SENTIMENT ANALYSIS OF COMMENTS IN UZBEK LANGUAGE

Niyazmetova Kumushoy Ergash kizi
Teacher of the Urganch branch of TATU named after Muhammad al-Khwarizmi

https://doi.org/10.5281/zenodo.10654855

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 algorithm.*

Introduction.

The effectiveness of natural language processing (NLP) techniques in many applications depends on the amount of data specified. Sentiment analysis is the analysis of opinions expressed by consumers. Consumers usually post their reviews of places/food on popular apps like Google Map, Yelp, etc. They often encourage consumers to actively participate in reviews and allow consumers to fully satisfy their needs based on user-generated feedback [1]. And it helps entrepreneurs to provide personalized services in real time.

In addition, restaurant reviews represent the composition of customers' emotional needs and are an important source of information about consumer choice. Currently, feedback has achieved very high accuracy rates after applying deep learning techniques, especially for high source languages [2]. The Uzbek language is an agglutinative language, in which one word can be a meaningful sentence. To our knowledge, there is insufficient previous work on feedback-based sentiment classification problems in the restaurant domain [3]. So, for this article, the following are taken into account:

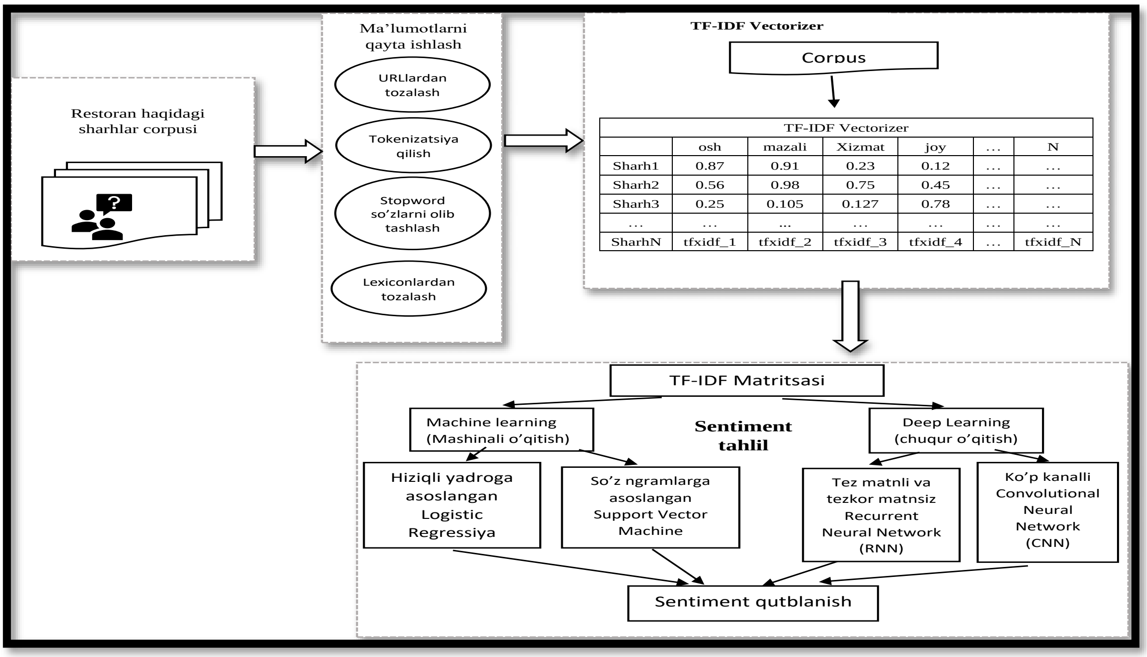
An annotated corpus of the restaurant domain is created for sentiment analysis, which is collected based on the location of Uzbek cuisine from Google Maps, where local national food reviews are the main target. The dataset contains 4500 positive and 3710 negative reviews after manual removal and cleaning of large errors. The annotation process is based on a 5-star feedback method allocated by Google Maps, where we consider 1 to 3 datasets as negative and 4 to 5 as positive. We found that some reviews are based on other languages such as English, Kyrgyz and Russian. We didn't want to ignore them, so we decided to translate them into Uzbek using the official Google Translate API.

An annotated corpus of the restaurant domain is created for sentiment analysis, which is collected based on the location of Uzbek cuisine from Google Maps, where local national food reviews are the main target. The dataset contains 4500 positive and 3710 negative reviews after manual removal and cleaning of large errors. The annotation process is based on a 5-star feedback method allocated by Google Maps, where we consider 1 to 3 datasets as negative and 4 to 5 as positive. We found that some reviews are based on other languages such as English, Kyrgyz and

Russian. We didn't want to ignore them, so we decided to translate them into Uzbek using the official Google Translate API.

- dividing the word into adverbs;
- morphological analysis of words.

In recent years, several works have been done in the field of NLP for the Uzbek language, including a sentiment analysis dataset created by collecting and analyzing Google Play app reviews, two types of data: medium o size manually annotated datasets and large datasets automatically translated from English [5]. Another similar article studied the influence of characteristics based on the category of news text in the opinion classification of Uzbek texts [6]. A semantic evaluation dataset of semantic similarity and relatedness scores in word pairs, as well as its analysis for the Uzbek language, is presented in a recent work. There is a growing trend in NLP to use artificial intelligence-based methods, as can be seen in the Uzbek work with neural transformers - an architecture-based language model [7].


Restoran haqidagi sharhlar corpusi
?
Ma'lumotlarni qayta ishlash
URLlardan tozalash
Tokenizatsiya qilish
Stopword so'zlarni olib tashlash
Lexiconlardan tozalash
TF-IDF Vectorizer
Corpus
TF-IDF Matritsasi
Machine learning (Mashinali o'qitish)
Sentiment tahlil
Deep Learning (chuqur o'qitish)
Hiziqli yadroga asoslangan Logistic Regressiya
So'z ngramlarga asoslangan Support Vector Machine
Tez matnli va tezkor matnsiz Recurrent Neural Network (RNN)
Ko'p kanalli Convolutional Neural Network (CNN)
Sentiment qutblanish

TF-IDF Vectorizer

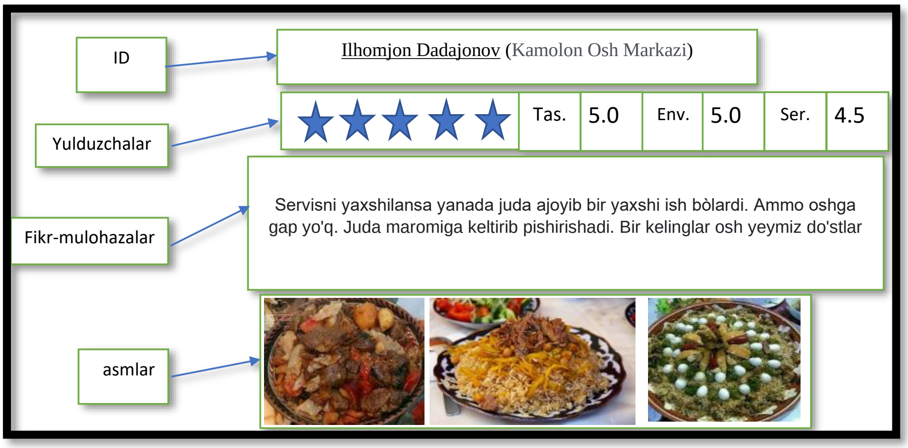
	osh	mazali	Xizmat	joy	…	N
Sharh1	0.87	0.91	0.23	0.12	…	…
Sharh2	0.56	0.98	0.75	0.45	…	…
Sharh3	0.25	0.105	0.127	0.78	…	…
…	…	...	…	…	…	…
SharhN	tfxidf_1	tfxidf_2	tfxidf_3	tfxidf_4	…	tfxidf_N

Figure 1. Scope of the study

There are works that use different methods of sentiment analysis, such as machine learning and deep learning, in their work with the idea of taking into account the differences in opinions and views of the field of sentiment analysis from a global perspective. Includes opinions from popular social platforms like Twitter, Reddit, Tumblr, and Facebook.

Methodology.

In this paper, we proposed a sentiment analysis framework based on machine learning and deep learning for the restaurant domain dataset (Figure 1) [8]. The framework is used for data collection, preprocessing (cleaning, stop words, lexicon-free stemming) [3], TF-IDF weight matrix generation [9], [10], sentiment analysis using a web browser. Includes implementation of ML and DL [11].



Servisni yaxshilansa yanada juda ajoyib bir yaxshi ish bòlardi. Ammo oshga gap yo'q. Juda maromiga keltirib pishirishadi. Bir kelinglar osh yeymiz do'stlar

Figure 2: An example of a review interface for a restaurant on the Google platform

Data collection.

We begin by reviewing the large number of datasets available for scanning in the Uzbek language.

However, conventional approaches such as Twitter or movie reviews are not suitable for the Uzbek language. That's why we decided to collect restaurant reviews because locals mostly like to give feedback about restaurants. In our opinion, this is logical, because Uzbek cuisine is one of the most popular cuisines throughout the Commonwealth of Independent States (CIS, CA countries). For example, in most cities in California, it is easy to find busy restaurants specializing in Uzbek cuisine. We have reviewed all local restaurants in Tashkent on Google Maps. First, we selected a list of more than 140 URLs with at least 3 reviews, and we obtained all the data shown in Figure 2. We reviewed Google's anti-spam and anti-DDOS policies when browsing because there are certain restrictions on data collection.

Data preprocessing.

A set of starred texts required manual correction during dataset validation. Comments containing only emojis, names, or other irrelevant content such as username mentions, URLs, or custom app names were removed. Texts written in languages other than Uzbek (mainly Russian and some English) are translated using the official Google translate API. Although people in Uzbekistan use the official Latin alphabet, the use of the old Cyrillic alphabet is equally popular, especially among adults [12]. Those notes written in Cyrillic were translated into Latin using the Uzbek machine transliteration tool. Next, we applied stop words to remove low-level information words from our comments to focus on important information. Our model is a proposed algorithm to automatically identify sets of single-word stop words using TF-IDF (Term Frequency - Inverse Document Frequency). After that, each word is processed into a lexicon-free base generator, a word size reduction algorithm due to prefixes and suffixes. The main idea is to use the combinatorial approach of matching words.

Evaluation.

The new data set collected is divided into train-set and test-set for evaluation in 8 x 2 ratio, respectively. Trainset is needed for building and training the model, and test set is needed for testing the built model. After the data cleaning process, we have the original data set as follows: where $\vec{x_i}$ represents the feature vectors and $\vec{y_i}$ represents the annotation labels:

$$(\vec{x_i}, \mathrm{yi}), \qquad i = 1, 2, 3, \ldots, N \ (1)$$

$$\vec{x_i} = (\mathrm{xi1}, \mathrm{xi2}, \ldots, \mathrm{xim}) \qquad i = 1, 2, 3, \ldots, N \ (2)$$

N and m are the number of views and the length of the feature vector, respectively. We then calculate the TF-IDF scores for each feature vector $\vec{x_i}$, which is vectorized by word extraction. Counting the frequency of a word in a given comment and the frequency between comments.

The final result of all $\vec{z_i}$ is defined as a sparse matrix.

$$\vec{z_i} = \mathrm{TF(xi)xIDF(xi)} \qquad i = 1, 2, 3, \ldots, N \quad (3)$$

Machine learning algorithms. Logistic regression model

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

$$\mathrm{P}(\mathrm{y} \,|\, \vec{z}\,) = \begin{cases} h(\vec{z}\,), & \text{if } \mathrm{y} = +1(\text{positive}) \\ 1 - h(\vec{z}\,), & \text{if } \mathrm{y} = -1(\text{negative}) \end{cases} \quad (4)$$

A logistic regression model is a classification algorithm known for its exponential and log-linear functions. It works with discrete values and any real-valued function displays 0's and 1's. Sentiment analysis shows that comments are positive or negative using formula (4).

Results and discussion.

This section provides a detailed description of the results obtained during the evaluation process using machine learning and deep learning techniques applied to the collected sentiment analysis dataset.

Experimental results

Table 1

Model name	**Sentiment**	**Accuracy %**	**Recall %**	**F1 %**	**Precision %**
Logistic regression (word-based n-gram)	positive	88	98	93	89
	negative	88	67	74	
Logistic regression (sign-based n-gram)	positive	87	51	92	87
	negative	83	97	64	
Logistic regression (n-gram based on words and characters)	positive	95	95	92	91
	negative	90	89	90	

The overall experimental results of the aforementioned evaluation were performed and the results can be seen in Table 1.

Conclusion.

In this paper, we presented a new dataset in the restaurant domain for the Uzbek language, with 8,210 reviews, annotated with positive or negative labels, which were retrieved from Google Maps using the URL addresses of all restaurant profiles in the capital city of Tashkent. and marked as their corresponding star rating. Next, we applied comprehensive preprocessing steps to the dataset to improve the accuracy of our underlying models. The best accuracy result in the data set (91%) was obtained using a logistic regression model with n-grams of words and symbols. In the near future, we plan to expand the work by collecting more data that can effectively analyze restaurant reviews at a practical level. Also, efforts are being made to eliminate the inaccuracy of the evaluation of educational experiments by using cross-validation methods in data sharing.

REFERENCES

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.