АЛГОРИТМИЧЕСКИЙ ПОДХОД К АНАЛИЗУ ЭМОЦИОНАЛЬНОЙ ОКРАСКИ И ПОЭТИЧНОСТИ УЗБЕКСКИХ ТЕКСТОВ

Барахнин Владимир Борисович
д.т.н., проф., Новосибирский государственный университет

Менглиев Давлатёр Бахтиярович
Ургенчский филиал ТУИТ
E-mail: shogunuz@gmail.com

Abstract: This article presents the development of an algorithm for recognizing poetic texts in the Uzbek language using a lexical approach. The research focuses on creating an extensive dictionary containing more than 10 thousand poetic words and expressions collected from contemporary literary works. The algorithm analyzes texts, comparing each word with the dictionary database, determining its emotional connotation and presence in the dictionary. A feature of the developed dictionary is the inclusion of full forms of words, which allows you to accurately preserve their meaning in the text. The results of testing the algorithm on a sample of texts showed its high.

Keywords: Uzbek poetry, recognition algorithm, vocabulary base, sentiment analysis, algorithm accuracy, artificial intelligence, lexicons, emotional richness, morphological structure, literary research.

Аннотация: Эта статья представляет разработку алгоритма для распознавания поэтических текстов на узбекском языке с использованием лексического подхода. Исследование фокусируется на создании обширного словаря, содержащего более 10 тысяч поэтических слов и выражений, собранных из современных литературных произведений. Алгоритм анализирует тексты, сравнивая каждое слово с базой данных словаря, определяя его эмоциональную окраску и наличие в словаре. Особенностью разработанного словаря является включение полных форм слов, что позволяет точно сохранять их значимость в тексте. Результаты тестирования алгоритма на выборке текстов показали его высокую эффективность, подтверждая его потенциал для глубокого анализа поэтических текстов.

Ключевые слова: Узбекская поэзия, алгоритм распознавания, словарная база, сентимент-анализ, точность алгоритма, искусственный интеллект, лексиконы, эмоциональная насыщенность, морфологическая структура, литературные исследования.

Введение

В области компьютерной лингвистики задача анализа поэтических текстов, особенно в условиях богатого многообразия узбекского языка, представляет собой сложную задачу. Поэзия отличается своими стилистическими и структурными элементами, такими как ритм, метр и образное использование языка, что требует тонкого анализа текста.

Предложенный подход основан на создании и использовании специализированных лексиконов, отражающих наиболее важные элементы поэзии, включая поэтическую лексику.

Статья начинается с введения и предоставления информации о узбекском языке, в частности, о его особенностях в поэтическом контексте. Далее в статье представлен обзор существующих исследований по данной тематике или близких к ней работ. В последующих разделах авторы излагают свое решение, а также результаты тестирования предложенного алгоритма. В заключительном разделе содержится обобщение результатов исследования, а также возможные направления для дальнейшего развития алгоритма.

Морфологические особенности

Агглютинативный характер узбекского языка подразумевает, что слова часто формируются путем последовательного добавления морфем (суффиксов)[1-2], каждый из которых выполняет определенную грамматическую или семантическую функцию[3-4]. Например, слово «kitoblaringizdan» (из ваших книг) демонстрирует, как существительное «kitob» (книга) может быть дополнено несколькими суффиксами, указывающими на притяжательность («laringiz» для вашего), множественное число («lar») и аблатив («dan» означает «из»).

Аналогичные работы

Несмотря на большое количество научных работ в области узбекской компьютерной лингвистики, авторам статьи не удалось найти научные исследования, связанные с распознаванием поэтических текстов на узбекском языке. К примеру, статья [5] посвящена разработке информационных систем для анализа поэтических текстов. Авторы рассматривают сложность анализа поэзии и используют в своих исследованиях как традиционные методы и алгоритмы классической математики, так и современные подходы, связанные с машинным обучением. Статья [6] посвящена анализу зависимости семантического содержания поэтических текстов от их метроритмических характеристик и строфической структуры. Авторы ставят целью сформулировать определение «текстуры» как структурной модели поэтического текста, которая уникально определяет метроритмический узор и строфическую структуру текста, а также разработку алгоритмов для автоматического определения текстуры.

Предлагаемое решение

Для обеспечения эффективного функционирования алгоритма была разработана словарная база, охватывающая более 10 тысяч поэтических слов и выражений. Источниками для составления словаря послужили литературные произведения современных авторов XX и XXI веков, тогда как произведения более ранних периодов в словарь не вошли в связи с их оригинальным созданием на персидском или арабском языке. Структурированный словарь представлен в виде табличного файла (.xlsx) с четырьмя столбцами:

1) идентификатор слова или фразы;
2) само слово;
3) часть речи слова;
4) уровень настроения (присущ только прилагательным, варьируется от -2 до 2, что указывает на спектр от очень отрицательной до очень положительной эмоциональной окраски).

Важно отметить, что слова в словаре представлены в полной форме с учетом всех аффиксов и приставок, чтобы избежать потери смысла при морфологическом анализе. Основная причина этого в том, что при образовании литературных слов образующиеся корни слов могут иметь совершенно разное значение, благодаря чему мы можем получить совершенно разный результат. Все элементы словаря были созданы вручную

Алгоритм работает следующим образом:

1. Входной текст разбивается на предложения, а затем на слова.
2. Каждое слово сравнивается со словарем, чтобы определить, существует оно там или нет.
3. Определяется эмоциональный оттенок слова, если он есть.
4. Подсчитывается общее количество обнаруженных стихотворных слов и словосочетаний в анализируемом тексте.
5. Численные значения тональных слов суммируются для оценки общего уровня тональностей в тексте.
6. Результаты представлены по количеству выявленных поэтических элементов и общему уровню тональности текста.

Тестирование и результаты алгоритма

Главным критерием отбора была литературная значимость и доступность текстов, в результате было отобрано 500 предложений из разных произведений. Этот образец охватывает широкий спектр тем и стилей, представляющих современную узбекскую поэзию, что позволяет оценить алгоритм в различных контекстах использования.

По результатам тестирования алгоритм демонстрирует высокую эффективность выявления поэтических элементов в тексте, достигая точности 99%. Это подтверждает его способность точно распознавать и классифицировать поэтические слова и фразы в узбекской поэзии. Кроме того, алгоритм также показал высокую точность оценки настроений, достигнув 89%. Эти результаты подтверждают эффективность разработанного алгоритма и его пригодность для анализа поэтических текстов на узбекском языке.

Наш словарь, хотя и включает более 10 тысяч поэтических слов и словосочетаний, не является исчерпывающим. В узбекской поэзии существует огромное разнообразие стилей и выражений, и некоторых уникальных слов или новых явлений может не быть в словаре. Это объясняет случаи, когда алгоритм не мог идентифицировать определенные поэтические элементы.

Кроме того, несмотря на высокую точность алгоритма определения настроений, достигающую 89%, он имеет небольшие недостатки. Основная причина – субъективность настроения, то есть эмоциональный оттенок слов может меняться в зависимости от контекста употребления, что делает его определение сложной задачей. В некоторых случаях тональность слова может интерпретироваться по-разному в зависимости от литературного контекста исходного текста.

Решением подобных проблем является изменение технологии расчета уровня настроенности предложений. Например, вместо словарного подхода использовать технологии искусственного интеллекта, хотя для этого потребуется совершенно другой тип и объем данных, значительно превышающий текущий объем размеченного словаря.

Подробности результатов показаны в Таблице I и Таблице II.

Таблица I.

Type of analyzing	Correctly detected words and phrases / all poetic words	Precision
Poetic words detection	990 / 1001	99%
Sentiment words detection	389 / 437	89%

Таблица II.

Type of sentiment words in text	Correctly detected words / all words	Precision
Very negative	23 / 35	66%
Negative	89 / 94	95%
Neutral	115 / 120	96%
Positive	92 / 99	93%
Very positive	70 / 89	79%
Overall	**389 / 437**	**89%**

Заключение

Алгоритм для распознавания поэтических текстов на узбекском языке, основанный на методе использования словарной базы, был разработан и протестирован. Алгоритм показал высокую точность в определении поэтических элементов и анализе сентиментов, достигая соответственно 99% и 89%. Эти результаты подтверждают значительный потенциал подхода, основанного на специализированных лексиконах для анализа текстов в стиле узбекской поэзии и их эмоциональной насыщенности.

Тем не менее, достигнутые результаты являются важным шагом для задач анализа поэтических текстов, а также их эмоциональной окраски. Дальнейшее развитие алгоритма, включая интеграцию с современными технологиями искусственного интеллекта, может открыть новые горизонты для исследований в области литературы.

Использованная литература:

1. D. Mengliev, E. Akhmedov, V. Barakhnin, Z. Hakimov, O. Alloyorov, "Utilizing Lexicographic Resources for Sentiment Classification in Uzbek Language," 2023 IEEE XVI International Scientific and Technical Conference Actual Problems of Electronic Instrument Engineering (APEIE), Novosibirsk, Russian Federation, pp. 1720-1724, 10-12 November 2023.
2. D. Mengliev, V. Barakhnin, M. Eshkulov, B. Palvanov, N. Abdurakhmonova, S. Khamraeva, "Dictionary-Based Medical Text Analysis in Uzbek: Overcoming the Low-Resource Challenge," 2023 IEEE Ural-Siberian Conference on Computational Technologies in Cognitive Science, Genomics and Biomedicine (CSGB), Novosibirsk, Russian Federation, pp. 85-89, 28-30 September 2023.
3. M. Sharipov, J. Mattiev, J. Sobirov, R. Baltayev, “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.

4. G. Oripova, “Uzbek poetry and the world literature in the years of independence”, Periodyk naukowy akademii polonijnej,, vol. 32, pp. 116-130, 2019.
5. O. Kozhemyakina, “Information systems for the analysis of poetic texts: history, methods and algorithms”, Journal of Computational Technologies, pp. 136-166, 2023.
6. V. Barakhnin, O. Kozhemyakina, I. Kuznetsova, V. Karpova, “The model of facture of Russian poetic texts”, Journal of Computational Technologies, pp. 107-117, 2021.