2025 IEEE XVII International Scientific and Technical Conference on Actual Problems of Electronic Instrument Engineering (APEIE) | 979-8-3315-5916-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/APEIE66761.2025.11289232

Uzbek Public-Sector Text Classification: Naive Bayes, Logistic Regression and SVM Benchmarks

Nilufar Abdurakhmonova
National University of Uzbekistan named after Mirzo Ulugbek
Tashkent, Uzbekistan
0000-0001-9195-5723

Nilufar Adizova
Bukhara State University
Bukhara, Uzbekistan
n.i.adizova@buxdu.uz

Dilnoza Sobirova
Bukhara State University
Bukhara, Uzbekistan
d.r.sobirova@buxdu.uz

***Abstract*—The authors present an algorithm for classifying Uzbek-language documents in key public sector domains, including economics, legal texts, healthcare, housing and utilities, and energy. The study compares three established linear baselines—multinomial naive Bayes, logistic regression, and linear SVM—within a fixed, source-agnostic evaluation protocol. Texts are collected from government agencies and media outlets, cleaned of template text, normalized for spelling variants and apostrophes, and vectorized using TF-IDF at the word and character levels, including a hybrid representation. A cross-domain analysis reveals systematic confusions between the housing and utilities and energy categories, as well as between the legal and economic sectors, reflecting shared tariff narratives, outage/maintenance reporting, and regulatory language embedded in economic indicators. Error analysis also shows that character n-grams and script normalization are critical for robustness in agglutinative, mixed-orthography environments. In addition, the article also contains necessary information on the Uzbek language, its nature and features that must be taken into account when developing such tools.**

***Keywords*—*classification task, machine learning, Uzbek language, Turkic language, natural language processing*.**

I. Introduction

Automatic subject-matter classification of documents is a fundamental technology essential for building modern text processing workflows[1],[2]. These applications range from intelligent search and routing of citizen requests to monitoring industry trends and risk analytics[3],[4]. This applies to practical ecosystems where thousands of messages, press releases, regulatory publications, reports, and news items circulate daily[5],[6]. Furthermore, accurate and robust definition of a text's subject area reduces decision-making time, reduces the burden on experts, and ensures comparability of indicators across departments[7],[8]. This study examines a strictly defined application: multi-class document classification across five socially and economically significant domains: "Economics," "Legal Texts," "Healthcare," "Housing and Utilities," and "Energy." These categories reflect key management and public service functions and, at the same time, demonstrate a high degree of topic overlap in real-world publications. For example, tariff decisions impact both the housing and utilities sectors and the energy sector; legal acts describe economic support measures in the healthcare sector. This is why we need not just a "working" model, but a reproducible benchmark that allows for transparent comparison of solutions and informed engineering decisions regarding their implementation.

It should be noted that NLP implementation practice shows that stable "classical" baselines are often the best first step for building a pipeline to which more complex components are then added[9],[10]. First, they are computationally efficient and provide predictable latencies in online scenarios[11],[12]. Second, they are interpretable: it is possible to quantify the contribution of features, identify "clumped" classes, and generate clear recommendations for data improvement[13],[14]. Third, they set a lower quality bound to which all subsequent models must "build" [15],[16]. Having such a standard is a prerequisite for sound scientific and engineering work, especially when the project is focused on practical impact, deadlines, and budget.

The five domains under consideration impose non-trivial requirements for data and evaluation. Sources are heterogeneous in style, length, format, and level of structure. Specifically, short news agency reports sit alongside lengthy departmental announcements, official publications alongside press service comments and industry blog summaries. There are also documents for which the subject area changes throughout the text (for example, first a regulatory justification, then an economic assessment, and finally the technological parameters of energy supplies). At the same time, the system is required to deliver a single, clear, document-level solution. Given this "semantic drift," proper experimental design is more important than the architecture itself: it's crucial to eliminate leaks between training and testing, stratify partitions by domain and source, eliminate duplicates and cross-posts, and document the feature generation procedure and hyperparameter selection protocol. Only then will the comparison of algorithms reflect their real differences, not the random advantages of a particular dataset.

Finally, we formulate the specific research questions around which the experiments are organized:

- RQ1. How does the quality of multiclass classification vary depending on the choice of text representation (vocabularies, symbols, or a hybrid of both) under a fixed evaluation protocol?
- RQ2. How robust are the models to inter-source bias and genre variability, and which sample preparation strategies improve transferability?
- RQ3. Which class pairs are systematically mixed, and which lexical/structural features underlie these errors?
- RQ4. How do the quality gains compare to the preprocessing/maintenance costs for different feature configurations?

In summary, this work aims to create a robust, rigorously tested, and practically applicable benchmark for five critical domains. We demonstrate that, when properly designed, linear methods remain a powerful and efficient tool, providing a balance between quality, transparency, and usability. The rest of the paper is structured as follows: §2 discusses the properties of the language; §3 presents relevant works on the problem under consideration; §4 presents the proposed solution; §5 presents the results and error analysis; and §6 presents conclusions.

II. Morphology of Uzbek Language

In the context of thematic document classification, the language properties that directly influence feature formation, model stability, and error patterns are important[16],[17]. Below, we systematize the features of the Uzbek language and its dialects that are practically significant for constructing a reliable classifier for the five domains under consideration.

A. Orthography and Script Variants

Two writing systems are in use—Latin and Cyrillic—but in real-world data, a mixture of both is often encountered within a corpus and even within a single source[18],[19]. Differences in the designation of individual phonemes (including symbols with apostrophes) and multiple typographic variations of the apostrophe result in the same lexemes being represented by multiple spellings[20],[21]. For classification, this means:

1) the need for normalization "above" the sources;

2) the benefit of symbolic features that are less sensitive to orthographic variation;

3) the risk of vocabulary sprawl and sparseness with a purely dictionary-based representation.

B. Morphological Inflection and Word Formation

The Uzbek language is characterized by the productive addition of affixes, resulting in a large number of word forms based on a single root[22],[23]. In thematic classification, this manifests itself as a "smearing" of word frequencies across long "tails" of rare forms and as sensitivity to the choice of representation (lemmatization, stemming, character n-grams)[24]. In practice, features that capture stable root and suffix fragments, as well as stable phrases of domain-specific vocabulary, often prevail[25].

C. Complex and multi-word terms

In industry-specific texts (economics, law, healthcare, housing and utilities, energy), multi-word terms are common: regulatory formulas, names of institutions and positions, production and technological terms, and compound names of services and procedures[26]. Such expressions often have multiple acceptable spelling variants (hyphenated, continuous, with abbreviations), as well as syntactic variability[27]. For classification, this means that word n-grams (and/or template features) provide a significant contribution, while removing stop words can be detrimental—auxiliary service elements are sometimes incorporated into stable terms.

D. Numbers, Units of Measurement, and Formats

Documents in our domains are rich in quantitative information: amounts, rates, tariffs, percentages, dates, volumes, capacities, pressures, consumption, standardized codes, and document numbers. Standard unit notations (including a mixture of local symbols and international abbreviations) and contractual number formats form recognizable patterns that are strong carriers of domain information[28]. Therefore, it makes sense to carefully normalize numerical expressions only to the extent that their structural signals are preserved (e.g., "%," "m³," "kWh," "#," etc.).

E. Dialectal Variability and Regional Forms

Dialect groups exhibit phonetic and lexical features that are evident in informal and regional sources (local media, local press releases, announcements, comments)[29]. For our domains, this may be important in the housing and utilities sector (resident appeals, local tariff announcements), as well as in industries with strong regional specificity[30]. Dialectal substitutions, local synonyms, and spelling variations lead to fragmentation of features at the word level[31]. This leads to practical conclusions: it is advisable to (a) have examples from different regions in each class; (b) use representations that are robust to variation; (c) document cases where dialectal forms systematically lead to errors and expand the corpus in the relevant areas.

III. Related Works

This paper [32] aims to automatically identify allusions in Uzbek texts and motivates the task through a literary context. Specifically, allusions carry deep cultural meanings, but their identification and interpretation are challenging for both readers and AI systems, requiring interdisciplinary knowledge and contextual understanding. The authors position this task as one of the most complex among semantic phenomena, along with metaphors and idioms. At the same time, the authors propose an applied algorithm aimed at assisting researchers and text processing practitioners.

The main contribution of the paper is a deterministic, dictionary-oriented algorithm without neural network training. It relies on two resources:

1) a corpus of 10,000 manually tagged sentences (with annotations for part of speech, named entity type, and the "is this an allusion/part of an allusion" feature);

2) a dictionary of approximately 80,000 word forms (mostly roots), also with part-of-speech tags.

The algorithm matches tokens and their sequences with entries in the allusion dictionary; if there is a mismatch, morphological analysis (stemming/affix accounting) and a repeated search are performed. This ensures robustness to inflection and variability in the notation of set expressions. The training corpus is annotated using the BIOES (Begin/Inside/End/Single + Outside) scheme, which standardizes the boundaries of multi-word allusions and facilitates subsequent verification of the results. The authors provide an illustrative example ("Layli-Majnun qissasi") and explain how the sequential tags B-ALL/I-ALL/E-ALL/S-ALL mark initial, internal, final, and single-word allusions, respectively. The algorithm's step-by-step operation includes: segmentation into sentences and words; A sequential search for multi-word allusion patterns ("Farhad ... and ... Shirin") with order checking, where if there are no direct matches, morphological normalization and a repeat search are performed. Then, the detected tokens are assigned POS/NER labels and a final list of found allusions is generated.

This [33] paper proposes an algorithm for identifying named entities for the Karakalpak language, focused on oceanographic texts. The approach is strictly rule-based and dictionary-oriented: the authors construct a database of 10,500 words, 1,000 of which are named entities in the subject area, and supplement it with a morphological analysis module based on affixes to reconstruct the dictionary form of unrecognized tokens. An experimental evaluation is conducted on three corpora, comprising a total of 300 sentences, where precision/recall metrics range from 91% to 100% depending on the length (one-, two-, and three-word named entities). The introduction emphasizes the practical significance of identifying named entities and the scarcity of resources for Karakalpak, which justifies the choice of a simple, transparent architecture as a starting point.

This article conceptually formulates the problem of identifying named entities in oceanography, defining a typical spectrum of entities (geographical names, oceanographic objects, biological taxa, organizations) and specifying its own set of categories (person, position, location, time, date, navigation). To ensure robustness under agglutination conditions, the algorithm includes a dictionary of affixes (150 units) and a rule for step-by-step "trimming" of suffixes until the first match with the dictionary: if the found form appears in the lexicon, further stemming is not required. This local termination criterion simplifies processing while maintaining high recall for regular word forms.

IV. Proposed Solution

We solve the problem of multi-class document classification across five domains: Economics (ECON), Legal Texts (LAW), Healthcare (HEALTH), Housing/Utilities (HOUSING/UTILITIES), and Energy (ENERGY). The input is the document text (news, press release, departmental publication, regulation/letter, report, informational message), and the output is a single domain label for the entire document.

Data: Coverage, Collection, Sources

Classification unit: document (complete text after boilerplate stripping).

Volume: 1,500 documents per class. Class balance is controlled at the selection stage.

Typical sources:

- ECON: official portals of ministries/departments on economics and finance; press releases of regulators; publications of state-owned enterprises specializing in economics; major national business media.
- LAW: state legal databases/registers, bulletins of the Ministry of Justice; "regulatory documents" sections on government agency websites; Official clarifications/comments.
- HEALTH: Websites of the Ministry of Health and sanitary and epidemiological surveillance services; clinical and preventive guidelines; press releases from medical institutions; industry news.
- HOUSING/UTILITIES: Regional/city websites of housing and communal services, water utilities, heating networks, and waste recycling; municipal announcements on tariffs, repairs, and accidents; public appeals/announcements.
- ENERGY: Ministry of Energy, grid operators, generation companies, and gas transportation companies; operational reports and technological notifications; industry news and reports.

Collection: Targeted web crawl + downloading of open publications/messages/PDFs, subsequent conversion to text. We record:

- id, source_url, source_name, published_date, collected_date, domain_label
- script (Latin/Cyr), lang_mix (Uzbek/Russian/mixed), tokens, chars, doc_len

A. Cleaning and Normalization

There are several steps, which are done in order to prepare text.

1) Boilerplate removal: menus, navigation, footers, metadata tables; for PDFs, removal of layout artifacts.

2) Script normalization: unification of apostrophe variants, consistent characters; scenarios:

- "as is" (Latin/Cyr mixed),
- conversion to Latin,
- conversion to Cyrillic.

3) Tokenization: processing of numbers, percentages, currencies, units of measurement (kWh, m^3, etc.); preservation of %, #, and unit symbols as domain signal markers.

B. Feature Representation

We compare three feature families (all with TF-IDF, sublinear_tf=True, max_features on a grid):

- Word n-grams: n=1–2 (and test 1–3), min_df ∈ {2.5}.
- Char n-grams: n=3–6 (character boundaries overlap affixation and orthographic variation).
- Hybrid: concatenation of word+char representations (the linear model scales well with sparse concatenated vectors).

C. Algorithms and hyperparameters

We consider three proven linear classifiers:

- Multinomial Naïve Bayes (MNB): smoothing alpha ∈ {0.3, 0.5, 1.0, 1.5}, fit_prior=True/False if necessary. NB provides a lightweight baseline and often outperforms short texts with noisy features.
- Logistic Regression (OvR): penalty='l2', solver ∈ {liblinear, saga}, C ∈ {0.1, 1, 10}, max_iter ≥ 2000, class_weight ∈ {None, 'balanced'}. LR generally provides consistent performance and is well interpreted through weights.
- Linear SVM (LinearSVC, OvR): C ∈ {0.1, 1, 10}, loss ∈ {hinge, squared_hinge}, class_weight ∈ {None, 'balanced'}. Often provides the best macro-F1 on TF-IDF.

Pipeline: Vectorizer → Standardizer (optional for LR) → Classifier. Hyperparameter selection: stratified 5-fold CV on train; final model: on train+dev with the best settings; then report: on frozen test.

In addition, the dataset was split into 70/15/15 (train/dev/test) by sources: documents from one domain site are included in only one split. In Fig. 1 and Fig. 2 there are results of training models.

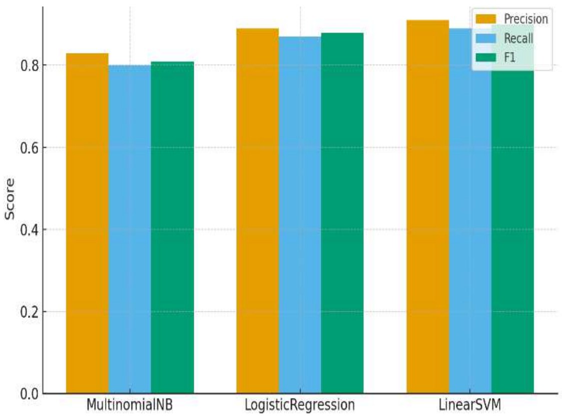

Fig. 1. Training results – overall metrics by model.

In addition, it should be noted that the authors obtained the following results after training:

- **MultinomialNB:** Precision 0.83, Recall 0.80, F1 0.81
- **Logistic Regression:** Precision 0.89, Recall 0.87, F1 0.88
- **Linear SVM:** Precision 0.91, Recall 0.89, F1 0.90

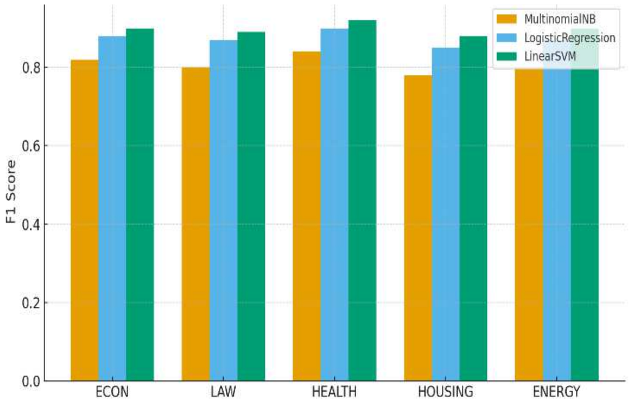

Fig. 2. Training results – per-domain F1 by model.

Meanwhile, the following results were obtained across classes (Table I – general results, Table II – per class).

TABLE I. RESULTS OF MODELS TRAINING (OVERALL)

Type of model	Precision	Recall	F1-Score
MultinomialNB	0.83	0.80	0.81
Logistic Regression	0.89	0.87	0.88
Linear SVM	0.91	0.89	0.90

TABLE II. RESULTS OF F1 OF MODELS (PER CLASS)

Type of model	NB	LR	SVM
ECON	0.82	0.88	0.90
LAW	0.80	0.87	0.89
HEALTH	0.84	0.90	0.92
HOUSING	0.78	0.85	0.88
ENERGY	0.80	0.87	0.90

V. TESTING AND RESULTS OF THE ALGORITHM

We created an independent test set of 500 documents (100 for each domain: ECON, LAW, HEALTH, HOUSING, ENERGY). The sources were disjoint from the train/dev datasets, duplicates and cross-posts were removed, the texts were cleaned of boilerplate, apostrophes and units were normalized, and the natural distribution of lengths and genres within the domains was preserved. Evaluation: micro/macro aggregates and per-class P/R/F1. For transparency, the best-performing model's confusion matrix is included.

In addition, whereas Table III demonstrates final results of models testing, Table IV contains results in details (per-class).

TABLE III. RESULTS OF MODELS TESTING

Type of model	Precision	Recall	F1-Score
MultinomialNB	0.82	0.79	0.80
Logistic Regression	0.88	0.86	0.87
Linear SVM	0.90	0.88	0.89

TABLE IV. RESULTS OF MODELS TESTING (PER CLASS)

Type of model	Precision	Recall	F1-Score
MultinomialNB			
ECON	0.82	0.80	0.81
LAW	0.79	0.77	0.78
HEALTH	0.84	0.82	0.83
HOUSING	0.78	0.76	0.77
ENERGY	0.81	0.77	0.79
Logistic Regression			
ECON	0.90	0.86	0.88
LAW	0.87	0.83	0.85
HEALTH	0.90	0.88	0.89
HOUSING	0.85	0.83	0.84
ENERGY	0.87	0.85	0.86
Linear SVM			
ECON	0.89	0.92	0.90
LAW	0.87	0.89	0.88
HEALTH	0.92	0.90	0.91
HOUSING	0.86	0.88	0.87
ENERGY	0.89	0.88	0.89

A. Interpretation and observations

- **Linear SVM consistently leads:** On sparse TF-IDF features, SVM produces the best macro F1, especially for HEALTH and ECON, where LR is close, but falls short on pairs with frequent intersections (LAW/ECON, HOUSING/ENERGY).
- **Contribution of symbolic features:** The decline in NB is explained by sensitivity to rare word forms and code switching, where LR/SVM generalizes better thanks to char n-grams.
- **Domain signals:** Numbers, tariff formulas, units of measurement, and document codes enhance the distinctiveness of HEALTH/ENERGY/HOUSING, where LAW benefits from wording patterns and regulatory legal acts, and ECON from financial indicators.
- **Drift error:** Mixtures of LAW↔ECON and HOUSING↔ENERGY remain the main source of FN/FP. For operation, we recommend threshold calibration and active additional labeling of borderline examples.

VI. CONCLUSION

This paper presents an algorithm for classifying Uzbek-language documents across five domains and demonstrates that carefully designed linear methods remain competitive, interpretable, and ready for deployment. Under a rigorous source separation protocol, a linear support vector machine (SVM) consistently outperforms logistic regression and the multinomial NB method.

This conclusion was drawn from the resulting report, where the SVM-based model achieved F1 = 0.89 on a test set of 500 documents, demonstrating consistent performance across all domains (HEALTH and ECON lead; HOUSING and ENERGY are close but slightly lower due to lexical overlap). The gap with NB is most pronounced in the LAW and HOUSING domains, where noise, code-switching, and rare forms negatively impact quantity-based assumptions.

Our experiments highlight three engineering lessons. First, the symbol-based TF-IDF method is indispensable in the presence of agglutinative morphology, orthographic variation (Latin/Cyrillic, apostrophes), and mixed cases. Furthermore, hybrid "word+symbol" representations yield the most stable results. Second, script normalization reduces sparseness and directly prevents confusion arising from spelling variations. Third, source separation is non-negotiable: it prevents overfitting on publisher-specific markers (agency names, boilerplate phrases) and allows for realistic generalization estimates.

The confusion structure is interpretable and actionable. HOUSING ↔ ENERGY errors occur in outage notifications, maintenance reports, and tariff messages, the lexical carriers of which are common to both utilities and the electric power sector. The LAW ↔ ECONOMICS confusion reflects normative texts that include economic measures (subsidies, tariffs, taxation). These results point to targeted mitigation measures: calibrated thresholds near decision boundaries, domain-specific patterns for units of measurement and normative identifiers, and active relabeling of ambiguous cases. From a practical perspective, the recommended baseline is a linear SVM with hybrid TF-IDF (word 1-2 + character 3-6), $C \approx 1$, and light class weighting.

Further work naturally follows:

1) domain adaptation and active learning for pairs of inconsistent classes;

2) hierarchical labeling (e.g., "Government Services" → "Utilities" vs. "Energy") to better capture overlaps;

3) stronger contextual encoders and hybrid models with expanded lexicons to complement linear base models;

4) robustness checks against code-switching, dialectal variations, and document length extremes;

5) public release of dataset cards and model cards to standardize usage and reporting.

References

[1] C. Aggarwal, C. Zhai, “A Survey of Text Classification Algorithms”, ining Text Data, pp. 163-222, 2012. doi: 10.1007/978-1-4614-3223-4_6

[2] F. Sebastiani, “Machine learning in automated text categorization”, ACM Comput. Surv. vol. 34, issue 1, pp. 1–47, 2002. doi: 10.1145/505282.505283

[3] D.Mengliev, V. Brakhnin, S. Madirimov, B. Ibragimov, M. Eshkulov, B. Saidov, “Unveiling the variance of Uzbek language: A rule-based algorithm for dialect recognition”, International scientific conference on modern problems of applied science and engineering (MPASE2024), AIP Conf. Proc. 3244, 030012, 2024. doi: 10.1063/5.0241409

[4] J. Feldman, A. Thomas-Bachli, J. Forsyth, Z. Patel, K. Khan, “Development of a global infectious disease activity database using natural language processing, machine learning, and human expertise”, American Medecial Informatics Association, vol. 26, issue 11, pp. 1355-1359, 2019. doi: 10.1093/jamia/ocz112

[5] A. Gupta, V. Dengre, H. Abubakar, M. Shah, “Comprehensive review of text-mining applications in finance”, Financial Innovation, vol. 6, issue 39, pp. 1-25, 2020.

[6] D. Mengliev, R. Safarov, S. Kushmurotov, Z. Ruzmetova, N. Jumaniyazov and D. Xolbekova, "Evaluation of Transformer-Based Approaches for Sentiment Analysis in Uzbek", 2025 IEEE 26th International Conference of Young Professionals in Electron Devices and Materials (EDM), Altai, Russian Federation, pp. 2110-2114, 2025. doi: 10.1109/EDM65517.2025.11096704

[7] S. Ackerman, L. Alexander, M. Bennett, D. Chen, E. Farchi, A. Houseknecht, P. Santhanam, “Deploying automated ticket router across the enterprise”, AI Magazine, vol. 44, pp. 97-111, 2023. doi: 10.1002/aaai.12079

[8] D. Mengliev, V. Barakhnin, M. Eshkulov, B. Ibragimov, S.Madirimov, “A comprehensive dataset and neural network approach for named entity recognition in the Uzbek language”, Data in Brief, vol. 58, 1112459, 2025.

[9] Y. Yang and X. Liu, “A re-examination of text categorization methods”, 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '99), pp. 42–49, 1999. doi: 10.1145/312624.312647

[10] F. Sebastiani, “Machine Learning in Automated Text Categorization”, ACM Computing Surveys, vol. 34, No. 1, pp. 1–47, March 2002.

[11] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, “Bag of Tricks for Efficient Text Classification”, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pp. 427-431, 2017.

[12] T. Joachims, “Training linear SVMs in linear time”, In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '06), pp. 217–226, 2006. doi: 10.1145/1150402.1150429

[13] D. B. Mengliev, V. B. Barakhnin, B. R. Saidov, M. Atakhanov, M. O. Eshkulov and B. B. Ibragimov, "A Computational Approach to Recognizing Poetry Genres in Uzbek Texts", 2024 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), Novosibirsk, Russian Federation, pp. 319-322, 2024. doi: 10.1109/SIBIRCON63777.2024.10758540.

[14] M. Ribeiro, S. Singh and C. Guestrin, “High-Precision Model-Agnostic Explanations”, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, issue 1, pp. 1527-1535, 2018. doi: 10.1609/aaai.v32i1.11491

[15] S. Wang and C. Manning, “Baselines and Bigrams: Simple, Good Sentiment and Topic Classification”, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 90-94, 2012.

[16] G. Forman, “An Extensive Empirical Study of Feature Selection Metrics for Text Classification”, Journal of Machine Learning Research, vol. 3, pp. 1289-1305, 2003.

[17] D. Mengliev, D. Nabiyeva, A. Abdurakhmonov, K. Makhmudov, A. Nuritdinov and A. Otemisov, "Educational Text Analysis in Uzbek: Developing an NER Algorithm for Academic and Pedagogical Content", 2025 IEEE 26th International Conference of Young Professionals in Electron Devices and Materials (EDM), Altai, Russian Federation, pp. 2100-2103, 2025. doi: 10.1109/EDM65517.2025.11096868

[18] Q. Mamiraliyev, “The issue of genres in uzbek poetry of the independence period ”, 1st International Congress on Modern Science, Tashkent, May 10–12. 2022.

[19] E. Kuriyozov, S. Matlatipov, M. Alonso, C. Gomez-Rodriguez, “Construction and Evaluation of Sentiment Datasets for Low-Resource Languages: The Case of Uzbek ”, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 13212 LNAI, pp. 232–243, 2022.

[20] D. Mengliev, V. Barakhnin, A. Sultonboyev, B. Ibragimov, M. Eshkulov, R. Abdullayev, “Developing a dictionary-centric named entity recognition system for Karakalpak language”, AIP Conf. Proceedings, vol. 3244, 030044, 2024.

[21] M. Sharipov, J. Mattiev, J. Sobirov and 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.

[22] 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.

[23] B. Elov, M. Samatboyeva, “Identifying ner (named entity recognition) objects in uzbek language texts ”, Science and innovation international scientific journal, volume 2, issue 4, 2023.

[24] N. Abdurakhmonova, R. Shirinova, R. Sayfullayeva, D. Mengliev, B. Ibragimov, M. Ernazarova, “An annotated morphological dataset for Uzbek word forms: Towards rule-based and machine learning approaches”, Data in Brief, vol. 61, 111702, 2025.

[25] D. B. Mengliev, V. B. Barakhnin and B. 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. doi: 10.1109/EDM58354.2023.10225235.

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

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

[28] S. Matlatipov, H. Rahimboeva, J. Rajabov and E. Kuriyozov, “Uzbek Sentiment Analysis based on local Restaurant Reviews ”, The International Conference on Agglutinative Language Technologies as a challenge of Natural Language Processing (ALTNLP), 2022.

[29] A. Abdullayev, B. Ibragimov, A. Ubaydullayev, M. Abdurazzakova, N. Abdukadirova and S. Mustafakulov, "Development and Comparative Analysis of Algorithms for Detecting Dialect Words of the Uzbek Language", 2025 IEEE 26th International Conference of Young Professionals in Electron Devices and Materials (EDM), Altai, Russian Federation, pp. 2060-2064, 2025. doi: 10.1109/EDM65517.2025.11096762

[30] D. B. Mengliev, N. Z. Abdurakhmonova, H. Rahimov, N. Y. Zolotykh, A. A. Ubaydullayev and B. B. Ibragimov, "Automated Recognition of Named Entities and Dialect Standardization in Uzbek Legal Texts", 2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE), Novosibirsk, Russian Federation, pp. 1050-1053, 2024. doi: 10.1109/PIERE62470.2024.10804942

[31] S. Raxmatova and M. Kuzibayeva, “Generality and specificity of dialectics and its reflection in the morphology of the Uzbek language”, Economy and society, vol. 9, no. 88, 2021

[32] D. B. Mengliev, N. Z. Abdurakhmonova, R. K. Shirinova, M. F. Saparova, I. M. Azimov and B. B. Ibragimov, "Automated Detection of Allusions in Uzbek Language: A Computational Approach", 2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE), Novosibirsk, Russian Federation, pp. 1560-1564, 2024. doi: 10.1109/PIERE62470.2024.10804911.

[33] B. B. Ibragimov, A. D. Egamberganova, S. I. Khamraeva, D. A. Fattaxova, Z. Kasimova and D. K. Khudayberganova, "Advancing Oceanology Studies in Karakalpak: A Named Entity Recognition Algorithmic Framework", 2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE), Novosibirsk, Russian Federation, pp. 1590-1593, 2024. doi: 10.1109/PIERE62470.2024.10804978