Development of an Algorithm for Automatic Analysis of Sentiment in School Essays of the Uzbek Language

Davlatyor B. Mengliev
Novosibirsk State University
Novosibirsk, Russia;
Urgench branch of Tashkent University of Information Technologies named after Muhammad al-Khwarizmi
Urgench, Uzbekistan
0000-0003-3969-1710

Nilufar Z. Abdurakhmonova
Department of Computer linguistics National University of Uzbekistan named after Mirzo Ulugbek
Tashkent, Uzbekistan
0000-0001-9195-5723

Vladimir B. Barakhnin
Federal Research Center for Information and Computational Technologies
Novosibirsk, Russia
0000-0003-3299-0507

Aybibi R. Iskandarova
Department of Computer linguistics National University of Uzbekistan named after Mirzo Ulugbek
Tashkent, Uzbekistan
0000-0002-5090-7909

Feruza R. Topildiyeva
Department of Computer linguistics National University of Uzbekistan named after Mirzo Ulugbek
Tashkent, Uzbekistan
0009-0008-9956-0601

Ergash Yu. Akhmedov
IT Department
Novosibirsk State University
Novosibirsk, Russia
0000-0003-4915-4089

2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE) | 979-8-3315-1632-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/PIERE62470.2024.10804909

***Abstract*—In this study, the authors presented an algorithm for automatic sentiment analysis in school essays written in the Uzbek language. The algorithm is implemented on the basis of a on a convolutional neural network architecture, designed to classify text using TensorFlow and Keras. Authors created a training dataset consisting of almost 5000 sentences and phrases, most commonly used in everyday communication. The text data underwent preprocessing, including punctuation removal and conversion to lowercase, before being transformed into numerical representations using an embedding layer that was trained simultaneously with the model. Besides, the authors tested the effectiveness of the model, where the evaluation was carried out using such metric as precision. As a result of testing, the precision reached 88 for sentiment analysis in 50 essays, which consist of 811 sentences overall. Moreover, the authors conducted a comparative analysis of existing works and proposed further options for the development of the algorithm.**

Keywords—sentiment analysis, named entity recognition, Uzbek language, convolutional neural networks, essay analysis, emotion detection, text processing, low-resource languages, custom model, text classification

I. Introduction

There is currently a strong interest in the field of text analysis, which intersects linguistics, psychology and education[1]. Such analysis was of interest before, however, due to the limitations of digital resources and computing power, research in this area was noticeably less[2]. Though, nowadays it is possible to analyze text data, especially in educational contexts[3].

While Natural language processing (NLP) tools have developed significantly for common languages such as English or Russian, low-resource languages such as Uzbek are still low-resourced in both digital resources and solutions (tools) for processing texts[4]. If we consider the concept of an essay from an educational perspective, we can note that it is more than just academic writing, since it also reflects the inner world of students, their psychological state[5]. To solve such non-trivial problems, we need reliable algorithms that can extract the objects we need from texts.

The developed algorithm is able to identify the most frequently used positive and negative words or phrases. Moreover, within framework of the education, sentiment analysis helps assess the emotional state of students, which can be an indicator of their engagement and well-being[6]. Generally, frequent use of words with negative sentiment can reflect problematic areas, prompting timely intervention. Meanwhile, positive sentiment usually reflects a healthy attitude towards learning[7].

The authors have formed the following structure of the article: Sections 1 and 2 introduce readers to the scientific field under study, as well as to the morphology of the Uzbek language. The third section discusses similar or related solutions to the problem under consideration. The fourth and fifth sections contain information on the proposed algorithm, as well as the results of its testing. The sixth section presents the conclusion of the study, as well as proposals for further development of the algorithm.

II. Morphology of Uzbek Language

Uzbek is an agglutinative language, which means that word forms are formed by concatenating affixes to the root of the word[8]. The presence of this property is due to the fact that it is part of the Turkic language family, where, however, all languages have a similar property[9]. To understand the current study, a slightly more detailed acquaintance with the Uzbek language is proposed, which is presented below.

Agglutinative nature of the Uzbek language

As mentioned above, the Uzbek language is agglutinative, which means that words are formed by combining morphemes. Each of the morphemes carries a specific grammatical function, without changing the root word[10]. These morphemes usually include various suffixes that

Authorized licensed use limited to: Tashkent University of Information Technologies. Downloaded on January 24,2025 at 04:53:24 UTC from IEEE Xplore. Restrictions apply.

indicate tense, number, case, attraction and other grammatical categories. Unlike fusional languages, where one affix can encode several grammatical features, each suffix in the Uzbek language usually has one specific function[11]. For example, consider the word "kitoblarimdan" (from my books):

- kitob (book) is the root of the word
- lar (plural) is the plural suffix
- im (my) is the possessive suffix
- dan (from) is the ablative suffix

In this single word, four different morphemes combine to convey the meaning "from my books". Each morpheme adds a specific grammatical meaning to the root word without changing its base form.

Word formation in Uzbek

Word formation in Uzbek involves a combination of derivational and inflectional processes[12]. Derivational affixes are used to create new words from existing ones, often changing the word class in the process. On the other hand, inflectional affixes modify words to express grammatical relationships without changing the underlying meaning.

Derivational morphology

Derivational affixes are often used to form nouns, adjectives, and verbs from existing roots[13]. For example:

- o‘quv (related to study) + chi (agentive suffix) = o‘quvchi (student)
- yashil (green) + lamoq (verbal suffix) = yashillamoq (become green)

In these examples, the derivational suffixes “chi” and “la” transform the root words into a noun and a verb, respectively. By manipulating affixes in this way, you can change the meaning of a word, thereby transforming it from one part of speech to another.

III. Related Works

The article [14] is devoted to the development of a morphological analysis model, which includes stemming (selecting the word stem), lemmatization (reducing the word to its initial form) and extraction of morphological features taking into account morpho-phonetic exceptions in the Uzbek language. The main focus is on the analysis of words, their stems, as well as endings that add various grammatical meanings.

The proposed approach involves creating a set of word forms, affixes, and built-in rules for morphological analysis. At the same time, it should be noted that the authors point out the uniqueness of their methodology in connection with the studies conducted regarding the grammatical rules (and affixes) used for word stemming. The model was tested on a specially selected data corpus consisting of 40 documents and 11,952 words. The model showed high accuracy of 91% when analyzing word forms. The developed tool is available as a web application and as a Python library, which makes it easy to integrate the model into other applications. However, the work contains a number of contradictions, for example, instead of the term methodology (approach), the term model is used, which is not noticeable at first glance. Moreover, the work lacks any connection with machine learning technologies and language models in particular, instead, the author offers pre-created data sets with word endings and morphological information. It should be noted that such an approach is really effective for morphological analysis, although it is difficult to adapt for other tasks due to its strong limitations. Moreover, it is difficult to adapt it for NER.

The authors of the article[15] have developed a corpus of the Uzbek language. The main goal of the work is to solve a pressing problem in resources for the Uzbek language, namely, the lack of a tagged corpus for text analysis tasks.

Moreover, this research paper also describes the creation of a new set of tags that can be used to mark parts of speech (POS) and syntax in Uzbek texts. The above-mentioned sets contain 102 morphological and 14 syntactic tags, which allow for both morphological and syntactic analysis of texts written in Uzbek. In addition, the authors have implemented a web application that can be used to manually mark texts. It should be noted that this tool can help in creating custom marked corpora, which can then be used to train machine learning models.

The language corpus developed by the authors was created in the above-mentioned web application, where it was used to mark more than 10,000 words and 1,200 sentences from various texts. Judging by the final part of the article, the authors continue to mark texts, regularly replenishing its volume. The authors also note that in the future they plan to develop automatic tagging algorithms for the Uzbek language, where machine learning technology will be implemented. Moreover, they also plan to implement a similar tool for other Turkic languages. Despite the great significance of the work done, the authors offer only a corpus, without any tools for working with text.

The article [16] presents a scientific work on creating a list of stop words for the Uzbek language. This list is intended to improve text processing, such as opinion analysis, information retrieval, text clustering and other NER tasks. Regarding the methodology, the authors collected a corpus created on the basis of school textbooks in the Uzbek language, containing more than 731,000 words.

At the same time, stop words were identified using three methods: unigrams, bigrams and the collocation method. As a result of the work, three lists were compiled: unigram stop words, bigram stop words and bigram stop words with collocations. These methods themselves are not new, but their combination and the language to which these methods were applied may be of interest. Each of these methods is implemented using an individual approach, for example, the unigram method uses TF-IDF to determine the frequency of words. About 5% of the words that best fit the criteria of stop words were selected. As a result, a list of 2,348 words was created. At the same time, the Bigram method analyzes word pairs and determines stop words based on their frequency of occurrence. The list includes 4,548 word pairs. In addition, the last collocation method uses an approach similar to the bigram method, but in this case all possible word pairs are taken into account. The list includes 24,490 word pairs.

It should be noted that the created lists of stop words can be used for automatic text processing, including preprocessing and cleaning of the source text. Despite the relevance of the work done, it can only be used as one of the mini-modules for solving a specific subtask (cleaning the text from meaningless words and phrases).

The authors of the article [17] conducted a small study on the analysis and classification of reviews of applications obtained from the Google Play Marketplace. The focus sample was 100 of the most popular applications in Uzbekistan. The authors selected several popular neural network architectures for the analysis, and also conducted three experiments that will help to make an objective assessment of the effectiveness of each model. Regarding the datasets, the first dataset was data

Authorized licensed use limited to: Tashkent University of Information Technologies. Downloaded on January 24,2025 at 04:53:24 UTC from IEEE Xplore. Restrictions apply.

that was manually annotated by the authors. The second dataset contains data that was translated into Uzbek by automatic translators. As a result of testing, it was found that the best result of working with the first dataset was shown by convolutional neural networks (accuracy 88.8%), and with the translated dataset - logistic regression (89.5%). Despite the usefulness of the information, it should be noted that the authors conducted a review study, where the task was to identify the best efficiency of work on sentiment analysis. This work can serve as a good source of knowledge for further, narrower scientific research.

IV. Proposed Solution

To solve the problem of sentiment analysis of school essays in the Uzbek language, the authors developed an algorithm based on a convolutional neural network. The proposed algorithm is focused on using a model built from scratch, which is technically a pipeline for text classification using TensorFlow and Keras. In Fig. 1 there is a test process of sentence: *Tushumda men matematika fani o'qituvchimizni ko'rdim, undan qo'rqaman, sabibi doir kech qolganimda u meni uradi*. It can be translated like: *In my dream, I saw a math teacher, I am afraid of him, because he usually beats me when I am late for lesson*.

```
t = "Tushumda men matematika fani o'qituvchimizni ko'rdi
# Convert text to a sequence of numbers
t_sequence = tokenizer.texts_to_sequences([t])
# Fill sequence to maximum length
t_sequence = pad_sequences(t_sequence, maxlen=max_sequer

# Sentiment detection
#pass the transformed sequence to the model using model.
predictions = model.predict(t_sequence)
sentiment = "Positive" if predictions[0][1] > predictior

print("Tonality of the text:", sentiment)
```

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 29ms/step
Tonality of the text: Negative
```

Fig. 1. A moment from model's work.

Researchers of the paper built a training dataset, which consist of 3,100 positive and 1,600 negative sentences, and phrases that are most often used in everyday communication. The texts were pre-processed to remove punctuation and convert to lowercase. Then, for each word in the texts, numeric representations were created using an embedding layer that was trained together with the model.

The algorithm includes the following steps:

1. The texts of phrases are converted into numeric sequences using a tokenizer, which creates a dictionary containing all the unique words in the training dataset. Each word in the texts is replaced with a number representing its position in this dictionary. This allows text data to be converted into a numeric form suitable for feeding to the model.
2. The model includes the following layers:
   - Embedding layer: Converts input words into fixed-size multidimensional vectors.
   - Conv1D: Extracts local features from texts using convolutions of size 5. The convolutional layer helps to identify patterns in the data, such as N-grams, that may be important for the classification task.
   - GlobalMaxPooling1D: Reduces the dimensionality of features, leaving only the most significant ones.
   - Fully connected layers (Dense): The first layer trains the model to combine the features extracted by the convolutional layers. The second layer performs the final classification using the sigmoid activation function.
   - Dropout layer: Regularizes the model, preventing overfitting by randomly "turning off" 20% of the neurons.
3. The model is compiled with the binary_crossentropy loss function and the Adam optimizer. Training is performed for 30 epochs using a batch size of 128.

The training and validation process is shown in Fig. 2

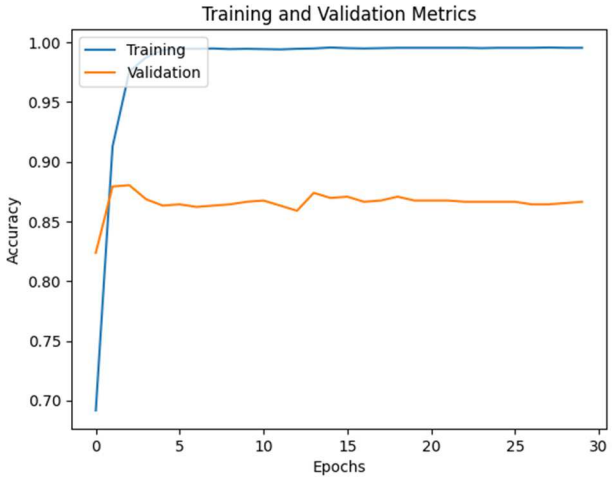

Fig. 2. Training and validation process.

The graph above shows the dynamics of accuracy for the training and validation samples during the model training process over 30 epochs. As you can see, the model shows high accuracy rates already at the initial stages, in particular, the accuracy rate is more than 85% on the validation sample. It can be concluded that the model quickly learns to find significant patterns in the data.

V. Testing and Results of the Algorithm

The testing included a sample of 50 school essays, which consisted of 811 sentences, each of which was either a positive or a negative sentence. As a result of testing, the model successfully classified 88% of the sentences (713 out of 811). One of the main problems for correct classification was the words in different forms (a wide variety of affix combinations), so the model could not recognize them. More detailed information about the testing results can be found in Table I.

TABLE I. Result of Model's Testing

Object of detection	Correctly detected / detected	Precision
Sentiment	713 / 811	88%

VI. Conclusion

In this study, an algorithm for automatic sentiment analysis in school essays written in the Uzbek language was developed. The authors substantiated the need to create such a solution and conducted a comparative analysis of existing works. The developed algorithm is based on a convolutional neural network architecture, which is technically a pipeline for text classification using TensorFlow and Keras. The model was trained on a sample, which consist of 3,100 positive and 1,600 negative sentences.

Authorized licensed use limited to: Tashkent University of Information Technologies. Downloaded on January 24,2025 at 04:53:24 UTC from IEEE Xplore. Restrictions apply.

In addition, the authors tested the model on 50 essays, which consisted of 811 sentences, each of which was either a positive or a negative sentence. As a result of testing, it was revealed that the algorithm achieved 88% accuracy in the task of classifying sentences on positive or negative respectively.

The ability to automatically analyze sentiment in student essays can provide educators with a deeper understanding of students' emotional and psychological states, thereby informing more targeted educational interventions. Future work will focus on expanding the dataset to include a wider range of topics, thereby improving the generalizability of the model.

REFERENCES

[1] J. Pennebaker, C. Chung, "Language: A window into the psychological and social worlds", Handbook of Psychology: Personality and Social Psychology, pp. 131-151, 2012.

[2] A. Mukhamadiyev, M. Mukhiddinov, I. Khujayarov, M. Ochilov, J. Cho, "Development of Language Models for Continuous Uzbek Speech Recognition System", Sensors, 23, p. 1145, 2023.

[3] E.Kuriyozov, D. Vilares, C. Gomez-Rodriguez, "BERTbek: A Pretrained Language Model for Uzbek", Special Interest Group on Under-resourced Languages workshop (SIGUL-2024), Torino, Italy, May 2024.

[4] D. Mengliev, V. Barakhnin, M. Eshkulov, B. Palvanov, N. Abdurakhmonova, S. Khamraeva, "Dictionary-Based Medical Text Analysis in Uzbek: Overcoming the Low-Resource Challenge", IEEE Ural-Siberian Conference on Computational Technologies in Cognitive Science, Genomics and Biomedicine, pp. 85-89, September 2023.

[5] Y. Tausczik, J. Pennebaker, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods", Journal of Language and Social Psychology, 29(1), pp. 24-54, 2010.

[6] C. Grimalt, M. Usart, "Sentiment analysis for formative assessment in higher education: a systematic literature review", Computing in Higher Education, 2023.

[7] D. Mengliev, V. Barakhnin, N. Abdurakhmonova, M.Eshkulov, "Developing named entity recognition algorithms for Uzbek: Dataset insights and implementation", Data in Brief, 54, 110413, 2024.

[8] D. Mengliev, V. Barakhnin, B. Ibragimov, "Rule-Based Syntactic Analysis for Uzbek Language: An Alternative Approach to Overcome Data Scarcity and Enhance Interpretability," 2023 IEEE 24th International Conference of Young Professionals in Electron Devices and Materials (EDM), Novosibirsk, Russian Federation, pp. 1910-1915, 2023.

[9] H. Türkmen, O. Dikenelli, C. Eraslan, M. C. Çalli and S. S. Ozbek, "Developing Pretrained Language Models for Turkish Biomedical Domain," 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI), Rochester, MN, USA, pp. 597-598, 2022.

[10] G. Dushaeva, "Phonological System of Modern Uzbek Language", Pindus Journal of Culture, Literature, and ELT, vol. 2, no. 5, 2022.

[11] F. Rakhmatov, "Linguoculturological and semantic features of poetic terms in English and Uzbek languages", Science and Innovation International scientific journal, vol. 4, 2022.

[12] Kh. Madatov, S.Sattarova, "Creation of a Corpus for Determining the Intellectual Potential of Primary School Students", 2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM), Altai, Russian Federation, pp. 2420-2423, 2024.

[13] I. Bakaev, T. Shafiyev, "Morphemic analysis of Uzbek nouns with Finite State Techniques", Journal of Physics: Conference Series, 1546, 2020.

[14] U.Salayev, "UzMorphAnalyser: A Morphological Analysis Model for the Uzbek Language Using Inflectional Endings", Computer Science, Computation and Language section, arXiv:2405.14179, June 2024.

[15] M. Sharipov, J. Mattiev, J. Sobirov, R. Boltayev, "Creating a morphological and syntactic tagged corpus for the Uzbek language", The International Conference and Workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing (ALTNLP), June 7-8, 2022.

[16] K. Madatov, S. Bekchanov, J. Vici, "Dataset of stopwords extracted from Uzbek texts", Data in Brief, vol. 43, 108351, 2023.

[17] E. Kuriyozov, S. Matlatipov, "Building a New Sentiment Analysis Dataset for Uzbek Language and Creating Baseline Models", Proceedings, vol. 21, issue 37, 2019.

Authorized licensed use limited to: Tashkent University of Information Technologies. Downloaded on January 24,2025 at 04:53:24 UTC from IEEE Xplore. Restrictions apply.