data

MDPI

Data Descriptor

Dual-Source Synthetic Uzbek Corpora for Sentiment Analysis and NER with Controlled Emoji Signals

Bobur Saidov [1,*], Vladimir Barakhnin [1,2], Shohrux Madirimov [1,3], Umid Ibragimov [1], Shakhboz Meylikulov [4], Sultonbek Normamatov [5], Feruza Bahodirova [6], Javlonbek Matnazarov [7] and Zarnigor Fayzullaeva [8]

1 Faculty of Mechanics and Mathematics, Novosibirsk State University, 1 Pirogova str., Novosibirsk 630090, Russia; bar@ict.nsc.ru (V.B.); s.madirimov@g.nsu.ru (S.M.); u.ibragimov@g.nsu.ru (U.I.)
2 Federal Research Center for Information and Computational Technologies, Novosibirsk 630090, Russia
3 Tashkent Institute of Textile and Light Industry, Tashkent 100100, Uzbekistan
4 Department of Information Technology and Exact Sciences, Termez University of Economics and Service, 38-B, Ibn-Sino str., Termez 190100, Uzbekistan; shaxboz_meyliqulov@tues.uz
5 Department of Computer Linguistics and Digital Technologies, Faculty of Social and Humanitarian Sciences, Alisher Navo'i Tashkent State University of Uzbek Language and Literature, 103, Yusuf Xos Khojib Str., Tashkent 100013, Uzbekistan; normamatovsulton1@gmail.com
6 Department of Interfaculty Foreign Languages, Urgench State University, 14, Kh. Alimdjan str., Urgench 220100, Uzbekistan; bahodirovaferuza215@gmail.com
7 Department of Language and Literature, Mamun University, 2, Bol-xovuz str., Khiva 220901, Uzbekistan; javlon_m@mamunedu.uz
8 Department of Software Engineering, Tashkent University of Information Technologies Named After Muhammad al-Khwarizmi, Tashkent 100084, Uzbekistan; zarnigor18z02@gmail.com
* Correspondence: saidovboburbek9629@gmail.com

Abstract

This data descriptor presents two fully synthetic corpora for sentiment analysis and named entity recognition (NER) in Uzbek. The first corpus contains 12,000 hybrid synthetic sentences generated from templates with lexical randomization, automatic insertion of named entities (PER/ORG/LOC), lexicon-based polarity scoring, and a controlled emoji distribution. The second corpus includes 3000 "manual-style" sentences designed to resemble short, naturally structured messages. Although the manual-style subset was initially intended to be emoji-free, the released version includes a 39.6% emoji presence (sentences containing at least one emoji) to maintain comparability in emotional markers across corpora. Both corpora are released in CSV, XLSX, and JSONL formats and share a unified schema (id, text, sentiment, entities, entity_type, polarity_score, polarity_source, token_count, emojis, emoji_position, emoji_sentiment, conflict_flag, sentiment_from_polarity_score, split). The dataset is publicly available via Mendeley Data (DOI: 10.17632/y2d5pcyrzz.3).

Dataset: DOI: 10.17632/y2d5pcyrzz.3 (Mendeley Data, Version 3); https://data.mendeley.com/datasets/y2d5pcyrzz/3 (accessed on 24 January 2026).

Keywords: Uzbek language; sentiment analysis; named entity recognition; synthetic datasets; low-resource languages; NLP; emoji signals

Academic Editor: Francesco M. Donini

Received: 10 January 2026
Revised: 24 January 2026
Accepted: 29 January 2026
Published: 1 February 2026

1. Summary

We release two openly accessible, fully synthetic Uzbek corpora designed for (i) sentiment (polarity) classification and (ii) named entity recognition with three entity types (PER,

ORG, LOC). The first corpus ("hybrid") contains 12,000 template-based sentences produced via linguist-defined patterns with slot filling and lexical randomization; it includes automatic entity insertion, lexicon-driven polarity scoring, and a controlled emoji distribution to mimic emotional and stylistic variability. The second corpus ("manual-style") contains 3000 shorter messages intended to resemble naturally structured text; the final version includes a controlled 39.6% emoji presence to maintain comparability in emotional markers across corpora.

Both corpora are distributed in CSV, XLSX, and JSONL formats under a unified field schema (id, text, sentiment, entities, entity_type, polarity_score, polarity_source, token_count, emojis, emoji_position, emoji_sentiment, conflict_flag, sentiment_from_polarity_score, split), enabling straightforward integration into machine learning pipelines and reproducible benchmarking. The resource is intended as an ethical alternative to social-media-derived datasets that are often constrained by privacy and redistribution policies, while still supporting controlled experiments on low-resource Turkic languages. All data files and accompanying metadata are hosted on Mendeley Data (DOI: 10.17632/y2d5pcyrzz.3).

Recent Uzbek NLP studies have addressed named entity recognition for tone/sentiment-related analysis and proposed NER selection methods [1]. Public dataset and implementation efforts for Uzbek NER have also been reported in data papers [2]. Uzbek sentiment resources for low-resource settings have been constructed and evaluated in prior work [3]. Additional NER research includes hybrid approaches for historical Uzbek texts [4]. Supporting Uzbek NLP tools such as rule-based POS tagging have been developed to facilitate preprocessing and downstream tasks [5]. Studies on dialect recognition further highlight linguistic variability that may affect model robustness [6]. For baseline modeling, transformer-based architectures such as Bidirectional Encoder Representations from Transformers (BERTs) are widely used in language understanding tasks [7]. Uzbek is a Turkic language, which motivates continued dataset development for computational linguistics in this language family [8]. The release is aligned with FAIR-oriented data stewardship principles to support findability and reuse [9]. Emoji symbols and categories used in the corpora follow the Unicode Consortium's emoji specification [10]. Related Uzbek NLP applications include computational approaches to poetry genre recognition [11]. Tools for closely related regional languages, such as syntax checking in Karakalpak, further demonstrate the need for reusable linguistic resources [12]. Additional Uzbek resources include annotated morphological datasets supporting both rule-based and machine learning approaches [13]. Social-media-derived datasets (e.g., Twitter-based corpora) illustrate the value of public datasets while also motivating synthetic alternatives due to redistribution and privacy constraints [14,15]. A comprehensive Uzbek NER dataset and neural approach have also been published in the data literature [16]. Finally, synthetic dataset releases in other languages demonstrate the utility of controlled generation for reproducible benchmarking [17].

2. Data Description

The deposited resource contains two separate synthetic Uzbek corpora created for sentiment (polarity) classification and named entity recognition (NER). Both corpora share a unified field schema and file formats, but differ in generation strategy, entity density, emoji presence, and overall text style.

2.1. Repository Structure

The dataset repository is organized into two main directories: (i) a hybrid synthetic subset containing 12,000 sentences and (ii) a manual-style subset containing 3000 sentences.

Each subset is provided in both CSV, XLSX, and JSONL formats and follows the same column structure to ensure straightforward cross-dataset use.

2.2. General Data Structure

A single schema is used for both corpora. Each record includes a unique identifier, the Uzbek sentence text, the sentiment label, named entities (surface forms) and their types (PER/ORG/LOC) stored as parallel JSON lists, and auxiliary fields supporting reproducibility (polarity_score, polarity_source, token_count, and emojis, where the emoji list is empty [] if none are present). This unified structure is intended to make the corpora easy to integrate into machine learning pipelines and to support consistent benchmarking across subsets. The shared field schema is summarized in Table 1.

Table 1. Example of data structure (shared schema).

Field Name	Description
id	Unique identifier for each record
text	Synthetic Uzbek sentence
sentiment	Polarity class: positive/neutral/negative
entities	JSON list of detected named entities
entity_type	JSON list of entity types (PER/ORG/LOC)
polarity_score	Numeric polarity score derived from lexicon
polarity_source	Rule-based polarity assignment source
token_count	Number of tokens in the sentence
emojis	JSON list of emojis detected/inserted in the sentence; an empty list [] indicates no emoji. This field is present in both subsets.
emoji_position	Categorical string: none/start/ middle/after_word/end (“none” if no emoji)
emoji_sentiment	Sentiment class for emoji signal: none/positive/neutral/negative
conflict_flag	1 if text polarity and emoji_sentiment conflict, else 0
sentiment_from_polarity_score	Label mapped from polarity_score alone (before emoji tie-break)
split	Predefined split: train/val/test

2.3. Dataset A: Hybrid Synthetic Dataset (12,000 Records)

The hybrid subset consists of 12,000 template-based synthetic sentences with slot filling and random lexical substitution. Compared with the manual-style subset, it uses richer patterns and explicitly models both named entities and emojis. Key corpus-level characteristics are summarized in Table 2, including typical sentence length, entity density, emoji frequency, and sentiment distribution.

Table 2. Statistical parameters of Dataset A (hybrid, 12k).

Parameter	Value
Total records	12,000
Avg. sentence length	6–18 tokens
Entities per sentence	1–3
Emoji presence	30.0% of sentences
Sentiment distribution	≈43% positive, 26% neutral, 31% negative
Named entity types	PER, ORG, LOC
File formats	CSV, XLSX, JSONL
Encoding	UTF-8

2.4. Dataset B: Manual-Style Synthetic Dataset (3000 Records)

The manual-style subset contains 3000 shorter messages designed to resemble naturally structured text. Although originally planned as an emoji-free baseline, the released version includes a controlled 39.6% emoji presence to keep both subsets comparable in emotional markers; sentiment is balanced at approximately 47/23/30 (positive/neutral/negative). The main corpus-level characteristics are provided in Table 3.

Table 3. Statistical parameters of Dataset B (manual-style, 3k).

Parameter	Value
Total records	3000
Avg. sentence length	5–13 tokens
Entities per sentence	1–2
Emoji presence	39.6% of sentences
Sentiment distribution	≈47% positive, 23% neutral, 30% negative
Named entity types	PER, ORG, LOC
File formats	CSV, XLSX, JSONL
Encoding	UTF-8

To complement the corpus-level statistics in Tables 2 and 3, Figure 1 visualizes the approximate sentiment balance in Dataset A (hybrid) and Dataset B (manual-style).

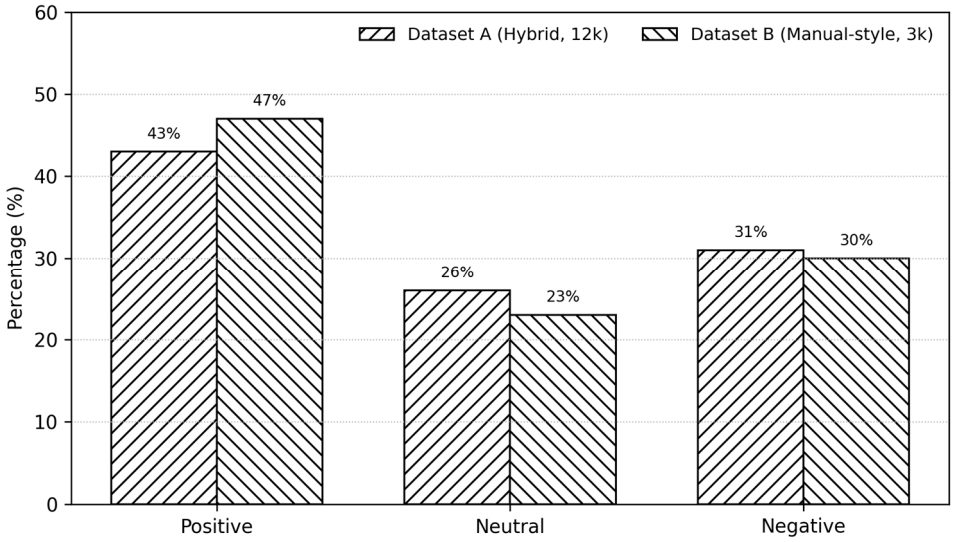

Figure 1. Sentiment distribution in Dataset A and Dataset B (percentages reported in Tables 2 and 3).

2.5. Comparison of the Two Datasets

To support dataset selection for different experimental goals, Table 4 summarizes key differences between the hybrid (12k) and manual-style (3k) subsets, including template variety, entity density, emoji usage, text style, and polarity-assignment approach.

Table 4. Dataset-level characteristics distinguishing the Hybrid and Manual-Style corpora.

Feature	Hybrid (12k)	Manual-Style (3k)
Templates used	High variety (9–12)	Low variety (4–6)
Emotions	Rich, including emojis	Emoji-supported; balanced (see Table 3)

Table 4. *Cont.*

Feature	Hybrid (12k)	Manual-Style (3k)
Emojis	Yes (30.0%)	Yes (39.6%)
Entity density	Up to 3 per sentence	Up to 2 per sentence
Text style	Synthetic but realistic	Human-like simple
Polarity assignment	Lexicon-based	Simplified, rule-based
Use cases	ML, DL, hybrid models	Baselines, linguistic tests

To improve interpretability and facilitate reuse, Table 5 provides representative annotated examples from both subsets, including an English translation and the corresponding JSON-style entity and emoji fields aligned with the released schema.

Table 5. Representative annotated sentence examples from Dataset A (hybrid) and Dataset B (manual-style), with English translations and JSON-formatted annotation fields. **Note**: Entities and entity_type are parallel lists; the *i*-th entity corresponds to the *i*-th type.

Subset	Uzbek Text	English Translation	Sentiment	Entities	Entity_Type	Emojis
Hybrid (12k)	Ali UzbekBank xizmatidan juda mamnun bo'ldi 👍😍	Ali was very satisfied with UzbekBank's service.	positive	["Ali", "UzbekBank"]	["PER", "ORG"]	["👍","😍"]
Hybrid (12k)	Samarqanddagi tadbir Navoi uchun muvaffaqiyatli bo'ldi 😳	The event in Samarkand was successful for Navoi.	positive	["Samarqand", "Navoi"]	["LOC", "PER"]	["😳"]
Hybrid (12k)	Botir Toshkentdagi uchrashuvdan norozi bo'ldi 👿	Botir was dissatisfied with the meeting in Tashkent.	negative	["Botir", "Toshkent"]	["PER", "LOC"]	["👿"]
Manual-style (3k)	Nukus bugun sokin, hammasi joyida.	Nukus is calm today; everything is fine.	neutral	["Nukus"]	["LOC"]	[]
Manual-style (3k)	Dilnoza TATU haqida yaxshi fikr aytdi 😳	Dilnoza said a good opinion about TATU.	positive	["Dilnoza", "TATU"]	["PER", "ORG"]	["😳"]
Manual-style (3k)	Andijon safarim juda charchatdi 😞	My trip to Andijon was very tiring.	negative	["Andijon"]	["LOC"]	["😞"]

These examples are representative; full corpora are available in the released CSV/XLSX files under the unified schema.

3. Methods

3.1. Overview of the Workflow

This study focuses on assessing the quality and reusability of two synthetic Uzbek corpora (12,000 hybrid sentences and 3000 manual-style sentences) for sentiment analysis and named entity recognition (NER).

The overall workflow consisted of the following: (i) preprocessing and normalization; (ii) automatic annotation of sentiment, named entities, and emojis; (iii) quality control and de-duplication; (iv) harmonization of all fields into a unified tabular schema; and (v) descriptive statistical analysis of corpus properties (e.g., sentence length and entity density).

Figure 2 summarizes the end-to-end workflow used to generate, annotate, and validate the dual-source corpora, and it maps each stage to the corresponding Methods subsections.

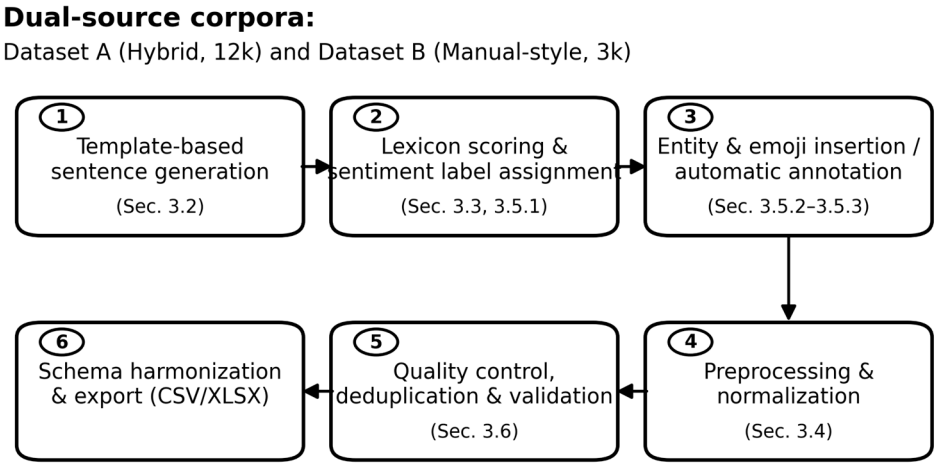

Figure 2. Workflow for generating and validating the dual-source synthetic Uzbek corpora.

3.2. Template- and Rule-Based Sentence Generation

Synthetic sentences were generated using reusable, linguist-designed templates that combine a fixed syntactic skeleton with typed slot placeholders (e.g., {PER}, {ORG}, {LOC}, {ACTION/VERB}, {OBJECT}, {ATTRIBUTE}). Templates were designed to reflect common short-message- and news-like Uzbek constructions and to cover three sentiment polarities (positive/negative/neutral) through sentiment-bearing predicates and modifiers.

To reduce ungrammatical combinations, compatibility constraints were applied (e.g., person–role phrases, organization–activity verbs, location–event verbs). For each slot, curated dictionaries were compiled: gazetteers for PER/ORG/LOC, sentiment lexicons for evaluative words, and neutral vocabularies to increase topical diversity. Sentence generation scripts randomly sampled slot values under fixed seeds for reproducibility, optionally inserted emoji tokens using emoji-to-sentiment rules, and applied post-generation filters including duplicate removal and rule-based validity checks.

The final release includes (i) a larger "hybrid" subset with broader template coverage and controlled emoji signals and (ii) a smaller "manual-style" subset using fewer and shorter templates to resemble concise human messages.

Since Uzbek is an agglutinative language, template patterns were designed to accommodate common noun-case suffixes and basic agreement endings to preserve minimal grammatical coherence in synthetic outputs. In particular, we implemented lightweight rules for frequent case realizations (e.g., '-ga', '-da', '-dan', and object marking '-ni') and simple possessive/attributive constructions (e.g., {PER}ning + noun). When a slot required an inflected surface form, the generator applied the corresponding suffix using a small rule set, reducing ungrammatical combinations while keeping the pipeline lightweight and reproducible.

3.3. Lexicon-Based Sentiment Scoring and Label Assignment

For each generated sentence, a rule-based lexicon procedure assigns an intermediate polarity signal stored in polarity_score. In the released dataset, polarity_score is a discrete integer in $\{-1, 0, +1\}$, representing negative, neutral, and positive polarity evidence derived from sentiment-bearing tokens and rules (e.g., negation handling and polarity flips).

Aggregation rule (token-to-sentence): To compute the intermediate score in a fully reproducible manner, we first assign each token a base polarity value from the lexicon ($p_i \in \{-1, 0, +1\}$ after normalization). We then apply simple composition rules for local modifiers (e.g., negation flips the sign of the nearest sentiment-bearing token within a small

window; intensifiers increase the magnitude of that token-level signal). The sentence-level raw score is computed by aggregating token signals:

$$s = \frac{1}{\max(1, m)} \sum_{i=1}^{n} p_i$$

where m is the number of non-zero polarity tokens (to avoid diluting the signal in longer neutral sentences). Finally, we discretize s into `polarity_score` $\in \{-1, 0, +1\}$ using a zero-centered threshold (negative if $s < 0$, positive if $s > 0$, neutral otherwise), and map it to the categorical label in sentiment. If emoji cues are present, they are treated as an additional polarity signal and can reinforce or override weak textual cues, consistent with the tie-breaking rules described in the sentiment annotation procedure.

The final categorical label in `sentiment` (positive/neutral/negative) is obtained by mapping this polarity evidence to a class label and applying tie-breaking logic when multiple cues conflict. In particular, emoji cues may reinforce or override weak textual cues in ambiguous cases, consistent with the controlled emoji setting used during generation. The field `polarity_source` records the labeling configuration/rule set used for generating that record.

3.4. Preprocessing and Normalization

Both corpora underwent the same preprocessing steps: case normalization, Uzbek-aware tokenization (including apostrophe handling), and removal of extraneous characters. We also validated the JSON-formatted fields (entities, entity_type, and emojis) and ensured consistent JSON serialization across records.

In addition, we computed auxiliary statistics (e.g., token_count, entity counts, and emoji presence) and removed duplicate and near-duplicate sentences to improve corpus uniqueness.

3.5. Automatic Annotation of Sentiment, Named Entities, and Emojis

3.5.1. Sentiment Annotation

Aggregation rule: We compute a lexicon-based `polarity_score` by summing token-level sentiment weights and mapping the result to $\{-1, 0, +1\}$. In parallel, we compute an emoji_score by summing emoji sentiment weights and mapping the result to $\{-1, 0, +1\}$. The final label is assigned as follows:

(i) if `polarity_score` = 0 and `emoji_score` $\neq 0 \rightarrow$ use `emoji_score`;
(ii) if `emoji_score` = 0 and `polarity_score` $\neq 0 \rightarrow$ use `polarity_score`;
(iii) if `polarity_score` and `emoji_score` have the same sign $\rightarrow$ use that sign;
(iv) if they conflict (opposite signs) $\rightarrow$ assign negative as a conservative tie-breaker.

3.5.2. Named Entity (NER) Annotation

Named entities were inserted and recorded automatically using verified gazetteers for three categories: persons (PER), organizations (ORG), and locations (LOC).

To ensure coverage, the resource includes more than 15,000 Uzbek personal names, over 700 geographic names, and about 500 organization names. Entities were stored in standardized fields as parallel JSON arrays: entities (surface forms) and `entity_type` (PER/ORG/LOC).

3.5.3. Emoji Annotation and Usage

Emojis were treated as an independent emotional signal. For each sentence, emoji tokens were automatically identified and classified into three categories (positive/negative/neutral), and their contribution was considered when text and emoji signals conflicted (priority to the stronger emotional cue).

Emojis appeared in 30.0% of hybrid sentences and 39.6% of manual-style sentences, where emoji frequency was intentionally increased to maintain comparability between the two corpora.

3.6. Quality Control and Reproducibility

Quality control was performed via automated validation on the full v3 release. Across both subsets, JSON parsing success was 100% for the JSON-formatted fields (entities, entity_type, and emojis). Schema alignment between entities and entity_type lists was 100% (equal list lengths for all records). Emoji presence was consistent with emoji_position (emoji_position = "none" iff emojis = []). In addition, duplicate and near-duplicate sentences were removed, and field-level constraints (allowed values and required columns) were checked using a public validation script provided with the dataset.

The identified schema-alignment and emoji-field inconsistencies were corrected prior to the final release.

3.7. Technical Validation (Downstream Utility)

To provide an indicative validation that the released corpora are learnable, we provide reproducible baseline scripts and recommend standard checks (e.g., TF-IDF + linear classifier for sentiment, and a CRF/transformer tagger for NER). Because the corpora are fully synthetic, these checks are reported as technical validation rather than real-world benchmarks.

4. User Notes

The dataset is provided in both CSV, XLSX, and JSONL formats. A complete data dictionary and field constraints are provided in Appendix A (Table A1). Each row corresponds to one synthetic Uzbek sentence and includes sentiment labels and NER annotations. The fields entities, entity_type, and emojis are stored as JSON lists; when reading the CSV, these fields should be parsed as arrays/objects (e.g., using json.loads() in Python 3.13.9). An example Python snippet for parsing JSON-formatted columns is provided in Algorithm 1. The sentiment field contains three classes (positive/neutral/negative). For NER, the dataset uses three entity types, PER, ORG, and LOC, stored in parallel arrays (entities[i] corresponds to entity_type[i]).

For practical use, we recommend the following:

1. Train/validation/test split with stratification by sentiment and (optionally) by the presence of entities and emojis to preserve class balance.
2. Tokenization using Uzbek-aware rules (apostrophes and Latin/Cyrillic variants, if applicable).
3. NER conversion: If a sequence-labeling format is required (BIO/BILOU), map each entity surface form to token entity strings in text and generate token-level tags; when multiple identical entity mentions appear, disambiguate by span matching.
4. Limitations: The corpora are fully synthetic and therefore may not capture all distributional properties of real social media or news texts; the results obtained on these corpora should be interpreted as controlled benchmarks rather than direct estimates of real-world performance.
5. Recommended baseline checks: (i) A simple sentiment baseline (e.g., TF-IDF + linear classifier) and (ii) a standard NER baseline (e.g., CRF or a transformer-based tagger) can be used to confirm that labels and entity fields are consistent with the expected learning signal.

Usage notes (parsing JSON-formatted fields). In the released CSV/XLSX files, the fields `entities`, `entity_type`, and `emojis` are stored as JSON-formatted strings (e.g.,

["Toshkent","Ali"]). For programmatic reuse, these columns should be parsed into Python lists. The snippet below demonstrates a minimal, reproducible loading procedure in Python.

Algorithm 1. Parsing JSON-formatted columns in the CSV/XLSX release (Python example).

```
import json
import pandas as pd
# Load the dataset (CSV example)
df = pd.read_csv(''dataset.csv'')
# Columns stored as JSON strings
json_cols = [''entities'', ''entity_type'', ''emojis'']
# Parse JSON safely (handle empty/missing values)
for c in json_cols:
df[c] = df[c].fillna(''[]'').apply(json.loads)
# Optional sanity check:  entities and entity_type must be aligned
assert (df[''entities''].str.len() == df[''entity_type''].str.len()).all()
print(df.head(3))
```

Note: The lists in `entities` and `entity_type` are parallel; the *i*-th entity corresponds to the *i*-th entity type.

Author Contributions: Conceptualization, B.S. and V.B.; Methodology, B.S., V.B., and S.M. (Shohrux Madirimov); Software, S.M. (Shohrux Madirimov) and U.I.; Data curation, U.I., S.N., F.B., and J.M.; Resources (gazetteers/lexicons and annotation materials), S.M. (Shakhboz Meylikulov), S.N., F.B., J.M., and Z.F.; Validation, B.S., S.M. (Shohrux Madirimov), and Z.F.; Formal analysis, B.S. and V.B.; Visualization, S.M. (Shohrux Madirimov) and B.S.; Writing—original draft, B.S.; Writing—review and editing, B.S., V.B., S.M. (Shohrux Madirimov), U.I., S.M. (Shakhboz Meylikulov), S.N., F.B., J.M., and Z.F.; Supervision, V.B.; Project administration, B.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by the state assignment of the Ministry of Science and Higher Education of the Russian Federation for the Federal Research Center for Information and Computational Technologies. The funding number is not applicable/was not provided for this state assignment. The APC was self-funded by the authors.

Data Availability Statement: The dataset generated and analyzed during the current study is publicly available on Mendeley Data at DOI: 10.17632/y2d5pcyrzz.3 (Version 3), https://data.mendeley.com/datasets/y2d5pcyrzz/3, accessed on 26 January 2026. The dataset is released under the CC BY 4.0 license.

Acknowledgments: The research was partially supported by the state assignment of Ministry of Science and Higher Education of the Russian Federation for Federal Research Center for Information and Computational Technologies. AI assistance was used for language editing. During the preparation of this manuscript, the authors used ChatGPT (OpenAI) only for grammar and spelling checks. All scientific content, analyses, and conclusions were produced and verified by the authors, who take full responsibility for the publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

APC	Article Processing Charge
FAIR	Findable Accessible Interoperable Reusable
NLP	Natural Language Processing

NER	Named Entity Recognition
PER	Person
ORG	Organization
LOC	Location
JSON	JavaScript Object Notation

Appendix A

Appendix A.1. Full Data Dictionary and Field Constraints

Table A1. Full data dictionary for the released XLSX/CSV files (both subsets).

Field	Type	Allowed Values/Format	Description	Example
`id`	integer	1…N	Unique record identifier	4
`text`	string	UTF-8	Synthetic Uzbek sentence	`Ali ...` 👍😍
`sentiment`	categorical string	`positive, neutral, negative`	Final sentiment label (may incorporate emoji/tie-break rules)	`positive`
`entities`	JSON array (string in CSV/XLSX)	`["<ent1>", "<ent2>", ...]`	Surface forms of inserted named entities in the sentence	`["Ali","UzbekBank", "Samarqand"]`
`entity_type`	JSON array (string in CSV/XLSX)	`["PER","ORG","LOC"]` (parallel to `entities`)	Entity type for each element in entities (same index)	`["PER","ORG","LOC"]`
`polarity_score`	integer	$-1, 0, 1$	Lexicon-derived polarity score (intermediate numeric signal)	1
`polarity_source`	categorical string	manual subset: manual_style_v3; hybrid subset: synthetic_lexicon_v3	Which polarity/labeling configuration produced the record	synthetic_lexicon_v3
`token_count`	integer	≥ 1	Token count after preprocessing/tokenization described in Methods	15
`emojis`	JSON array (string in CSV/XLSX)	[] or [“😳”, ...]	Emoji list detected/inserted in the sentence (empty list if none)	[“👍”,”😍”]
`emoji_position`	categorical string	none, start, middle, after_word, end	Emoji position category (“none” if emojis = [])	none
`emoji_sentiment`	categorical string	none, positive, neutral, negative	Emoji sentiment signal derived from emoji list	positive
`conflict_flag`	integer	0 or 1	1 if polarity_score and emoji_sentiment conflict; else 0	0
`sentiment_from_polarity_score`	categorical string	positive, neutral, negative	Sentiment label derived only from polarity_score (before emoji tie-break)	neutral
`split`	categorical string	train, val, test	Predefined split for reproducible experiments	train

Appendix A.2. Minimal Parsing Note (Recommended)

When reading the CSV (or XLSX), parse JSON-list fields (`entities`, `entity_type`, `emojis`) as arrays. In Python, for example: `json.loads(row["entities"])`. Ensure that `len(entities) == len(entity_type)` and interpret each `(entities[i],` entity_type[i]) as one annotation pair.

References

1. Saidov, B.R.; Barakhnin, V.B.; Rixsibayev, U.T.; Sharipov, E.J.; Sobirov, O.O.; Bekchanov, M.K. Methods of Automatic Selection of Named Entities (NER) in Uzbek Language for Text Tone Analysis. In Proceedings of the 2025 IEEE 26th International Conference of Young Professionals in Electron Devices and Materials (EDM), Altai Republic, Russia, 27 June–1 July 2025; IEEE: Piscataway, NJ, USA, 2025. [CrossRef]
2. Mengliev, D.; Barakhnin, V.; Abdurakhmonova, N.; Eshkulov, M. Developing named entity recognition algorithms for Uzbek: Dataset insights and implementation. *Data Brief* **2024**, *54*, 110413. [CrossRef] [PubMed]

3. Kuriyozov, E.; Matlatipov, S.; Alonso, M.A.; Gómez-Rodríguez, C. Construction and Evaluation of Sentiment Datasets for Low-Resource Languages: The Case of Uzbek. In *Human Language Technology. Challenges for Computer Science and Linguistics (LTC 2019)*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2022; Volume 13212, pp. 232–243. [CrossRef]
4. Mengliev, D.; Abdurakhmonova, N.; Allamov, O.; Ibragimov, B.; Saidov, B.; Boltayev, N. Development of a Hybrid Algorithm for Identifying Named Entities in 20th Century Uzbek Texts. In Proceedings of the AIP Conference Proceedings, Wollongong, Australia, 1–5 December 2025; Volume 3356, p. 050002. [CrossRef]
5. Sharipov, M.; Kuriyozov, E.; Yuldashev, O.; Sobirov, O. UzbekTagger: The Rule-Based POS Tagger for the Uzbek Language. *arXiv* **2023**, arXiv:2301.12711. [CrossRef]
6. Mengliev, D.; Barakhnin, V.; Madirimov, S.; Ibragimov, B.; Eshkulov, M.; Saidov, B. Unveiling the Variance of Uzbek Language: A Rule-Based Algorithm for Dialect Recognition. In Proceedings of the AIP Conference Proceedings, Melbourne, Australia, 22–23 May 2024; Volume 3244, p. 030012. [CrossRef]
7. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [CrossRef]
8. Johanson, L.; Csató, É.Á. *The Turkic Languages*; Routledge: London, UK, 2015. [CrossRef]
9. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef] [PubMed]
10. Unicode Consortium. Full Emoji List, v17.0. 2025. Available online: https://unicode.org/emoji/charts/full-emoji-list.html (accessed on 8 January 2026).
11. Mengliev, D.; Barakhnin, V.B.; Saidov, B.R.; Ibragimov, B.B. A Computational Approach to Recognizing Poetry Genres in Uzbek Texts. In Proceedings of the 2024 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), Novosibirsk, Russia, 30 September–2 October 2024; IEEE: Piscataway, NJ, USA, 2024. [CrossRef]
12. Saidov, B.R.; Barakhnin, V.B.; Sharipov, E.J.; Maksetbaev, A.B.; Ruzimov, J.O.; Abdullayev, R.M. Development and Realization of Software Application for Syntax Checking of Karakalpak Language Text. In Proceedings of the 2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE), Novosibirsk, Russia, 15–17 November 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1580–1588. [CrossRef]
13. Abdurakhmonova, N.; Shirinova, R.; Sayfullayeva, R.; Mengliev, D.; Ibragimov, B.; Ernazarova, M. An annotated morphological dataset for Uzbek word forms: Towards rule-based and machine learning approaches. *Data Brief* **2025**, *61*, 111702. [CrossRef] [PubMed]
14. Nia, Z.M.; Bragazzi, N.L.; Wu, J.; Kong, J.D. A Twitter dataset for Monkeypox, May 2022. *Data Brief* **2023**, *48*, 109118. [CrossRef] [PubMed]
15. de Melo, T.; Figueiredo, C.M.S. A first public dataset from Brazilian Twitter and news on COVID-19 in Portuguese. *Data Brief* **2020**, *32*, 106179. [CrossRef] [PubMed]
16. Mengliev, D.; Barakhnin, V.; Eshkulov, M.; Ibragimov, B.; Madirimov, S. A comprehensive dataset and neural network approach for named entity recognition in the Uzbek language. *Data Brief* **2024**, *58*, 111249. [CrossRef] [PubMed]
17. Al-Shameri, N.; Al-Khalifa, H. Arabic paraphrased parallel synthetic dataset. *Data Brief* **2024**, *57*, 111004. [CrossRef] [PubMed]