha va inglizcha so'zlar aralash qo'llanadi.*

Shu sababli o'zbek tili uchun dastlabki qayta ishlashda alifbolarni yagona ko'rinishga keltirish, qo'shimchalarni ajratuvchi morfologik tahlilchi va stop-so'zlarning maxsus ro'yxatidan foydalanish tavsiya etiladi.

6. Natijalarni baholash

Tahlil algoritmining sifati quyidagi ko'rsatkichlar yordamida baholanadi:

- *Aniqlik (Accuracy) — to'g'ri tasniflangan hujjatlarning umumiy ulushi.*
- *Precision — ijobiy deb topilgan hujjatlar ichida haqiqatan ijobiylarining ulushi.*
- *Recall — barcha haqiqiy ijobiy hujjatlar ichida to'g'ri topilganlarining ulushi.*
- *F1-ko'rsatkich — precision va recall ning garmonik o'rtachasi:*

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall)$$

Ma'lumotlar sinflar bo'yicha notekis taqsimlangan holatda faqat aniqlikka tayanish noto'g'ri xulosaga olib kelishi mumkin, shuning uchun F1-ko'rsatkichdan foydalanish maqsadga muvofiq.

7. Xulosa

Matnlarni tahlil qilish algoritmi dastlabki qayta ishlash, xususiyatlarni ajratib olish, modelni qo'llash va natijani baholash kabi bosqichlardan iborat yaxlit jarayondir.

Sodda masalalar uchun TF-IDF va Naive Bayes yoki SVM kabi klassik usullar yetarli bo'lsa, murakkab kontekstni tushunishni talab qiladigan vazifalarda transformer asosidagi modellar afzalroq. O'zbek tili uchun morfologik tahlil, alifbolarni bir xillashtirish va katta korpuslar yaratish alohida e'tibor talab qiladigan yo'nalishlardir. Kelgusida o'zbek tiliga moslashtirilgan o'qitilgan modellarni ishlab chiqish bu sohadagi asosiy vazifalardan biri bo'lib qoladi.

Foydalanilgan adabiyotlar

1. Jurafsky D., Martin J. H. Speech and Language Processing. 3rd ed. draft.
2. Manning C. D., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge University Press, 2008.
3. Vaswani A. va boshq. Attention Is All You Need. NeurIPS, 2017.
4. Devlin J. va boshq. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL, 2019.
5. Казанцев Т. Искусственный интеллект и Машинное обучение. Основы программирования на Python / Т. Казанцев - «ЛитРес: Самиздат», 2020.
6. LeCun Y., Bengio Y., Hinton G. Deep learning //nature. – 2015. – Т. 521. – №. 7553. – С. 436-444.
7. Dhamodharan B. Beyond Traditional Methods: A Novel Approach to Anomaly Detection and Classification Using AI Techniques //Transactions on Latest Trends in Artificial Intelligence. – 2022. – Т. 3. – №. 3.