FORMATION OF A DATABASE FOR SENTIMENT ANALYSIS OF TEXTS IN THE UZBEK LANGUAGE

[1] Niyazmetova Kumushoy, [2] Raximov Komron, [3] Anvarova Dilrabo, [4] Bekjanov Ro'zimboy
[1] Teacher of the Urganch branch of TATU named after Muhammad al-Khwarizmi
[2,3,4] Student of Urganch branch of TATU named after Muhammad al-Khwarizmi
https://doi.org/10.5281/zenodo.10143621

Abstract. *In sentiment analysis of user comments, we first need to start with pre-processing the comments. Because the commentary texts were written by different people in different languages, with different spelling mistakes in the writings. If the input texts for classification algorithms in data mining are pre-processed, the accuracy of the sentiment analysis algorithm will increase and we can achieve the expected result. Solving such problems is an important task of natural language processing. 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.*
Keywords: *sentiment analysis, dataset, bag of words model, NLP, TF-IDF algorithm.*

The effectiveness of natural language processing (NLP) methods depends on a large amount of data. Sentiment analysis is the analysis of opinions expressed by people [1,2]. Reviews are usually posted by users about services, products, applications and other types of services on Google Map, Yelp, Play Market and other popular applications. They often encourage consumers to actively participate in reviews, and based on user-generated feedback, they help consumers better meet their needs and improve the quality of their product or service to win the competition and so on. they achieve. Such development allows entrepreneurs to develop without creating convenience for users. Entrepreneurs will lag behind in development if they do not take into account negative opinions about their products or services.

In addition, restaurant reviews represent the composition of customers' emotional needs and are an important source of information about consumer choice. Currently, sentiment analysis detection has achieved very high accuracy rates after applying deep learning techniques, especially for high source languages. Since the Uzbek language is an agglutinative language, one word can be a meaningful sentence. To our knowledge, there is insufficient previous work on feedback-based emotion classification problems. Thus, for this thesis, the following were considered:

A dataset based on the location of Uzbek cuisine was collected from Google Maps, where information was collected based on reviews of local national dishes. The collected comment texts were sorted into tokens [3,4] and it was observed that there were many gaps in the words. In this case, the presence of grammatical errors, dialect words, Uzbek, Russian and other languages in the text of the comments reduces the accuracy of the sentiment analysis algorithm. For example, the word "good" in the text "good", "good", "good", "good", "good" and many other uses of the same word in the text reduces the accuracy of the algorithm. Due to the fact that TF-IDF is used to determine the frequency of words in sentiment analysis, the occurrence of one word in many cases affects its frequency in the text.

Data preprocessing for train set and test set plays an important role for classification. Ready data for Machine Learning algorithms must be pre-trained. Classified based on 5-star feedback

provided by Google Maps. In this case, we consider a dataset of 1- and 2-star reviews as negative, and a dataset of 4- and 5-star reviews as positive. Since most of the comments are written in other languages such as English, Kyrgyz and Russian, it is important to translate them into Uzbek. It is advisable to translate such comments into Uzbek using the Google Translate API.

Dataset preprocessing is applied in two steps.

The first step is to remove URLs, punctuation, and lowercase letters.

The second step is to ignore the stop words in the data set based on the accuracy evaluation after generating a list of stop words using the TF-IDF algorithm;

The third step is to apply the stemming algorithm.

Based on the electronic dictionary of words in the Uzbek language, a combinatorial approach is used to make a conclusion for the Uzbek part of speech: noun, adjective, number, verb, participle, declension, and consonants. The advantages of using the algorithm are that it is lexical-free, and its complexity allows you to perform a single operation (referring to the dictionary of endings of the language):

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 are automatically translated from English. Bilingual vocabularies for the Turkish languages were obtained and used to cross-linguistically match word addition supported by a bilingual vocabulary induction assessment task. They showed that aligning words from a low-resource language can use resource-rich, closely related languages. Another similar article studied the influence of emoji-based features in the classification of Uzbek texts. 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 that uses techniques based on artificial intelligence ency exists, as can be seen in the Uzbek language work with neuron transformers - an architecture-based language model [11]. Sentiment analysis in the field of NLP, 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 differences in views and opinions from a global perspective. Includes comments from popular social platforms like Twitter, Reddit, Tumblr, and Facebook. In this thesis, a data processing, preprocessing engine for a sentiment analysis system based on machine learning and deep learning is developed for the restaurant domain review dataset. It includes web-browser data collection, preprocessing (cleaning, stopwords, lexicon-free stemming), TF-IDF weight matrix generation, and ML and DL implementation for sentiment analysis.

Figure 1: An example of feedback

Data collection. It begins with a review of the large number of datasets available for sentiment analysis in the Uzbek language [10].

Usual approaches like Twitter or movie reviews are not suitable for the Uzbek language. Therefore, it was decided to collect restaurant reviews, because locals mostly like to give feedback about restaurants. Uzbek cuisine is one of the most popular dishes in the Commonwealth of Independent States (CIS, CA countries). All local restaurants in Tashkent have been viewed from Google Maps. First, we select a list of more than 140 URLs with at least 3 reviews, and all the data shown in Figure 1 is obtained. Google's anti-spam and anti-DDOS policies have been reviewed, as there are certain restrictions on data collection.

Data preprocessing. A set of starred texts requires manual correction during dataset validation. Because automatic text correction algorithms and programs for the Uzbek language are not complete. Only comments containing emojis, names, or other irrelevant content such as username mentions, URLs, or custom app names will be removed. It is advisable to translate those written in languages other than Uzbek (mainly Russian and some in English) 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. Those comments written in the Cyrillic alphabet are translated into Latin using the Uzbek machine transliteration tool. Next, we used to 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, i.e., a tagging operation is performed [6]. The main idea is to use a combinatorial approach of matching words by removing irrelevant words [5]. Evaluation. The new data set collected is divided into train set and test set for evaluation in 8 x 2 ratio, respectively. Train Set 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 xi represent the feature vectors and yet the annotation labels:

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

$$\overrightarrow{x_i} = (x_{i1}, x_{i2} \ldots, x_{in}) \qquad i = 1, 2, 3, ..., N \ (2)$$

N and m are the number of views and the length of the feature vector, respectively. Then we compute the TF-IDF scores for each feature vector xi, which vectorizes by extracting words [7]. Counting the frequency of a word in a given comment and the frequency between comments.

The final result of all z's is defined as a sparse matrix.

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

Machine learning algorithms. Logistic regression model [8].

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

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

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).

Conclusion. In order to analyze the sentiment of Uzbek texts using logistic regression model, we need to remove as much as possible from the training texts. That is, we increase the accuracy of the model by making all the texts the same.

REFERENCES

1. Niyazmetova K., Quriyozov E. “Restoran sohasidagi o ‘zbek tilidagi matnlarning sentiment tahlili” //computer linguistics: problems, solutions, prospects. – 2023. – Т. 1. – №. 1.
2. Рахимов Х. К. и др. “O‘zbek tili sentiment analizning nazariy masalalari” //международный журнал искусство слова. – 2023. – Т. 6. – №. 1.
3. Bakaev, Ilkhom. "Creating a tokenization algorithm based on the knowledge base for the Uzbek language." 2022 International Conference on Information Science and Communications Technologies (ICISCT). IEEE, 2022.
4. Sharipov, Maksud, et al. "UzbekTagger: The rule-based POS tagger for Uzbek language." arXiv preprint arXiv:2301.12711 (2023).
5. Mahmudjonova G. “Nomuhim so ‘zlar tushunchasi va uning ahamiyati” //computer linguistics: problems, solutions, prospects. – 2023. – Т. 1. – №. 1.
6. Elov B. et al. “O‘zbek, turk va uyg ‘ur tillarida POS teglash va stemming” //Uzbekistan: Language and Culture. – 2023. – Т. 1. – №. 1.
7. Madatov K., Bekchanov S., Vičič J. “Uzbek text summarization based on TF-IDF” //arXiv preprint arXiv:2303.00461. – 2023.
8. Alisher o’g’li R. S. “Logistik regressiya modeli” //formation of psychology and pedagogy as interdisciplinary sciences. – 2023. – Т. 2. – №. 16. – С. 61-66.
9. Yusupov D.F., Abdullayeva G., Aliyev O., Hamrayeva S. Management of the different systems of oil extraction enterprise based on models of current and calendar planning// AIP Conference Proceedings 2402, 050002 (2021), 050002-1 – 050002-
10. Atanazarovich M. S., Saparbayevna R. L., Ilkhomovna A. X. DETERMINING THE KNOWLEDGE LEVEL OF PUPILS IN THE" SMART SCHOOL" INFORMATION SYSTEM //International Journal of Contemporary Scientific and Technical Research. – 2022. – С. 86-90.
11. Atanazarovich, Masharipov Sanatbek, S. Q. Iskandarov, and R. B. Sharifboyeva. "SUN’IY INTELLEKT ASOSIDA YOSHLARNI KASBGA TO‘G‘RI YO‘NALTIRISH TIZIMINI ISHLAB CHIQISH." Komputer texnologiyalari 1.10 (2022)