RESTORAN SOHASIDAGI O‘ZBEK TILIDAGI MATNLARNING SENTIMENT TAHLILI

Niyazmetova Kumushoy Ergash qizi
UrDU magistranti,
kumushoy.n@urdu.uz
Quriyozov Elmurod Rajabboy o‘g‘li
UrDU o‘qituvchisi,
elmurod1202@urdu.uz

Annotatsiya. Foydalanuvchilar tomonidan yozilgan ko‘plab sharhlarni sentimental tahlil qilish va tasniflash muammolarini yechich uchun foydali ma’lumotlarni olish tabiiy tilni qayta ishlashning muhim vazifasidir. Bu nafaqat mijozlar ehtiyojini qondirish uchun, balki kompaniyaning keyingi rivojlanishiga ham ta’sir qilishi mumkin. Ushbu maqolada biz Google xaritasidagi Toshkent shahrida joylashgan restoranlarga bildirilgan fikr-mulohazalaridan foydalanib Dataset tayyorlandi va logistic regressiyaga asoslangan modellaridan foydalanilib Sentiment tahlil qilindi. Umumiy baholash natijalari shuni ko‘rsatadiki, agglyutinativ tillar uchun stemming kabi dastlabki ishlov berish bosqichlarini bajarish orqali tizim yaxshi natijalar beradi va natijada eng yaxshi ishlaydigan modelda 91% aniqlikka erishiladi.

Kalit so‘zlar: *NLP, Sentiment tahlil, Dataset, Bag of word modeli, TF-IDF algoritmi.*

Abstract. Sentiment analysis of many comments written by users and extracting useful information for solving classification problems is an important task of natural language processing (NLP). This can affect not only customer satisfaction, but also the further development of the company. In this article, we have prepared a Dataset using feedback given to restaurants located in the city of Tashkent on the Google map and analyzed Sentiment using Logistic Regression models. Overall evaluation results show that the system performs well by performing pre-processing steps such as stemming for agglutinative languages, resulting in 91% accuracy in the best performing model.

Keywords: *NLP, Sentiment Analysis, Dataset, Bag of word model, TF-IDF algoritm.*

Аннотация. Анализ тональности многих комментариев, написанных пользователями, и извлечение полезной информации для решения задач классификации является важной задачей обработки естественного языка (NLP). Это может повлиять не только на удовлетворенность клиентов, но и на дальнейшее развитие компании. В этой статье мы подготовили набор данных, используя отзывы, полученные от ресторанов, расположенных в городе Ташкент, на карте Google, и проанализировали настроения с использованием

моделей логистической регрессии. Общие результаты оценки показывают, что система хорошо работает, выполняя этапы предварительной обработки, такие как выделение корней для агглютинативных языков, что обеспечивает точность 91% в самой эффективной модели.

Ключевые слова: *NLP, анализ тональности, набор данных, модель «мешок слов», алгоритм TF-IDF.*

Kirish.

Tabiiy tilni qayta ishlash (NLP) usullarining samaradorligi ko‘plab ilovalarda belgilangan ma’lumotlar miqdoriga bog‘liq. Sentiment tahlil qilish - bu iste’molchilar tomonidan bildirilgan fikrlarni tahlil qilish. Iste’molchilar odatda joylar/ovqatlar haqidagi fikr-mulohazalarini Google Map, Yelp va boshqalar kabi mashhur ilovalarga joylashtiradilar. Ular ko‘pincha iste’molchilarni sharhlarda faol ishtirok etishga undaydi va foydalanuvchi tomonidan yaratilgan fikr-mulohazalar asosida iste’molchilarga o‘z ehtiyojlarini to‘liq qondirish imkonini beradi [1]. Tadbirkorlarga esa real vaqt rejimida shaxsiylashtirilgan xizmatlarni taqdim etishda yordam beradi.

Bundan tashqari, restoran sharhlari mijozlarning hissiy ehtiyojlari tarkibini ifodalaydi va iste’molchilar tanlovi haqida muhim ma’lumot manbayi hisoblanadi. Hozirgi vaqtda fikr-mulohaza yuritish, ayniqsa, yuqori manba tillari uchun chuqur o‘rganish usullarini qo‘llaganidan keyin juda yuqori aniqlik ko‘rsatkichlariga erishdi [2]. O‘zbek tili agglyutinativ til bo‘lib, unda bir so‘z ma’noli gap bo‘la oladi. Bizning ma’lumotlarimizga ko‘ra, restoran domenidagi fikr-mulohazalarga asoslangan hissiyotlarni tasniflash muammolari bo‘yicha oldingi ishlar yetarli emas [3]. Shunday qilib, ushbu maqola uchun quyidagilarni hisobga olinadi:

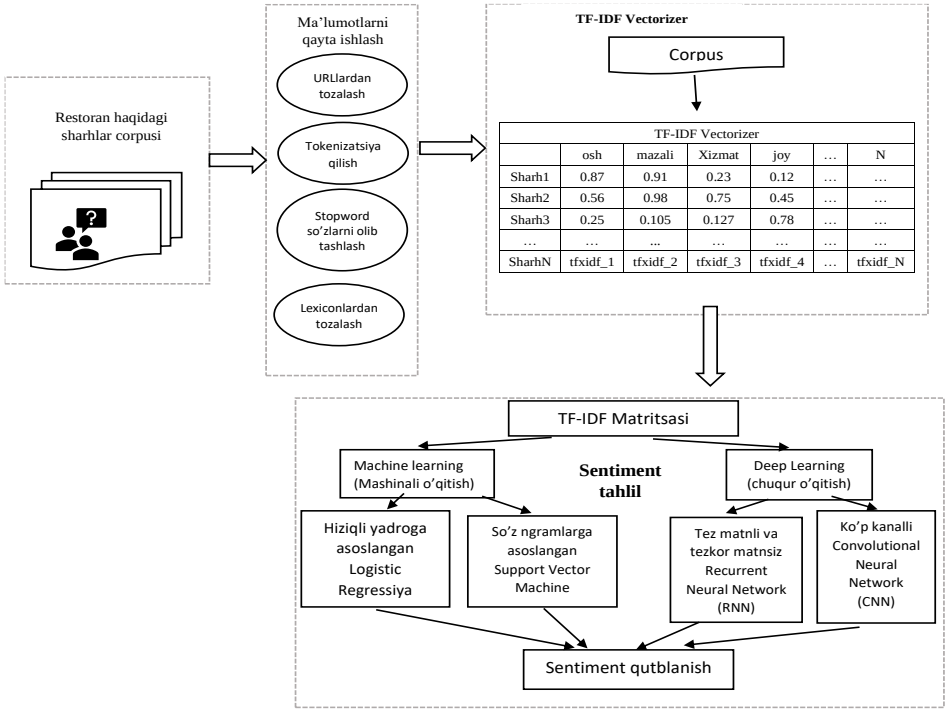
• Restoran domenining izohli korpusi his-tuyg‘ularni tahlil qilish uchun yaratilgan bo‘lib, u Google Xaritalardan o‘zbek oshxonasi joylashuvi asosida to‘planadi, bu yerda mahalliy milliy taomlar sharhlari asosiy maqsad hisoblanadi. Dataset katta xatolarni qo‘lda olib tashlash va tozalashdan keyin 4500 ijobiy va 3710 salbiy sharhlarni o‘z ichiga oladi. Annotatsiya jarayoni Google Xaritalar tomonidan ajratilgan 5 yulduzli fikr-mulohaza usuliga asoslanadi, bunda 1 dan 3 gacha ma’lumotlar to‘plamini salbiy va 4 dan 5 gacha ijobiy deb hisoblaymiz. Biz ba’zi sharhlar ingliz, qirg‘iz va rus kabi boshqa tillarga asoslanganligini aniqladik. Biz ularni e’tiborsiz qoldirishni istamadik, shuning uchun rasmiy Google Translate API yordamida ularni o‘zbek tiliga tarjima qilishga qaror qildik.

• Datasetni oldindan qayta ishlash ikki bosqichda qo‘llaniladi. Birinchi qadamlar URL manzillarini, tinish belgilarini va kichik harflarni olib tashlashdir. Ikkinchi qadam TF-IDF algoritmidan foydalangan holda to‘xtash so‘zlar ro‘yxatini yaratgandan so‘ng aniqlikni baholashga asoslangan ma’lumotlar to‘plamidagi to‘xtatuvchi so‘zlarni e’tiborsiz qoldirish [4]; Keyin, biz stemming algoritmini qo‘lladik.

O‘zbek tilidagi so‘zlarni elektron lug‘ati asosida o‘zbek tili nutqining bo‘lagi: ot, sifat, son, fe’l, kesim, mayl, undoshlar uchun xulosa chiqarish kombinatsion yondashuvdan foydalaniladi. Algoritmni qo‘llashning afzalliklari leksikadan xoli va uning murakkabligi bitta operatsiyani (tilning oxirlari lug‘atiga ishora) bajarishga imkon beradi:

- so‘zni qo‘shimchalarga bo‘lish;
- so‘zlarni morfologik tahlil qilish.

So‘nggi yillarda o‘zbek tili uchun NLP sohasida bir qancha ishlar amalga oshirildi, jumladan Google Play ilovasi sharhlarini to‘plash va tahlil qilish natijasida yaratilgan hissiyotlarni tahlil qilish ma’lumotlar to‘plami, ikki turdagi ma’lumotlar: o‘rta o‘lchamli qo‘lda izohlangan ma’lumotlar to‘plami va katta hajmdagi ma’lumotlar to‘plami ingliz tilidan avtomatik ravishda tarjima qilinadi [5]. Yana bir shunga o‘xshash maqola o‘zbek matnlarini fikr tasniflashda yangilik matninig kategoriyasi asosidagi xususiyatlarning ta’sirini o‘rganilgan [6]. Semantik baholash bo‘yicha ma’lumotlar to‘plami so‘z juftliklarida semantik o‘xshashlik va bog‘liqlik ballari, shuningdek, uning o‘zbek tili uchun tahlili so‘nggi ishda taqdim etilgan. NLP-da sun’iy intellektga asoslangan usullardan foydalanadigan o‘sib borayotgan tendensiya mavjud, buni o‘zbek tilidagi neyron transformatorlari - arxitekturaga asoslangan til modeli bilan o‘zbek tilidagi ishda ko‘rish mumkin [7].



1-rasm. Tadqiqot doirasi

Tuyg‘ularni tahlil qilish sohasiga global nuqtayi nazardan qarashlar va fikrlardagi farqlarni hisobga olish g‘oyasi bilan o‘z ishlarida hissiyotlarni tahlil qilishning turli usullaridan, masalan, mashinani o‘rganish va chuqur o‘rganish usullaridan foydalangan ishlar mavjud. Twitter, Reddit, Tumblr va Facebook kabi mashhur ijtimoiy platformalarda mavjud bo‘lgan fikrlardan iborat.

Metodologiya.

Ushbu maqolada biz restoran domeni ma’lumotlar to‘plami uchun mashinali o‘rganish va chuqur o‘rganishga asoslangan hissiyotlarni tahlil qilish tizimini taklif qildik (1-rasm) [8]. Ramka veb-brauzer yordamida ma’lumotlarni yig‘ish, oldindan ishlov berish (tozalash, to‘xtatuvchi so‘zlar, leksikonsiz stemming) [3], TF-IDF vazn matritsasi yaratish [9], [10], hissiyotlarni tahlil qilish uchun ML va DLni amalga oshirishni o‘z ichiga oladi [11].

2-rasm: Google platformasida restoran uchun fikr qoldirish interfeysi namunasi

Ma’lumotlarni yig‘ish.

Biz o‘zbek tilida skanerlash uchun mavjud bo‘lgan ko‘p sonli ma’lumotlar to‘plamini ko‘rib chiqishdan boshlaymiz. Biroq Twitter yoki kino sharhlari kabi odatiy yondashuvlar o‘zbek tiliga to‘g‘ri kelmaydi. Shu sababli, biz restoran sharhlarini to‘plashga qaror qildik, chunki mahalliy aholi asosan restoranlar haqida fikr-mulohaza bildirishni yaxshi ko‘rishadi. Bizning fikrimizcha, bu mantiqan to‘g‘ri, chunki o‘zbek taomlari Mustaqil Davlatlar Hamdo‘stligi (MDH, CA mamlakatlari) bo‘ylab eng mashhur taomlardan biri hisoblanadi. Masalan, Kaliforniyaning aksariyat shaharlarida o‘zbek oshxonasiga ixtisoslashgan band bo‘lgan restoranlarni topish oson. Biz Toshkentdagi barcha mahalliy restoranlarni Google Xaritalardan ko‘rib chiqdik. Birinchidan, biz kamida 3 ta ko‘rib chiqishga ega bo‘lgan 140 dan ortiq URL manzillar ro‘yxatini tanladik va biz 2-rasmda ko‘rsatilgan barcha ma’lumotlarni oldik. Ko‘rish chog‘ida Googlening spamga qarshi va DDOSga qarshi siyosatlarini ko‘rib chiqdik, chunki ma’lumotlarni yig‘ishda ma’lum cheklovlar mavjud.

Ma’lumotlarni oldindan qayta ishlash.

Ma’lumotlar to‘plamini tekshirishda yulduzcha bilan baholangan matnlar to‘plami qo‘lda tuzatish talab qilingan. Faqat kulgichlar, ismlar yoki foydalanuvchi nomi eslatmalari, URL manzillari yoki maxsus ilova nomlari kabi boshqa ahamiyatsiz kontentdan iborat sharhlar olib tashlandi. O‘zbek tilidan boshqa tillarda (asosan rus va ba’zilari ingliz tilida) yozilganlar rasmiy Google translate API

yordamida tarjima qilingan. O‘zbekistonda odamlar rasmiy lotin alifbosidan foydalanishsa-da, eski kirill alifbosidan foydalanish, ayniqsa, kattalar orasida bir xil darajada mashhur [12]. O‘sha izohlar kirill alifbosida yozilganlar o‘zbek mashinasi transliteratsiyasi vositasi yordamida lotinga o‘tkazildi. Keyin, muhim ma’lumotlarga e’tibor qaratish uchun sharhlarimizdan past darajadagi ma’lumot so‘zlarini olib tashlash uchun to‘xtash so‘zlarini qo‘lladik. Modelimiz TF-IDF (Term chastotasi - teskari hujjat chastotasi) yordamida bitta so‘zni to‘xtatuvchi so‘zlar to‘plamini avtomatik aniqlashning tavsiya etilgan algoritmidir. Shundan so‘ng, har bir so‘z prefiks va qo‘shimchalar tufayli so‘z hajmini kamaytirish algoritmi leksikadan xoli bo‘lgan asos yaratish vositasiga qayta ishlanadi. Asosiy g‘oya - mos keladigan so‘zlarning kombinatsion yondashuvidan foydalanish.

Baholash. To‘plangan yangi ma’lumotlar to‘plami mos ravishda 8 x 2 nisbatda baholash uchun *train-set* va *test-set* to‘plamlariga bo‘linadi. Trainset modelni qurish va o‘qitish uchun, testset esa qurilgan modelni testlab ko‘rish uchun kerak. Ma’lumotni tozalash jarayonidan so‘ng bizda asl ma’lumotlar to‘plami quyidagicha bo‘ladi:

bu yerda $\overrightarrow{x_i}$ xususiyat vektorlarini va $\overrightarrow{y_i}$ izohli teglarni ifodalaydi:

$$(\overrightarrow{x_i}, \text{yi}), \qquad i = 1, 2, 3, ..., N \ (1)$$

$$\overrightarrow{x_i} = (\text{xi1}, \text{xi2}, \ldots, \text{xim}) \qquad i = 1, 2, 3, ..., N \ (2)$$

N va m mos ravishda ko‘rib chiqishlar soni va xususiyat vektorining uzunligiga teng. Keyin biz har bir xususiyat vektori $\overrightarrow{x_i}$ uchun TF-IDF ballarini hisoblaymiz, bu esa so‘zlarni olish orqali vektorlashtiradi. Berilgan sharhdagi so‘zning chastotasini va sharhlar orasidagi chastotani hisobga olish.

Barcha $\overrightarrow{z_i}$ larning yakuniy natijasi siyrak matritsa sifatida aniqlanadi.

$$\overrightarrow{z_i} = \text{TF(xi)xIDF(xi)} \qquad i = 1, 2, 3, ..., N \ (3)$$

Mashinani o‘rganish algoritmlari. Logistik regressiya modeli

$$\text{h}(\vec{z}) = 1/(1 + \exp(-z)$$

$$\text{P}(\text{y} \mid \vec{z}) = \begin{cases} h(\vec{z}), & \text{if y} = +1\text{(ijobiy)} \\ 1 - h(\vec{z}), & \text{if y} = -1\text{(salbiy)} \end{cases} \quad (4)$$

Logistik regressiya modeli – bu tasniflash algoritmi bo‘lib, o‘zining eksponentsial va log chiziqli funksiyalari bilan mashhur. U dikret qiymatlar bilan ishlaydi va har qanday real qiymat funksiyasi 0 va 1 ni ko‘rsatadi. Hissiyotni tahlil qilish shuni ko‘rsatadiki, sharhlar (4) formula yordamida ijobiy yoki salbiy bo‘ladi.

Natijalar va muhokamalar.

Ushbu bo‘limda to‘plangan hissiyotlarni tahlil qilish ma’lumotlar to‘plamiga qo‘llaniladigan mashinani o‘rganish va chuqur o‘rganish usullaridan foydalangan holda baholash jarayonida olingan natijalarning batafsil tavsifi keltirilgan.

Tajriba natijalari

Madel nomlari	Sentiment	Accuracy %	Recall %	F1 %	Precision %
Logistik regressiya (so‘zlarga asoslangan n-gramm)	Ijobiy	88	98	93	89
	Salbiy	88	67	74	
Logistik regressiya (belgilarga asoslangan n-gramm)	Ijobiy	87	51	92	87
	Salbiy	83	97	64	
Logistik regressiya (so‘z va belgilarga asoslangan n-gramm)	Ijobiy	95	95	92	91
	Salbiy	90	89	90	

1-jadval

Yuqorida qayd etilgan baholashning umumiy eksperiment natijalari amalga oshirildi va natijalarni 1-jadvalda ko‘rish mumkin.

Xulosa. Ushbu maqolada biz o‘zbek tili uchun restoran domenidagi yangi ma’lumotlar to‘plamini ko‘rsatdik, 8210 ta sharh, ijobiy yoki salbiy yorliqlar bilan izohlangan, u Google Xaritalardan poytaxt Toshkentdagi barcha restoran profillarining URL manzillaridan foydalangan holda ko‘rib chiqilgan va ularga mos yulduz reytingi sifatida belgilangan. Keyin biz asosiy modellarimiz aniqligini oshirish uchun ma’lumotlar to‘plamiga oldindan ishlov berishning to‘liq bosqichlarini qo‘lladik. Ma’lumotlar to‘plamidagi eng yaxshi aniqlik natijasi (91%) so‘z va belgilar n-grammlari bilan logistik regressiya modeli yordamida olingan. Yaqin kelajakda biz restoran sharhlarini amaliy darajada samarali tahlil qila oladigan ko‘proq ma’lumotlarni to‘plash orqali ishni kengaytirishni rejalashtirmoqdamiz. Shuningdek, ma’lumotlarni bo‘lishda o‘zaro tekshirish usullarini qo‘llash orqali o‘quv eksperimentlarini baholashning noto‘g‘riligini bartaraf etish bo‘yicha ishlar olib borilmoqda.

Foydalanilgan adabiyotlar:

1. B. Pang, L. Lee, and others, “Opinion mining and sentiment analysis,” *Foundations and Trends® in information retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.
2. L. Zhang, S. Wang, and B. Liu, “Deep learning for sentiment analysis: A survey,” *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 8, no. 4, p. e1253, 2018.
3. U. Salaev, E. Kuriyozov, and C. Gómez-Rodríguez, “SimRelUz: Similarity and Relatedness scores as a Semantic Evaluation Dataset for Uzbek Language,” in *1st Annual Meeting of the ELRA/ISCA Special Interest Group on*

Under-Resourced Languages, SIGUL 2022 - held in conjunction with the International Conference on Language Resources and Evaluation, LREC 2022 - Proceedings, 2022, pp. 199–206. [Online]. Available: www.scopus.com

4. X. Madatov, M. Sharipov, and S. Bekchanov, “O`ZBEK TILI MATNLARIDAGI NOMUHIM SO`ZLAR,” *COMPUTER LINGUISTICS: PROBLEMS, SOLUTIONS, PROSPECTS*, vol. 1, no. 1, 2021.

5. E. Kuriyozov, S. Matlatipov, M. A. Alonso, and C. Gómez-Rodríguez, “Construction and evaluation of sentiment datasets for low-resource languages: The case of Uzbek,” in *Human Language Technology. Challenges for Computer Science and Linguistics: 9th Language and Technology Conference, LTC 2019, Poznan, Poland, May 17–19, 2019, Revised Selected Papers*, 2022, pp. 232–243.

6. E. Kuriyozov, U. Salaev, S. Matlatipov, and G. Matlatipov, “Text classification dataset and analysis for Uzbek language,” *arXiv preprint arXiv:2302.14494*, 2023.

7. J. Mattiev, U. Salaev, and B. Kavsek, “Word Game Modeling Using Character-Level N-Gram and Statistics,” *Mathematics*, vol. 11, no. 6, p. 1380, 2023.

8. S. Matlatipov, H. Rahimboeva, J. Rajabov, and E. Kuriyozov, “Uzbek Sentiment Analysis Based on Local Restaurant Reviews,” in *CEUR Workshop Proceedings*, 2022, pp. 126–136. [Online]. Available: www.scopus.com

9. K. Madatov, S. Bekchanov, and J. Vičič, “Uzbek text summarization based on TF-IDF,” *arXiv preprint arXiv:2303.00461*, 2023.

10. K. Madatov, S. Matlatipov, and M. Aripov, “Uzbek text’s correspondence with the educational potential of pupils: a case study of the School corpus,” *arXiv preprint arXiv:2303.00465*, 2023.

11. M. Sharipov and U. Salaev, “Uzbek affix finite state machine for stemming,” *arXiv preprint arXiv:2205.10078*, 2022.

12. M. Sharipov, E. Kuriyozov, O. Yuldashev, and O. Sobirov, “UzbekTagger: The rule-based POS tagger for Uzbek language,” *arXiv preprint arXiv:2301.12711*, 2023.