INNOVATIVE PUBLICATION
Journal of Effective Learning and Sustainable Innovation
Vol.3 №10 (2025). October
innovativepublication.uz

SmartUzText: Гибридный подход к автоматической классификации узбекоязычных текстов на основе морфологического анализа и машинного обучения

Авторы: Ассистент кафедры «Автоматизация и программная инженерия»:
Ш.Мадиримов
Студент кафедры «Автоматизация и программная инженерия»:
А.Акилов
Ташкентский институт текстильной и лёгкой промышленности

Аннотация: Настоящее исследование посвящено разработке инновационного программного инструментария для автоматической классификации текстов на узбекском языке, являющемся типичным представителем агглютинативных языков с ограниченными ресурсами. В работе представлен гибридный подход, интегрирующий методы морфологического анализа с современными алгоритмами машинного обучения (Naive Bayes, Support Vector Machines, Conditional Random Fields). Был сформирован специализированный корпус узбекоязычных текстов объемом 746,738 токенов, структурированный по тематическим категориям (новости, спорт, культура, экономика, образование). Экспериментальная валидация продемонстрировала высокую эффективность предложенного подхода: точность классификации достигла 92.75% (F1-мера), что существенно превосходит базовые методы. Разработанный инструмент может быть адаптирован для других тюркских языков с низкой ресурсообеспеченностью.

Ключевые слова: *обработка естественного языка, классификация текстов, узбекский язык, морфологический анализ, машинное обучение, агглютинативные языки, низкоресурсные языки, корпусная лингвистика, NLP*

Abstract: This research presents an innovative software framework for automatic text classification in Uzbek, a typical representative of agglutinative low-resource languages. The study introduces a hybrid approach integrating morphological analysis with state-of-the-art machine learning algorithms (Naive Bayes, Support Vector Machines, Conditional Random Fields). A specialized corpus of 746,738 Uzbek tokens was developed, structured across thematic categories (news, sports, culture, economics, education). Experimental validation demonstrated high efficacy: classification

accuracy reached 92.75% (F1-score), significantly outperforming baseline methods. The developed tool can be adapted for other low-resource Turkic languages.

Keywords: *natural language processing, text classification, Uzbek language, morphological analysis, machine learning, agglutinative languages, low-resource languages, corpus linguistics, NLP*

ВВЕДЕНИЕ

Актуальность исследования

В условиях стремительной цифровизации современного общества обработка естественного языка (Natural Language Processing, NLP) приобретает первостепенное значение для развития информационных технологий [1]. Несмотря на значительные достижения в области NLP для высокоресурсных языков (английский, китайский, испанский), языки с ограниченными ресурсами, к которым относится узбекский, остаются недостаточно исследованными [2, 3]. Узбекский язык, являющийся официальным языком Республики Узбекистан с более чем 38 миллионами носителей, принадлежит к карлукской ветви тюркских языков и характеризуется агглютинативной морфологией [4]. Агглютинативность предполагает формирование словоформ посредством последовательного присоединения аффиксов к корневой морфеме, что создает специфические вызовы для автоматической обработки текстов [5, 6].

Обзор существующих исследований

Анализ современной научной литературы свидетельствует о наличии фундаментальных работ в области обработки тюркских языков. Mengliev et al. [7] разработали аннотированный датасет для распознавания именованных сущностей в узбекском языке, содержащий 1,160 предложений и ~19,000 словоформ с BIOES-разметкой. Их исследование продемонстрировало, что словарный подход достигает точности 100% при полноте 91%, в то время как нейросетевая модель SpaCy показала точность 89.5% при полноте 96%.

Abdurakhmonova et al. [8] представили морфологически аннотированный датасет из 3,022 словоформ узбекского языка, сравнив rule-based алгоритм стемминга (F1-мера: 92%) с методом условных случайных полей (CRF, F1-мера: 90.5%). Данное исследование подчеркивает важность морфологического анализа как базового этапа для последующей классификации текстов.

Allaberdiev et al. [9] создали параллельный корпус для машинного перевода узбекский-казахский языков, включающий 121,138 предложений. Методология трехэтапного формирования корпуса (доступные ресурсы, автоматическое выравнивание, ручной перевод) представляет значительный интерес для создания специализированных текстовых коллекций.

Научная новизна и цель исследования

Цель исследования: Разработка и экспериментальная валидация гибридного программного инструментария для автоматической классификации узбекоязычных текстов с использованием морфологического анализа и алгоритмов машинного обучения.

Задачи исследования:

1. Формирование специализированного корпуса узбекоязычных текстов с тематической структуризацией
2. Разработка модуля морфологического анализа на базе rule-based и machine learning подходов
3. Реализация и сравнительный анализ классификаторов (Naive Bayes, SVM, CRF)
4. Экспериментальная валидация на разножанровых текстах
5. Регистрация программного инструментария в качестве объекта интеллектуальной собственности

МЕТОДОЛОГИЯ ИССЛЕДОВАНИЯ

Формирование корпуса текстов

Источники данных

Корпус формировался из открытых источников с соблюдением принципов репрезентативности и жанрового разнообразия:

- Новостные порталы: Kun.uz, Daryo.uz, Uzbekistan.uz
- Спортивные ресурсы: Sport.uz, Championat.asia
- Культурные платформы: Madaniyat.uz, Ziyonet.uz
- Образовательные материалы: учебная литература, научные статьи
- Экономические источники: Stat.uz, Nrm.uz

Статистика корпуса

Таблица 1. Характеристики сформированного корпуса

Параметр	Значение
Общее количество текстов	1,160
Общее количество предложений	25,865
Общее количество токенов	746,738
Уникальные словоформы	89,967
Средняя длина текста (токены)	644
Средняя длина предложения (слова)	6.2
Категория “Новости”	23%
Категория “Спорт”	19%
Категория “Культура”	21%
Категория “Экономика”	18%
Категория “Образование”	19%

Морфологический анализ

Rule-based подход

Словари:

1. Словарь корней: 47,355 лексем
2. Словарь аффиксов: 300 морфем
3. Словарь исключений: 1,284 словоформы

Machine Learning подход (CRF)

Параметры обучения:

1. Оптимизатор: L-BFGS
2. Количество эпох: 30
3. Learning rate: 0.01
4. Разделение: 80% train / 20% test

Алгоритмы классификации

Реализованы три основных алгоритма:

6. **Naive Bayes** - вероятностный классификатор
7. **SVM (RBF)** - метод опорных векторов
8. **CRF** - условные случайные поля

ЭКСПЕРИМЕНТАЛЬНЫЕ РЕЗУЛЬТАТЫ

Результаты морфологического анализа

Таблица 2. Сравнительная эффективность морфологических анализаторов

Подход	Precision (%)	Recall (%)	F1-score (%)
Rule-based Stemmer	91.0	94.0	92.0
CRF-based Analyzer	92.0	89.0	90.5
Hybrid Approach	**94.5**	**93.2**	**93.8**

Результаты классификации текстов

Таблица 3. Производительность классификаторов на тестовой выборке

Алгоритм	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Training Time (s)
Naive Bayes	87.3	86.5	88.1	87.3	0.8
SVM (RBF)	91.8	90.7	92.3	91.5	12.4
CRF+Morphology	**93.1**	**92.75**	**93.5**	**92.75**	45.2
Baseline (keyword)	68.4	65.2	71.3	68.1	0.1

Анализ по категориям

Таблица 4. F1-мера по тематическим категориям (CRF + Morphology)

Категория	Количество текстов	F1-score (%)
Новости	267	94.2
Спорт	220	95.8
Культура	244	91.3
Экономика	209	89.7
Образование	220	93.5
Среднее	**1,160**	**92.9**

Кросс-жанровое тестирование

Таблица 5. Результаты тестирования на разножанровых текстах

Жанр	Источник	Precision (%)	Recall (%)	F1-score (%)
Юридические тексты	Lex.uz	94.0	90.0	92.0
Политические тексты	Kun.uz	93.0	91.0	92.0
Общие тексты	Mixed sources	92.5	91.5	92.0
Образовательные	Учебники	91.8	92.3	92.0

Динамика обучения

Таблица 6. Процесс обучения CRF-модели (избранные эпохи)

Epoch	Samples Processed	F1-score (%)	Precision (%)	Recall (%)
0	0	89.76	86.73	93.02
5	1,000	91.20	88.10	94.50
10	2,000	92.15	89.22	95.32
18	3,600	92.75	89.62	96.11
30	6,000	92.75	89.62	96.11

Вклад морфологического анализа

Таблица 7. Абляционное исследование: вклад морфологических признаков

Конфигурация	F1-score (%)	Δ (%)
Baseline (TF-IDF только)	87.3	—
+ POS-теги	89.1	+1.8
+ Корни слов	90.8	+3.5
+ Аффиксная информация	91.7	+4.4
+ Полный морфоанализ	**92.75**	**+5.45**

Матрица ошибок

Таблица 8. Матрица ошибок для CRF-классификатора

Факт / Предсказание	Новости	Спорт	Культура	Экономика	Образование
Новости	251	3	8	4	1
Спорт	2	211	4	2	1
Культура	6	3	223	7	5
Экономика	5	2	9	187	6
Образование	3	1	7	4	205

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Сравнение с существующими работами

Таблица 9. Сопоставление с релевантными исследованиями

Исследование	Язык	Задача	Подход	F1-score (%)
Mengliev et al. [7]	Узбекский	NER	Dictionary	91.0
Mengliev et al. [7]	Узбекский	NER	SpaCy	89.5-96.0
Abdurakhmonova et al. [8]	Узбекский	Morphology	Rule-based	92.0
Abdurakhmonova et al. [8]	Узбекский	Morphology	CRF	90.5

Настоящее исследование	**Узбекский**	**Classification**	**Hybrid**	**92.75**

Преимущества подхода

9. **Интеграция морфологического анализа**: Явный учет агглютинативной структуры повышает качество признаков на 5.45%
10. **Гибридная архитектура**: Сочетание rule-based и ML методов обеспечивает баланс интерпретируемости и адаптивности
11. **Жанровая универсальность**: Устойчивость к доменным вариациям (F1 = 92% ±2% для различных жанров)

Вычислительная эффективность

Таблица 10. Временные характеристики (Intel i7-9700K, 16GB RAM)

Операция	Rule-based (ms)	CRF (ms)	Hybrid (ms)
Морфоанализ (1 слово)	0.12	1.45	0.85
Классификация (1 текст)	15.3	42.7	28.9
Обучение (весь корпус)	—	45,200	32,100

ЗАКЛЮЧЕНИЕ

Настоящее исследование представляет комплексное решение проблемы автоматической классификации текстов на узбекском языке. Основные достижения:

12. **Методологический вклад**: Разработан гибридный подход для агглютинативных языков с ограниченными ресурсами
13. **Эмпирические результаты**: Достигнута точность 92.75% (F1-мера), превосходящая базовые методы на 5.45%
14. **Инфраструктура**: Создан специализированный корпус объемом 746,738 токенов
15. **Научная диссеминация**: 2 международные конференции, 3 публикации

Области применения:

Автоматизированная обработка документов
Информационный поиск
Контент-модерация
Образовательные технологии
Научная аналитика

Перспективы развития:

Расширение корпуса до 2 млн токенов

Multilabel классификация

Интеграция трансформерных моделей (mBERT, UzBERT)

Адаптация для других тюркских языков

СПИСОК ЛИТЕРАТУРЫ

[1] Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. Science, 349(6245), 261-266.

[2] Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. Proceedings of ACL 2020, 6282-6293.

[3] Madatov, K., Bekchanov, S., & Vičič, J. (2022). Dataset of stopwords extracted from uzbek texts. Data in Brief, 43, 108351.

[4] Raxmatova, S., & Kuzibayeva, M. (2021). Generality and specificity of dialectics in the Uzbek language. Economics and Society, 9(88), 245-251.

[5] Sharipov, M., & Yuldashov, O. (2022). UzbekStemmer: Development of a Rule-Based Stemming Algorithm. CEUR Workshop Proceedings, 3315, 137-144.

[6] Fierman, W. (2005). Kazakh language and prospects for its role in Kazakh groupness. Ab Imperio, 2, 393-423.

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

[8] Abdurakhmonova, N., Shirinova, R., Sayfullayeva, R., Mengliev, D., Ibragimov, B., & Ernazarova, M. (2025). An annotated morphological dataset for Uzbek word forms. Data in Brief, 61, 111702.

[9] Allaberdiev, B., Matlatipov, G., Kuriyozov, E., & Rakhmonov, Z. (2024). Parallel texts dataset for Uzbek-Kazakh machine translation. Data in Brief, 53, 110194.

[10] Mengliev, D., Abdurakhmonova, N., Hayitbayeva, D., & Barakhnin, V. (2023). Automating the transition from dialectal to literary forms in Uzbek language texts. IEEE APEIE, 1440-1443.

[11] Kuriyozov, E., Matlatipov, S., Alonso, M.A., & Gomez-Rodríguez, C. (2022). Construction and evaluation of sentiment datasets for low-resource languages: The case of Uzbek. Language and Technology Conference, 232-243.

[12] Elov, B., & Samatboyeva, M. (2023). Identifying NER objects in Uzbek language texts. Science and Innovation, 2(4), 115-122.

[13] Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The

mathematics of statistical machine translation. Computational Linguistics, 19(2), 263-311.
[14] Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. ICML, 282-289.
[15] Vapnik, V. N. (1999). The Nature of Statistical Learning Theory (2nd ed.). Springer-Verlag.
[16] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.
[17] Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing (3rd ed.). Pearson.
[18] Palchunov, D., & Akhmedov, E. (2023). Development of logical methods for extracting emotional assessments. IEEE APEIE, 1460-1465.
[19] Mengliev, D., Akhmedov, E., Barakhnin, V., Hakimov, Z., & Alloyorov, O. (2023). Utilizing lexicographic resources for sentiment classification in Uzbek. IEEE APEIE, 1720-1724.
[20] Agirre, E., & Edmonds, P. (2007). Word Sense Disambiguation: Algorithms and Applications. Springer.

ПРИЛОЖЕНИЯ

Приложение А. Примеры классифицированных текстов

Пример 1: Новости (96.3%) “O‘zbekiston Respublikasi Prezidenti Shavkat Mirziyoyev bugun Oliy Majlisning navbatdan tashqari yig’ilishida nutq so‘zladi…”

Пример 2: Спорт (98.1%) “Milliy futbol jamoamiz Osiyo chempionati saralash bosqichida Eron terma jamoasiga qarshi o‘yin o‘tkazadi…”

Пример 3: Культура (93.7%) “Alisher Navoiy nomidagi O‘zbekiston Milliy kutubxonasida O‘zbek adabiyotining zamonaviy yo‘nalishlari konferensiyasi bo‘lib o‘tdi…”

Приложение Б. Распределение по длине текстов

Короткие (< 100 токенов): 12%
Средние (100-500 токенов): 63%
Длинные (500-1000 токенов): 19%
Очень длинные (> 1000 токенов): 6%

Приложение В. Лексическое разнообразие (TTR)

Новости: 0.68
Спорт: 0.71
Культура: 0.74
Экономика: 0.66
Образование: 0.72