UDK 004.89:811.512.133'322

O‘ZBEK TILIDA SENTIMENT TAHLIL UCHUN MATNNI TAYYORLASH

Abdulla Abdullayev

Urganch innovatsion universiteti Iqtisodiyot va axborot texnologiyalari kafedrasi o‘qituvchisi

E-mail: abdullayev7092@gmail.com

0009-0007-9037-1364

Annotatsiya: Ma’lumotlar sentiment tahlilining "xom ashyosi" hisoblanadi – agar ma’lumotlar sifatsiz, noto‘g‘ri yoki yetarli emas bo‘lsa, natijalar ham xato yoki foydasiz bo‘lib chiqadi. Sentiment tahlilda ma’lumotlarni boshlang‘ich qayta ishlash (preprocessing) jarayoni juda muhim rol o‘ynaydi, chunki bu bosqich ma’lumotlarning sifatini oshiradi va modelning aniqligini yaxshilaydi. Ma’lumotlarni boshlang‘ich qayta ishlash sentiment tahlilining asosiy qismidir, chunki xom matnlar ko‘pincha keraksiz ma’lumotlar, shovqin yoki tuzilmasiz elementlarni o‘z ichiga oladi. Ushbu maqolada matnlarni boshlang‘ich qayta ishlash bosqichlari ko’rib chiqiladi.

Kalit so‘zlar: Sentiment tahlil (Sentiment analysis), Text preprocessing, Lemmalash, Stemmlash, Cleaning, TF-IDF, Bag-of-Words, Word Embeddings, Tokenlash chuqur o‘rganish, nomuhim so‘zlar.

Аннотация: Данные являются "сырьем" для анализа настроения - если данные некачественные, неверные или недостаточные, результаты также будут ошибочными или бесполезными. В сентиментальном анализе процесс предварительной обработки данных играет очень важную роль, поскольку этот этап повышает качество данных и улучшает точность модели. Первичная обработка данных является ключевой частью анализа настроения, поскольку сырые тексты часто содержат ненужную информацию, шум или неструктурированные элементы. Эта статья рассматривает этапы первичной обработки текстов.

Ключевые слова: Sentiment analysis (Sentiment analysis), Text preprocessing, TP, TF-IDF, Bag-of-Words, Word Embeddings, Regular Expressions, Tokenization - углубленное изучение, неважные слова.

Abstract: Data is the "raw material" of sentiment analysis - if the data is poor quality, incorrect, or insufficient, the results will also be erroneous or useless. In sentiment analysis, the initial data processing (preprocessing) process plays a very important role, as this stage improves the quality of the data and the accuracy of the model. Primary data processing is a key part of sentiment analysis, as raw texts often contain unnecessary information, noise, or unstructured elements. This article examines the stages of primary text processing.

ISSN 3030-3907

Keywords: Sentiment analysis, Text preprocessing, TP, TF-IDF, Bag-of-Words, Word Embeddings, Regular Expressions, Tokenization, in-depth learning, unimportant words.

Kirish.

Sentiment tahlilning asosiy maqsadi – matn, audio, video yoki boshqa ma'lumot turlarida ifodalangan **his-tuyg'ular, kayfiyat, munosabat yoki fikrlarni avtomatik aniqlash, tahlil qilish va tasniflash**dan iborat. Sentiment tahlil orqali tashkilotlar, ta'lim muassasalari yoki shaxslar o'z faoliyatini samaraliroq yo'lga qo'yishi mumkin. Ushbu tahlil orqali ma'lumotlarda mavjud bo'lgan emotsional ohang (ijobiy, salbiy, neytral) va undagi nuanslar (masalan, ishonchsizlik, xursandchilik, g'azab) aniqlanadi. Quyidagi jadvalda s**entiment tahlilning** amaliy ahamiyati keltirilagan:

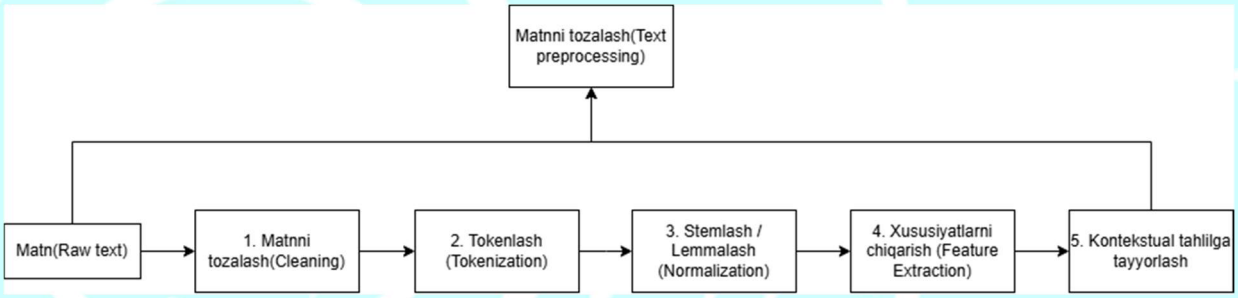
Maqsad	Namuna
Ijtimoiy fikr-mulohazalarni tushunish	Foydalanuvchilar, mijozlar yoki talabalar sharhlaridan shaxsiy fikr va kayfiyatlarni aniqlash.
Qaror qabul qilishni yaxshilash	Tahlil natijalari asosida kompaniya va tashkilotlarga mahsulot, xizmat yoki ta'lim jarayonlarini yaxshilash bo'yicha strategiyalar ishlab chiqishda yordam berish.
Brand imijini monitoring qilish	Ijtimoiy tarmoqlarda brend haqidagi salbiy yoki ijobiy munosabatlarni kuzatish.
Mijozlarga xizmatni optimallashtirish	Mijozlarning shikoyatlarini tezda aniqlab, ularga javob berish.
Raqobatdosh tahlil	Raqobatchilarning zaif va kuchli tomonlarini aniqlash.
Psixologik holatni baholash	Matnlarda stress, tashvish yoki depressiya belgilarini aniqlash (masalan, talabalarning insholari orqali).
Trend va tendentsiyalarni kuzatish	Ijtimoiy tarmoqlardagi, forumlardagi yoki boshqa manbalardagi kayfiyat o'zgarishlarini monitoring qilish orqali bozor va jamiyatdagi o'zgarishlarni aniqlash.

Odatda, sentiment tahlilda hissiyotlarning uch turini ajratib turadi - **salbiy, neytral** va **ijobiy**. U alohida gapga yoki uning qismiga qo'llanilishi mumkin, shuningdek hujjatlar tasnifi uchun ham ishlatilishi mumkin. Bu yerda hujjat sifatida *elektron pochta xabarlari, sharhlar, obzorlar, maqolalar* va boshqalar kabi matnli elementlar qamrab olinadi.

Matnni boshlang'ich qayta ishlash (Text preprocessing).

ISSN 3030-3907

Sentiment tahlilda ma’lumotlarni boshlang‘ich qayta ishlash jarayoni bir necha bosqichlardan iborat bo‘lib, bu bosqichlar tahlilning aniqligi va samaradorligini oshirish uchun muhimdir[1], chunki bu bosqich ma’lumotlarning sifatini oshiradi va modelning aniqligini yaxshilaydi. O‘zbek tilida bu jarayon tilning o‘ziga xos xususiyatlarini hisobga olgan holda amalga oshiriladi.



Ma’lumot tozalash (Cleaning) – Matndan keraksiz belgilar, stop-so‘zlar, maxsus belgilar va takroriy elementlarni olib tashlash.

Tokenlash (Tokenization) – Matnni alohida so‘z yoki iboralarga ajratish.

Stemlash va lemmalash – So‘zlarning asosiy shaklini aniqlash va tahlil uchun mos shaklga keltirish.

Xususiyatlarni chiqarish (Feature Extraction) – TF-IDF, Bag-of-Words, Word Embeddings kabi usullar yordamida matnni tahlil qilishga mos formatga o‘tkazish[2].

Kontekstual tahlilga tayyorgarlik – Sarkazm va ironiyani aniqlash, ma’no o‘zgarishlariga moslashish.

Ma’lumotlarni boshlang‘ich qayta ishlash sentiment tahlilining asosiy qismidir, chunki strukturlanmagan matnlar ko‘pincha keraksiz ma’lumotlar, shovqin yoki keraksiz elementlarni o‘z ichiga oladi. Quyidagi sabablar bu jarayonning muhimligini ko‘rsatadi:

Tozalik: Matnlar ichida sentimentga ta’sir qilmaydigan belgilar (masalan, ortiqcha tinish belgilari, raqamlar, ijtimoiy tarmoq teglari) bo‘lishi mumkin. Bularni olib tashlash modelning e’tiborini asosiy ma’noga qaratadi.

Bir xillik: O‘zbek tilida so‘zlar turli shakllarda yozilishi mumkin (masalan, "yaxshi" va "yahshi"). Normalizatsiya orqali bunday farqlar bartaraf qilinadi.

Tahlil uchun tayyorgarlik: Matnni tokenlash va keraksiz so‘zlarni olib tashlash modelga har bir so‘zning sentimentdagi o‘rnini aniqlashda yordam beradi. O‘zbek tilida sentiment tahlil uchun ma’lumotlarni qayta ishlash bir nechta muhim bosqichlardan iborat. Quyida har bir bosqichni misollar bilan keltirilgan:

Matnni tozalash (Text Cleaning) – bu sentiment tahlil jarayonida asosiy bosqich bo‘lib, matndan keraksiz, chalg‘ituvchi yoki tahlilga salbiy ta’sir ko‘rsatadigan qismlarni olib tashlash va uni toza, strukturaviy holga keltirishni nazarda tutadi. Bunda, ma’lumot manbasi va uning tarkibini tahlil qilib, mavjud xatoliklar, HTML teglari, maxsus belgilar va boshqa keraksiz elementlarni aniqlash

ISSN 3030-3907

lozim. Veb-saytlardan olingan matnlarda HTML teglari, JavaScript kodi yoki reklama bloklari bo‘lishi mumkin. Ularni olib tashlash orqali faqat asosiy matn qoladi[3].

Tokenlash[4] – matnni so‘z yoki so‘z birikmalariga ajratish jarayoni bo‘lib, bu modelga har bir so‘zning sentimentdagi rolini aniqlashga yordam beradi. Tokenlash – bu sentiment tahlil jarayonida matnni kichik, analizga yaroqli birliklarga ajratish bosqichidir. Tokenlash orqali matnni kompyuter tushunishi va tahlil qilishi mumkin bo‘lgan shaklga keltiriladi. Ushbu jarayon natijasida matn so‘zlar, iboralar, raqamlar va ba’zan hatto punktuatsion belgilar darajasiga bo‘linadi. Tokenlash jarayonida matn murakkabligini kamaytirilib, uni tahlil qilishga mos va soddalashtirilgan shaklga keltiriladi. Har bir token matndagi ma’no va hissiyotning aniqlanishida muhim rol o‘ynaydi.

Ba’zan tokenlash jarayonida murakkab iboralar yoki emojilar, hashtaglar kabi maxsus elementlarni ajratish talab etiladi. Regulyar ifodalar (Regular Expressions) vositasida matndagi maxsus belgilar va shablonlar asosida tokenlarni aniqlashda qo‘llaniladi.

Har bir tilning o‘ziga xos grammatikasi va ifodalash uslubi bo‘ladi. Masalan, o‘zbek tilida qo‘shimchalar va so‘zlarning turli shakllari tokenlash jarayonida inobatga olinishi kerak. Shuning uchun tilga mos maxsus algoritmlar yoki qoida asosida tokenizatsiya usullari tanlanadi. Har bir token sentiment tahlilda alohida ma’no va his-tuyg‘ularni ifodalaydi. To‘g‘ri amalga oshirilgan tokenlash modelga matnning semantik strukturasini aniqroq o‘rganish imkoniyatini beradi. Tokenlar asosida so‘z chastotasi, kontekstual bog‘lanishlar va boshqa xususiyatlar ajratilib, modelga uzatiladi. Bu sentiment tahlil natijasining sifatini oshiradi. Tokenlar yordamida, keyingi bosqichlarda xususiyatlar (features) ajratiladi, masalan, Bag-of-Words, TF-IDF yoki word embeddings shakllantiriladi va modelning ishlash samaradorligi sezilarli darajada yaxshilanadi.

Kichik harflarga o‘tkazish (Lowercasing)[5] – bu sentiment tahlilida matnni normallashtirishning oddiy, ammo juda muhim bosqichlaridan biridir. Ushbu jarayon matndagi barcha harflarni kichik shaklga o‘tkazishni nazarda tutadi va tahlil jarayonining samaradorligini oshiradi. Matnda "Yaxshi", "yaxshi" yoki "YAXSHI" kabi turlicha yozilgan so‘zlar bitta standart shaklga keltiriladi. Bu, model so‘zlarning takrorlanishini kamaytirish va yagona lug‘at (vocabulary) yaratishda yordam beradi. Katta va kichik harfli variantlar alohida token sifatida qaralmasligi uchun so‘zlarning umumiy soni (vocabulary size) qisqaradi, bu modelning tezligi va aniqligini oshiradi.

Nomuhim so‘zlarni olib tashlash (Stop-words Removal) – bu sentiment tahlilda matndan analizga sezilarli ma’no bermaydigan, kundalik tilning funksional qismlarini (masalan, bog‘lovchilar, yordamchi fe’llar, ko‘p ishlatiladigan qisqa

ISSN 3030-3907

so‘zlar) olib tashlash jarayonidir. Ushbu bosqichning asosiy maqsadi – tahlil jarayonida chalg‘ituvchi va ortiqcha ma’lumotlarni kamaytirish, shuningdek, muhim so‘zlarga e’tibor qaratishdir.

Nomuhim so‘zlar – matnda tez-tez uchraydigan, lekin o‘zida katta semantik yukga ega bo‘lmagan so‘zlardir. Masalan, “va”, “bu”, “lekin”, “bir”, “uchun” kabi so‘zlar ko‘plab til modellari va tahlil usullarida nomuhim so‘zlar ro‘yxatiga kiritiladi. Ushbu so‘zlar matnning grammatik tuzilishini ta’minlashga yordam beradi, ammo sentiment yoki asosiy ma’lumotni ifodalashda unchalik muhim emas.

Nomuhim so‘zlar ko‘pincha semantik jihatdan nisbatan kam ma’lumot beradi. Ularni olib tashlash, asosiy va muhim so‘zlarning ahamiyatini oshiradi. Matnni vektorlashtirish jarayonida nomuhim so‘zlar mavjud bo‘lsa, lug‘at (vocabulary) o‘lchami keraksiz ravishda oshib ketadi. Ularni olib tashlash esa hisoblash resurslarini tejaydi va modelning samaradorligini oshiradi. Sentiment tahlil uchun faqat muhim so‘zlar qolganda, keyingi bosqichlarda (masalan, Bag-of-Words, TF-IDF yoki word embeddings) modelga uzatiladigan xususiyatlar aniqroq ifodalanadi.

Lemmalash va stemlash – bu sentiment tahlilida matnni normallashtirishning muhim bosqichlaridan bo‘lib, ular so‘zlarni ularning asosiy shakliga (asos, o‘zak) keltirish orqali tahlil jarayonini soddalashtiradi.

Stemlash – bu so‘zlarning ildizini (stem) ajratib olish jarayoni. Ushbu jarayonda so‘zlardan qo‘shimchalar va morfemalar olib tashlanadi. Natijada hosil bo‘lgan ildizlar ko‘pincha haqiqiy lug‘aviy so‘zlar bo‘lmaydi, ammo ular ma’lumotlarni qisman normallashtirishga yordam beradi. Stemlash algoritmlari so‘z oxiridagi qo‘shimchalarni olib tashlash uchun oddiy qoidalar asosida ishlaydi. Stemlash jarayonida ba’zan so‘zning semantik ma’nosini yetarlicha aniqlay olmaydi. Hosil bo‘lgan stemlar haqiqiy so‘z bo‘lmasligi mumkin, bu esa tahlil natijalarida aniqlikni pasaytirishi mumkin.

Lemmalash – bu so‘zlarni ularning lug‘aviy asosiy shakliga (lemma) keltirish jarayoni. Bu usul so‘zning grammatik va semantik xususiyatlarini hisobga olib, uni lug‘atda mavjud bo‘lgan to‘g‘ri shaklga keltiradi. Masalan, “yugurdi”, “yugurayotgan” – bularning barchasi “yugur” shakliga keltiriladi. Lemmalash so‘zning lug‘aviy ma’nosi va grammatik tuzilishini inobatga oladi. Shuning uchun, bu jarayonda so‘z to‘g‘ri lemma shakliga keltiriladi. Lemmalash jarayonida so‘zlarning to‘g‘ri va kontekstual shakli aniqlanadi. Bu jarayonda so‘zlarning to‘g‘ri va kontekstual shakli aniqlanadi. Lemmalash natijasida turli shakldagi so‘zlar alohida token sifatida qaralmasligi uchun umumiy so‘zlar to‘plami (vocabulary) qisqaradi. Sentiment tahlilda asosiy ma’no muhim, qo‘shimchalar esa ko‘pincha sentimentga ta’sir qilmaydi.

ISSN 3030-3907

Har ikkala usul ham so‘zlarning turlicha shakllarini yagona formatga keltiradi. Masalan, “yaxshi”, “yaxshiroq”, “eng yaxshi” kabi shakllarni bitta asosiy shaklga – “yaxshi” ga keltirish, tahlil modeliga ma’lumotlarni aniqroq uzatishga yordam beradi. Matndan xususiyatlarni chiqarishda, normallashtirilgan so‘zlar yordamida, Bag-of-Words, TF-IDF yoki word embeddings kabi usullar samaraliroq ishlaydi. To‘g‘ri normallashtirish, ya’ni so‘zlarning lug‘aviy shaklini aniqlash, sentiment tahlil modelining kontekstual ma’noni tushunishini yaxshilaydi va natijalar aniqligini oshiradi.

Normalizatsiya jarayoni – bu sentiment tahlilida matnni bir xil va standart formatga keltirishni nazarda tutadi. Ushbu jarayon matndagi turli xil yozilish uslublari, belgilar, raqamlar va maxsus elementlarni yagona ko‘rinishga keltirish orqali keyingi bosqichlarda (masalan, tokenlash, xususiyatlarni ajratish, model o‘rgatish) yanada samarali tahlil qilish imkonini beradi.

Normalizatsiya – matndagi noqulayliklarni (masalan, imlo farqlari, qisqartmalar, sheva variantlari) yo‘qotib, uni bir xil shaklga keltirish jarayoni. Sentiment tahlilda bu jarayon so‘zlarning bir xil ma’noni bildiruvchi turli shakllarini birlashtirish uchun ishlatiladi, shunda model chalkashmaydi.

Normalizatsiya jarayoni quyidagi vazifalarni o‘z ichiga oladi:

So‘zlarni birlashtirish: Turli shakldagi so‘zlarni bir xil shaklga keltirish.

ketdim, ketadi, ketayotgan → **ketmoq**

Bu jarayonda so’zlar lug’aviy asosiy shakilga keltiriladi.

Matnni soddalashtirish: Matndagi keraksiz, takroriy yoki shovqinli qismlarni olib tashlash. Masalan:

– ortiqcha belgi va simvollarni olib tashlash
– reklama matnlarini chiqarib tashlash
– HTML teglarni olib tashlash
– takrorlangan bo‘laklarni tozalash
– emoji, raqamlar yoki stop-so‘zlarni olib tashlash (zarur bo‘lsa)

Ma’lumotlarni standartlashtirish: Matnni tahlil qilish uchun mos shaklga keltirish.

O‘zbek tilida bir xil so‘z turli uslubda yoki shevada yozilishi mumkin ("qo‘shiq" va "koshiq"). Adabiy til bilan shevalar o‘rtasida katta farq bor. Shevadagi bazi so’zlar adabiy tilda umuman boshqacha shaklda qo’llanadi. Masalan Xorazm shevasida mavjud bo‘lgan “nichiksan?” so’zi aslida adabiy tilda mavjud va “sog’lig’ing yaxshimi?” kabi qo’llanadi. Bu kabi holatlarni ko’plab uchratishimiz mumkin. Biz matnni manbasi sifatida qaysi platformani qarasak shunga qarab bu kabi holatlar ham o’zgarib boradi. Masalan ijtimoiy tarmoqdagi izohlarni qarasak bunday holatlarga juda ko’p duch kelishimiz mumkin. Forumlar yoki boshqa ilmiy sohaga oid web platformalarda ko’proq adabiy izohlarni

ISSN 3030-3907

kuzatishimiz mumkin. Normalizatsiya bu farqlarni yo‘qotadi. Normalizatsiya natijasida model "yaxshi" va "yahshi"ni bir xil so‘z sifatida taniydi, bu sentiment tasnifini yaxshilaydi.

Xulosa. Matn turli manbalardan yig‘ilgan bo‘lsa, unda yozilish uslublari, katta-kichik harflar, maxsus belgilar va raqamlar turlicha bo‘lishi mumkin. Bir xil formatga keltirilgan matnni keyingi bosqichlarda aniqroq analiz qilish mumkin, bu esa modelning samaradorligini va aniqligini oshiradi. Keraksiz yoki chalg‘ituvchi belgilar, raqamlar, maxsus ifodalar va qisqartmalarni bartaraf etish orqali asosiy semantik ma’lumotga urg‘u beriladi.

ADABIYOTLAR RO‘YXATI

1. Pavlopoulos, Ioannis. Aspect based sentiment analysis. Diss. Athens University Economics and Business, Greece, 2014.
2. Do, Hai Ha, et al. "Deep learning for aspect-based sentiment analysis: a comparative review." Expert systems with applications 118 (2019): 272-299.
3. Javdan, Soroush, and Behrouz Minaei-Bidgoli. "Applying transformers and aspect-based sentiment analysis approaches on sarcasm detection." Proceedings of the second workshop on figurative language processing. 2020.
4. Guzman, Emitza, and Walid Maalej. "How do users like this feature? a fine grained sentiment analysis of app reviews." 2014 IEEE 22nd international requirements engineering conference (RE). Ieee, 2014.
5. Kumar, Akshi, et al. "Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data." Information Processing & Management 57.1 (2020): 102141.

ISSN 3030-3907