Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An annotated Uzbek social media corpus for emotion and sentiment analysis

Botir Elov

elov@navoiy-uni.uz

Alisher Navo'i Tashkent State University of Uzbek Language and Literature https://orcid.org/0000-0001-5032-6648

Alayev Ruhillo

National University of Uzbekistan named after Mirzo Ulugbek. https://orcid.org/0000-0003-3757-7711

Zilola Xusainova

Alisher Navoi Tashkent State University of Uzbek Language and Literature https://orcid.org/0000-0003-4357-7515

Data Note

Keywords: Uzbek NLP, emotion detection, sentiment analysis, social media corpus, low-resource language, affective computing, sarcasm detection, domain adaptation

Posted Date: May 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9595755/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

UzEDSA is an annotated corpus for emotion detection and sentiment analysis in Uzbek social media comments. It contains 304,364 unique comments after batch merging, invalid-record removal, and duplicate filtering. Each comment is labelled with a multi-attribute schema covering primary emotion, multi-emotion annotation, sentiment polarity, emotional intensity, sarcasm, and domain. The emotion inventory includes JOY, SAD, ANG, FEA, SUR, TRU, DIS, ANT, MIX, and NEU. The corpus preserves social media features such as spelling variation, Uzbek Latin-Cyrillic script mixing, Uzbek-Russian code-switching, emojis, slang, repeated characters, markup artifacts, short reactions, and long comments. It covers general, personal, news, technology, medical, education, transport, service, and marketing domains. This Data Descriptor documents corpus construction, preprocessing, annotation, deduplication, validation, repository records, and usage notes. UzEDSA supports reusable research on low-resource Uzbek affective NLP, including emotion classification, polarity classification, intensity prediction, sarcasm-aware modelling, domain adaptation, and corpus-based analysis.

Background & Summary

Social media comments provide direct evidence of how users express opinions, emotional reactions, service complaints, trust, fear, hope, and disappointment in everyday digital communication. Unlike formal documents, these comments are typically short, informal, noisy, and highly contextual. They include spelling variation, emojis, repeated characters, code-switching, mixed scripts, slang, and abbreviated forms. These properties make social media comments useful for affective natural language processing, while also making automatic emotion and sentiment analysis difficult [1–3].

Sentiment analysis and emotion detection are related but distinct tasks. Polarity labels identify whether a text is positive, negative, neutral, or mixed, whereas emotion labels distinguish more specific affective states such as joy, sadness, anger, fear, surprise, trust, disgust, and anticipation. This distinction is useful because negative comments may express very different meanings, such as anger toward a service provider, fear about uncertainty, or sadness about personal loss. The UzEDSA annotation scheme therefore records both polarity and emotion categories.

The design of the label set is informed by categorical and dimensional theories of emotion. Basic-emotion approaches motivate discrete categories such as joy, sadness, anger, fear, surprise, and disgust [4]. Dimensional approaches highlight valence and intensity [5], while psychoevolutionary approaches emphasize relationships among emotions and mixed affective states [6]. UzEDSA operationalizes these ideas with ten labels: JOY, SAD, ANG, FEA, SUR, TRU, DIS, ANT, MIX, and NEU.

Several affective corpora have supported progress in emotion-aware NLP, including SemEval affective tasks and GoEmotions [7–10]. However, high-quality emotion datasets remain concentrated in high-resource languages. Uzbek is a low-resource language with Latin-Cyrillic script variation, rich morphology, and frequent Uzbek-Russian code-mixing in online discourse. Existing Uzbek resources such as UzBERT, BERTbek, and UzABSA provide important models or sentiment-related datasets, but

they do not describe a large, multi-domain, fine-grained emotion and sentiment corpus for noisy Uzbek social media comments [15–17].

UzEDSA addresses this data-resource gap by providing 304,364 unique annotated Uzbek social media comments. Each record contains a comment and six reusable labels: Emotion_primary, Emotion_multi, Polarity, Intensity, Sarcasm, and Domain. The anonymized and deduplicated dataset is available through Hugging Face and Mendeley Data [30, 31]. The present article documents how the dataset was constructed, what it contains, how it was validated, and how it can be reused.

Table 1
Positioning of UzEDSA among related affective NLP resources.

Resource	Language	Data source	Main focus	Relevance
SemEval-2007 Affective Text	English	News headlines	Emotion and valence annotation	No Uzbek social media data
SemEval-2018 Affect in Tweets	English, Arabic, Spanish	Tweets	Emotion intensity, valence, emotion classification	No Uzbek data
GoEmotions	English	Reddit comments	Fine-grained emotion annotation	English-only
UzBERT	Uzbek	General Uzbek text	Pretrained Uzbek BERT model	Model resource, not an emotion corpus
BERTbek	Uzbek	Uzbek text	Pretrained Uzbek language model	Model resource, not an affective dataset
UzABSA	Uzbek	Restaurant reviews	Aspect-based sentiment analysis	Domain-specific sentiment
UzEDSA	Uzbek	Social media comments	Emotion, polarity, intensity, sarcasm, domain	Dataset described in this article

Methods

Input data and scope

The input data consisted of Uzbek-language and Uzbek-centric user-generated comments collected from publicly visible online discussions. The comments cover telecommunication services, mobile internet quality, customer support, mobile applications, public events, entertainment, politics, education, healthcare, transport, finance-related discussions, family issues, and general everyday reactions.

The corpus intentionally preserves linguistic characteristics of Uzbek online communication, including Uzbek Latin script, Uzbek Cyrillic script, Russian Cyrillic text, Uzbek-Russian code-mixing, occasional

English expressions, informal transliteration, emojis, repeated letters, spelling errors, colloquial expressions, abbreviations, punctuation-based emphasis, and markup artifacts. These features were treated as part of the social media signal rather than as noise to be fully removed.

Inclusion and exclusion criteria

Table 2
Inclusion and exclusion criteria applied during corpus construction.

Category	Criterion	Decision
Uzbek or Uzbek-centric content	Written fully or mostly in Uzbek or related to Uzbek users, services, media, or public discourse	Included
Code-mixed online discourse	Uzbek mixed with Russian, English, emojis, or transliteration	Included
Informal social media writing	Spelling errors, slang, repeated letters, emojis, punctuation emphasis	Included
Empty or uninterpretable text	Blank rows, whitespace-only comments, or meaningless symbols	Excluded
Exact duplicates	Same normalized comment already present in the merged corpus	Excluded
Severe ambiguity	Insufficient information to infer any label reliably	Reviewed or excluded

Preprocessing and deduplication

Preprocessing was conservative. Empty rows, broken fragments, repeated whitespace, and non-informative artifacts were removed. Emojis, mixed scripts, informal spelling, repeated characters, and punctuation-based affective cues were retained unless they interfered with readability or annotation.

Thirteen annotated batches were merged into a single candidate corpus. The merged raw candidate set contained 321,419 records. After removing 50 invalid or empty records, 321,369 valid records entered deduplication. Exact and repeated duplicate comments were removed, resulting in 304,364 unique annotated records. The consolidation flow is shown in Fig. 2.

Annotation schema

Each comment was annotated using a compact multi-attribute EDSA schema. The main target label for single-label emotion classification is Emotion_primary. The Emotion_multi field supports multi-label emotion detection. Polarity supports sentiment classification, Intensity supports ordinal affective strength prediction, Sarcasm supports sarcasm-aware modelling, and Domain supports domain-aware analysis. The annotation framework is summarized in Fig. 3.

Table 3
Dataset schema used in UzEDSA.

Field	Type	Description
Text_uz	String	Original or minimally cleaned Uzbek social media comment
Emotion_primary	Categorical	Dominant emotion label
Emotion_multi	Multi-label string	Primary emotion plus optional secondary emotions separated by semicolon
Polarity	Categorical	Overall sentiment orientation: POS, NEG, NEU, or MIX
Intensity	Ordinal integer	Emotional strength: 0, 1, or 2
Sarcasm	Binary integer	0 = absent, 1 = present
Domain	Categorical	Thematic domain of the comment

Table 4
Emotion tagset and default polarity mapping.

Code	Meaning	Default polarity
JOY	Joy or satisfaction	POS
SAD	Sadness or disappointment	NEG
ANG	Anger, irritation, or complaint	NEG
FEA	Fear, worry, or uncertainty	NEG
SUR	Surprise or astonishment	Context-dependent
TRU	Trust or confidence	POS
DIS	Disgust, rejection, or contempt	NEG
ANT	Anticipation, hope, or expectation	POS
MIX	Mixed positive and negative affect	MIX
NEU	Factual or emotionally unmarked text	NEU

Annotation decision rules

The dominant emotion was assigned as Emotion_primary. If only one emotion was present, Emotion_multi was set equal to Emotion_primary. If multiple emotions were expressed, secondary emotions were added to Emotion_multi using semicolon-separated labels. Sarcasm was marked only when surface wording and intended meaning clearly diverged. Factual information was marked as NEU with NEU polarity and intensity 0. A representative set of decision rules and examples is provided in Supplementary Fig. S3.

Ethical and privacy considerations

The study uses publicly visible user-generated comments for aggregate-level computational analysis. No direct interaction with users, intervention, recruitment, interview, or survey was conducted. Raw source traces, usernames, account identifiers, private profile links, phone numbers, and other personally identifiable information are not released. The dataset is intended for aggregate-level research and must not be used for individual profiling, surveillance, harmful targeting, or sensitive personal inference.

Data Records

The anonymized and deduplicated UzEDSA dataset is publicly available from two repositories: Hugging Face and Mendeley Data [30, 31]. The Hugging Face record supports direct use in machine-learning workflows, while the Mendeley Data record provides a versioned and citable research-data archive. The dataset is distributed under a CC BY 4.0 license.

Table 5
Public release and documentation structure.

Component	Purpose	Contents
Hugging Face Dataset	Direct NLP reuse and dataset viewer access	DOI: 10.57967/hf/8617; CSV format; 304k rows
Mendeley Data	Citable archival repository record	Version 1; DOI: 10.17632/jxm9ynjg66.1
Main data file	Annotated corpus	Text_uz and six annotation fields
Tagset documentation	Label definitions	Emotion, polarity, intensity, sarcasm, and domain definitions
Correction log sample	Annotation harmonization documentation	Representative corrected values and reasons
Statistics summary	Corpus overview	Counts for labels, domains, and validation checks

The core data file contains one row per comment and the fields defined in Table 3. Users should cite the repository version they use. The dataset records in external repositories should be treated as the authoritative access points. Data citations are included in the reference list as refs. 30 and 31.

Data Overview

This optional overview provides compact descriptive information to help readers understand the scale and composition of the dataset. Users should download the repository files for full exploration and replication. The summary in Fig. 4 shows the primary label, polarity, intensity, and domain distributions.

Table 6
Compact corpus overview.

Item	Value
Final unique records	304,364
Valid non-empty records before deduplication	321,369
Invalid or empty records removed	50
Duplicate records removed	17,005
Deduplication rate among valid records	5.29%
Largest emotion classes	JOY: 104,783; NEU: 98,382
Sarcasm-positive records	1,861 (0.61%)
Largest domain	general: 198,438 (65.20%)

Technical Validation

Technical validation was performed to ensure that the released corpus is structurally valid, internally consistent, annotation-aware, and suitable for downstream affective NLP experiments. Validation was not limited to spreadsheet completeness; it also examined whether relationships among Emotion_primary, Emotion_multi, Polarity, Intensity, Sarcasm, and Domain were semantically plausible.

Table 7
Rule-based validation checks.

Check	Rule
Emotion validity	Emotion_primary must be one of the ten official labels
Multi-emotion validity	Emotion_multi must contain valid labels separated by semicolons
Polarity validity	Polarity must be POS, NEG, NEU, or MIX
Intensity validity	Intensity must be 0, 1, or 2
Sarcasm validity	Sarcasm must be 0 or 1
Domain validity	Domain must be one of the nine official domain categories
Emotion-polarity consistency	JOY/TRU/ANT -> POS; SAD/ANG/FEA/DIS -> NEG; NEU -> NEU; MIX -> MIX
Neutral consistency	NEU comments normally have intensity 0
Duplicate control	Exact duplicates were removed

Field-level corrections were recorded in a correction log. The largest correction categories were Emotion_multi (72,724), Emotion_primary (58,212), Intensity (55,658), Polarity (40,995), Domain (27,179),

and Sarcasm (2,804). These counts are field-level corrections, so a row could contribute more than one correction. The validation procedure supports transparency but does not replace future inter-annotator agreement reporting. For studies that reuse or extend the corpus, a double-annotated subset should be assessed using metrics such as Cohen's kappa, Krippendorff's alpha, Jaccard overlap, or task-specific F1 [21–23].

The corpus is naturally imbalanced, which is typical of social media data. Evaluation protocols should therefore report macro-F1 and per-class metrics rather than accuracy alone [24, 25]. Sarcasm is especially imbalanced, with 1,861 sarcasm-positive records and 302,503 non-sarcastic records. Sarcasm detection should be evaluated with positive-class F1, precision, recall, and area under the precision-recall curve.

Usage Notes

UzEDSA can be used for primary emotion classification, multi-label emotion detection, polarity classification, intensity prediction, sarcasm detection, domain classification, domain adaptation, and multi-task affective modelling. Raw text should be treated as the primary benchmark setting because it reflects real social media conditions. Normalized versions may be used as ablation experiments, but preprocessing choices should be reported.

Table 8
Recommended reuse tasks and evaluation metrics.

Task	Input	Target	Recommended primary metric
Primary emotion classification	Text_uz	Emotion_primary	Macro-F1
Multi-label emotion detection	Text_uz	Emotion_multi	Micro-F1 and macro-F1
Sentiment polarity classification	Text_uz	Polarity	Macro-F1
Intensity prediction	Text_uz	Intensity	Macro-F1 or weighted kappa
Sarcasm detection	Text_uz	Sarcasm	Positive-class F1
Domain classification	Text_uz	Domain	Macro-F1
Domain adaptation	Text_uz + domain	Target-domain labels	Cross-domain macro-F1

Researchers should report the dataset DOI, repository version, access date, preprocessing strategy, split construction, model checkpoint, hyperparameters, random seed, metrics, and per-class results. Duplicate and near-duplicate comments should not appear across train, development, and test partitions. Supplementary Fig. S5 provides a stratified split template; Supplementary Figs. S6-S9 provide reporting

templates for per-class F1, confusion matrices, domain-transfer analysis, and preprocessing robustness. These templates do not contain empirical model results and should be populated only after controlled experiments on fixed splits.

Declarations

Data Availability

The anonymized and deduplicated UzEDSA dataset is publicly available through two repositories. The machine-learning-oriented release is available on Hugging Face Dataset under DOI 10.57967/hf/8617 [30]. A citable archival release is available on Mendeley Data, Version 1, under DOI 10.17632/jxm9ynjg66.1 [31]. The released dataset contains 304,364 unique annotated Uzbek social media comments with the fields Text_uz, Emotion_primary, Emotion_multi, Polarity, Intensity, Sarcasm, and Domain. The dataset is distributed under the CC BY 4.0 license. Raw source traces, usernames, account identifiers, platform metadata, private profile links, phone numbers, and other personally identifiable information are not released.

Code Availability

The dataset itself is publicly available through Hugging Face and Mendeley Data [30,31]. Scripts used for preprocessing, deduplication, label validation, statistics, train/dev/test splitting, and baseline modelling will be released in a versioned public repository after final code review. Until the public code repository is released, the rule-based validation logic and recommended experimental protocols are described in the Methods, Technical Validation, and Usage Notes sections.

Acknowledgements

The authors thank the annotators and reviewers who contributed to the construction, verification, and refinement of the UzEDSA corpus and acknowledge the Uzbek NLP research community for its contributions to low-resource language technologies.

Author Contributions

Conceptualization: Botir Elov and Ruhillo Alaev. Data collection and annotation guideline design: Botir Elov, Ruhillo Alaev, and annotator team. Data annotation and validation: annotator team. Data curation and statistical analysis: Botir Elov. Manuscript drafting: Botir Elov. Review and editing: all authors. Supervision: Botir Elov. All authors reviewed and approved the manuscript.

Competing Interests

The authors declare no competing interests.

Funding

This research received no external funding. This statement should be replaced before submission if project-specific funding is available.

Ethics Statement

This study used publicly available or publicly accessible user-generated social media comments for aggregate-level computational analysis. The released dataset does not include usernames, account identifiers, phone numbers, addresses, or other direct personal identifiers. UzEDSA is intended for aggregate-level Uzbek affective NLP research and must not be used for individual profiling, surveillance, targeting, or sensitive personal inference. The authors should confirm institutional ethics review or exemption requirements before journal submission if required by their institution or local regulations.

Declaration of Generative AI and AI-assisted Technologies

During manuscript preparation, AI-assisted tools were used for language polishing, formatting support, and editorial organization. The authors reviewed and edited the final manuscript and are responsible for all scientific content, data descriptions, tables, figures, and references.

References

1. Picard RW (1997) Affective Computing. MIT Press
2. Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2:1–135
3. Nandwani P, Verma R (2021) A review on sentiment analysis and emotion detection from text. Soc Netw Anal Min 11:81
4. Ekman P (1992) An argument for basic emotions. Cogn Emot 6:169–200
5. Russell JA (1980) A circumplex model of affect. J Pers Soc Psychol 39:1161–1178
6. Plutchik R (1980) Academic Press,. A general psychoevolutionary theory of emotion. In Emotion: Theory, Research, and Experience Vol. 1 (eds Plutchik, R. & Kellerman, H.) 3–33
7. Strapparava C, Mihalcea R (2007) SemEval-2007 Task 14: Affective Text. In Proc. Fourth International Workshop on Semantic Evaluations 70–74ACL
8. Mohammad SM, Bravo-Marquez F, Salameh M, Kiritchenko S (2018) SemEval-2018 Task 1: Affect in Tweets. In Proc. 12th International Workshop on Semantic Evaluation 1–17ACL
9. Demszky D et al (2020) ACL,. GoEmotions: A dataset of fine-grained emotions. In Proc. 58th Annual Meeting of the Association for Computational Linguistics 4040–4054
10. Bostan L-A-M, Klinger R (2018) An analysis of annotated corpora for emotion classification in text. In Proc. COLING 2018 2104–2119ACL
11. Hedderich MA, Lange L, Adel H, Strotgen J, Klakow D (2021) A survey on recent approaches for natural language processing in low-resource scenarios. In Proc. NAACL-HLT. 2545–2568 (ACL, 2021)

12. Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In Proc. NAACL-HLT 4171–4186 (ACL, 2019)
13. Conneau A et al (2020) ACL,. Unsupervised cross-lingual representation learning at scale. In Proc. ACL 2020 8440–8451
14. Xue L et al (2021) mT5: A massively multilingual pre-trained text-to-text transformer. In Proc. NAACL-HLT. 483–498 (ACL, 2021)
15. Mansurov B, Mansurov A, UzBERT (2021) Pretraining a BERT model for Uzbek. Preprint at https://arxiv.org/abs/2108.09814
16. Kuriyozov E, Vilares D, Gomez-Rodriguez C, BERTbek (2024) A pretrained language model for Uzbek. In Proc. SIGUL 2024 33–44ELRA and ICCL
17. Matlatipov SG, Rajabov J, Kuriyozov E, Aripov M (2024) UzABSA: Aspect-based sentiment analysis for the Uzbek language. In Proc. SIGUL 2024 394–403ELRA and ICCL
18. Bender EM, Friedman B (2018) Data statements for natural language processing: Toward mitigating system bias and enabling better science. Trans Assoc Comput Linguist 6:587–604
19. Gebru T et al (2021) Datasheets for datasets. Commun ACM 64:86–92
20. Wilkinson MD et al (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3:160018
21. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46
22. Krippendorff K (2011) Computing Krippendorff's alpha-reliability. (University of Pennsylvania
23. Artstein R, Poesio M (2008) Inter-coder agreement for computational linguistics. Comput Linguist 34:555–596
24. Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Inf Process Manag 45:427–437
25. Powers DMW, Evaluation (2011) From precision, recall and F-measure to ROC, informedness, markedness and correlation. J Mach Learn Technol 2:37–63
26. Joulin A, Grave E, Bojanowski P, Mikolov T (2017) Bag of tricks for efficient text classification. In Proc. EACL 2017 427–431ACL
27. Blitzer J, Dredze M, Pereira F, Biographies (2007) Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In Proc. ACL 440–447 (ACL, 2007)
28. Ribeiro MT, Singh S, Guestrin C (2016) Why should I trust you? Explaining the predictions of any classifier. In Proc. KDD 2016 1135–1144ACM
29. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30:4765–4774
30. Elov B, Alaev R, UzEDSA (2026) Uzbek Emotion and Sentiment Analysis Dataset. Hugging Face https://doi.org/10.57967/hf/8617
31. Elov B, Alaev R, UzEDSA (2026) Uzbek Emotion and Sentiment Analysis Dataset. Mendeley Data https://doi.org/10.17632/jxm9ynjg66.1

Figures

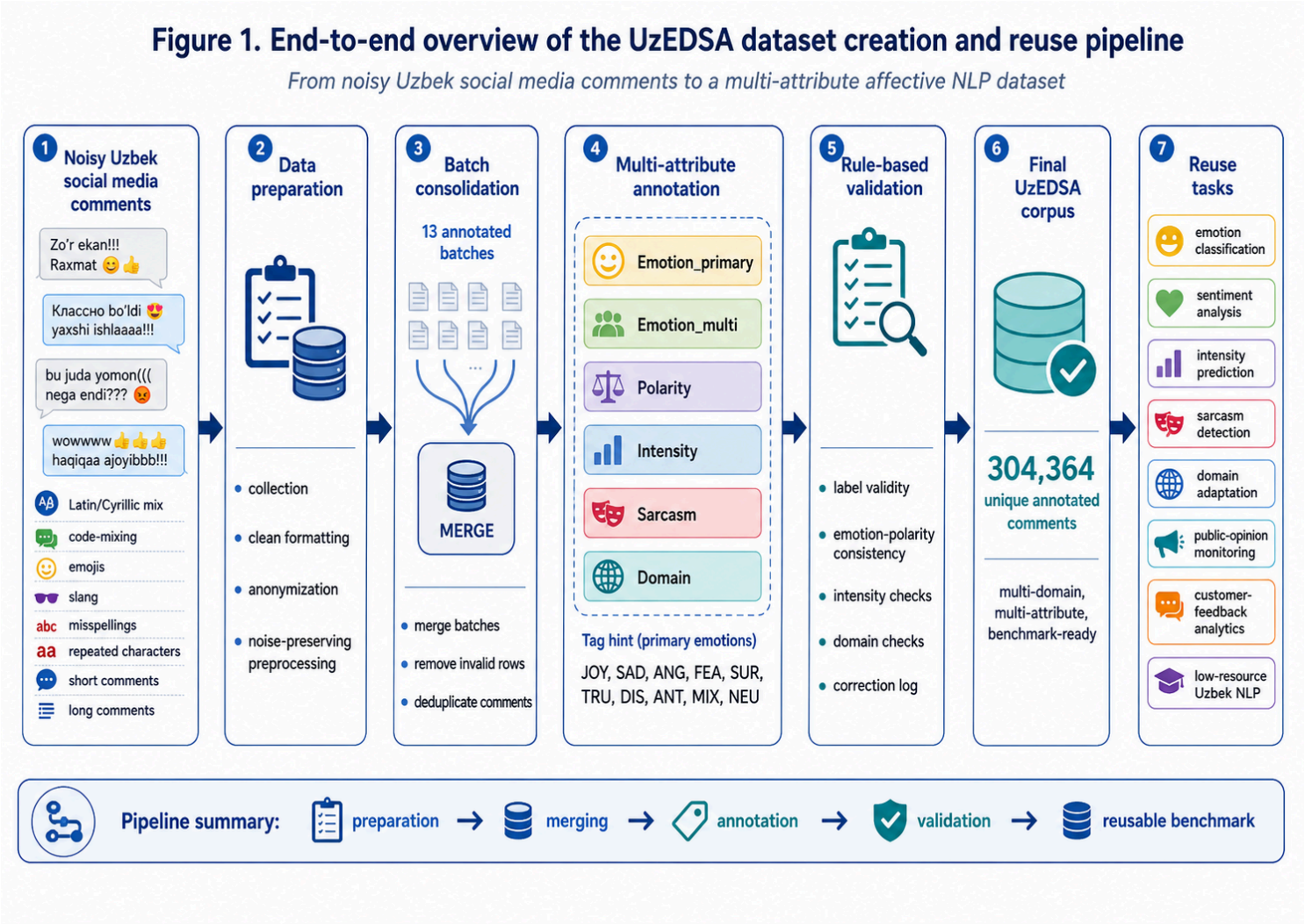


Figure 1

End-to-end overview of the UzEDSA dataset creation and reuse pipeline. The diagram summarizes the transition from noisy Uzbek social media comments to a multi-attribute affective NLP dataset, including preparation, merging, annotation, validation, and reuse tasks.

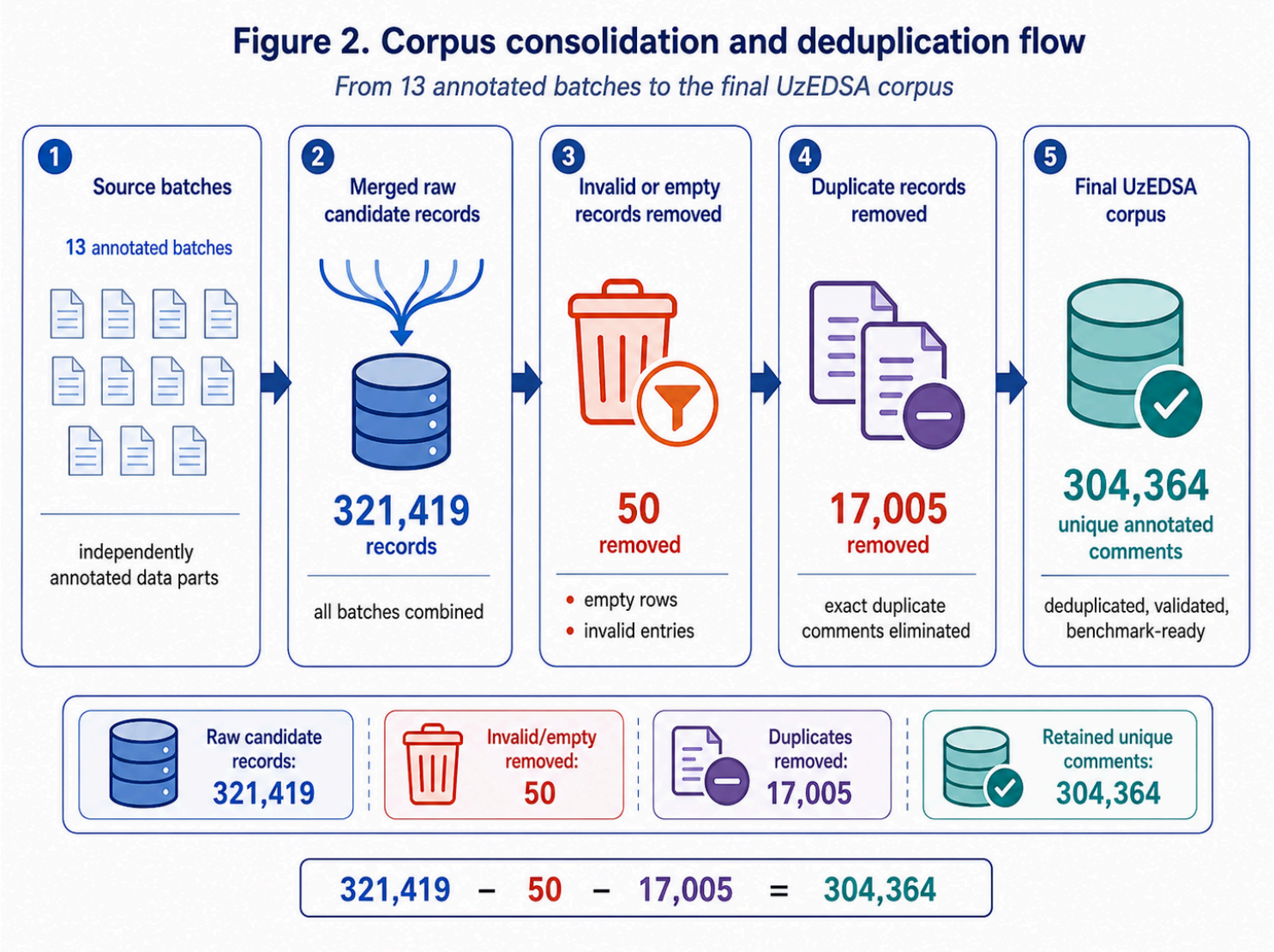


Figure 2

Corpus consolidation and deduplication flow. Thirteen annotated batches were merged, 50 invalid or empty records and 17,005 duplicate records were removed, and 304,364 unique annotated comments were retained.

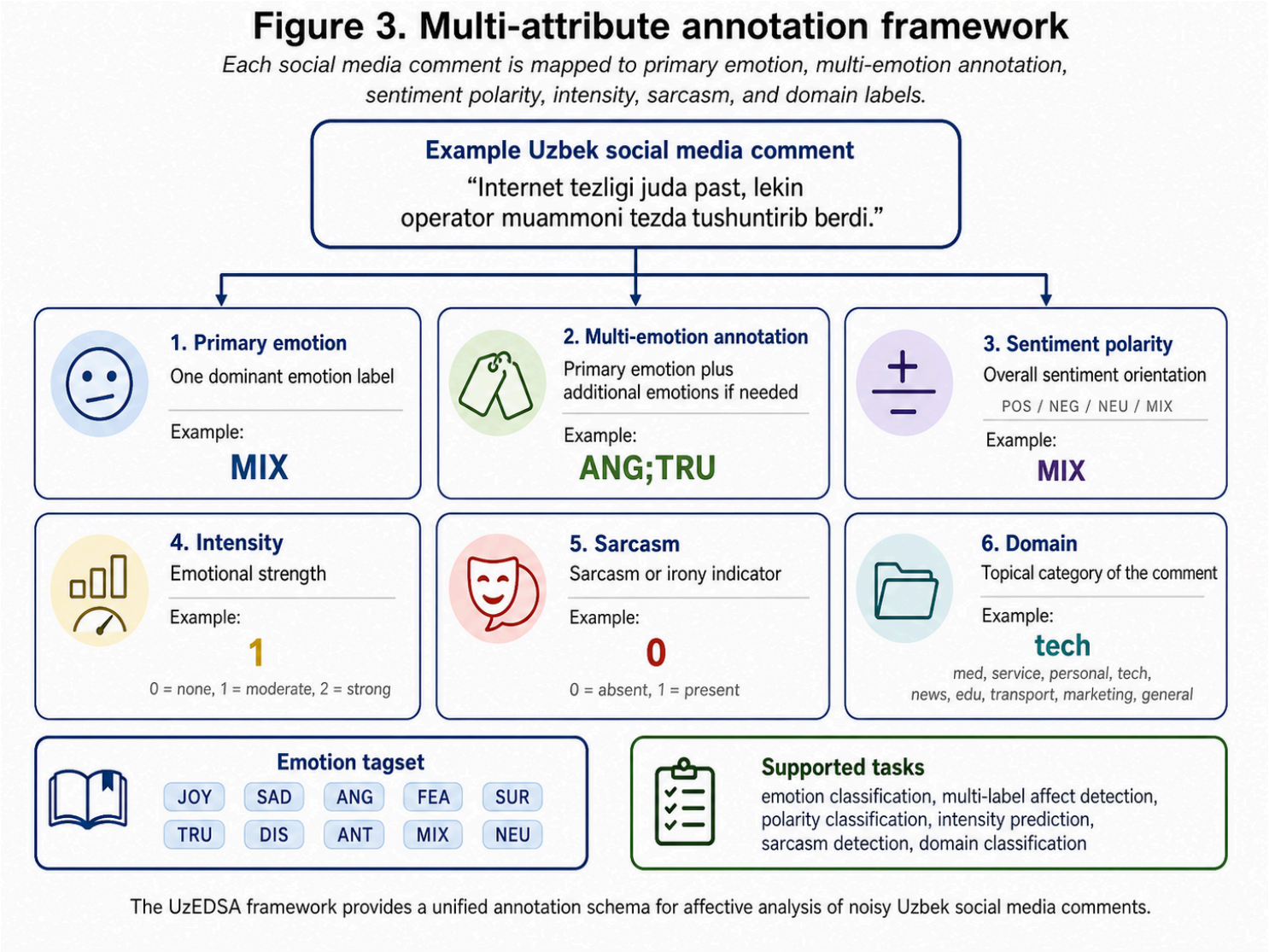


Figure 3

Multi-attribute annotation framework. Each social media comment is mapped to primary emotion, multi-emotion annotation, sentiment polarity, intensity, sarcasm, and domain labels.

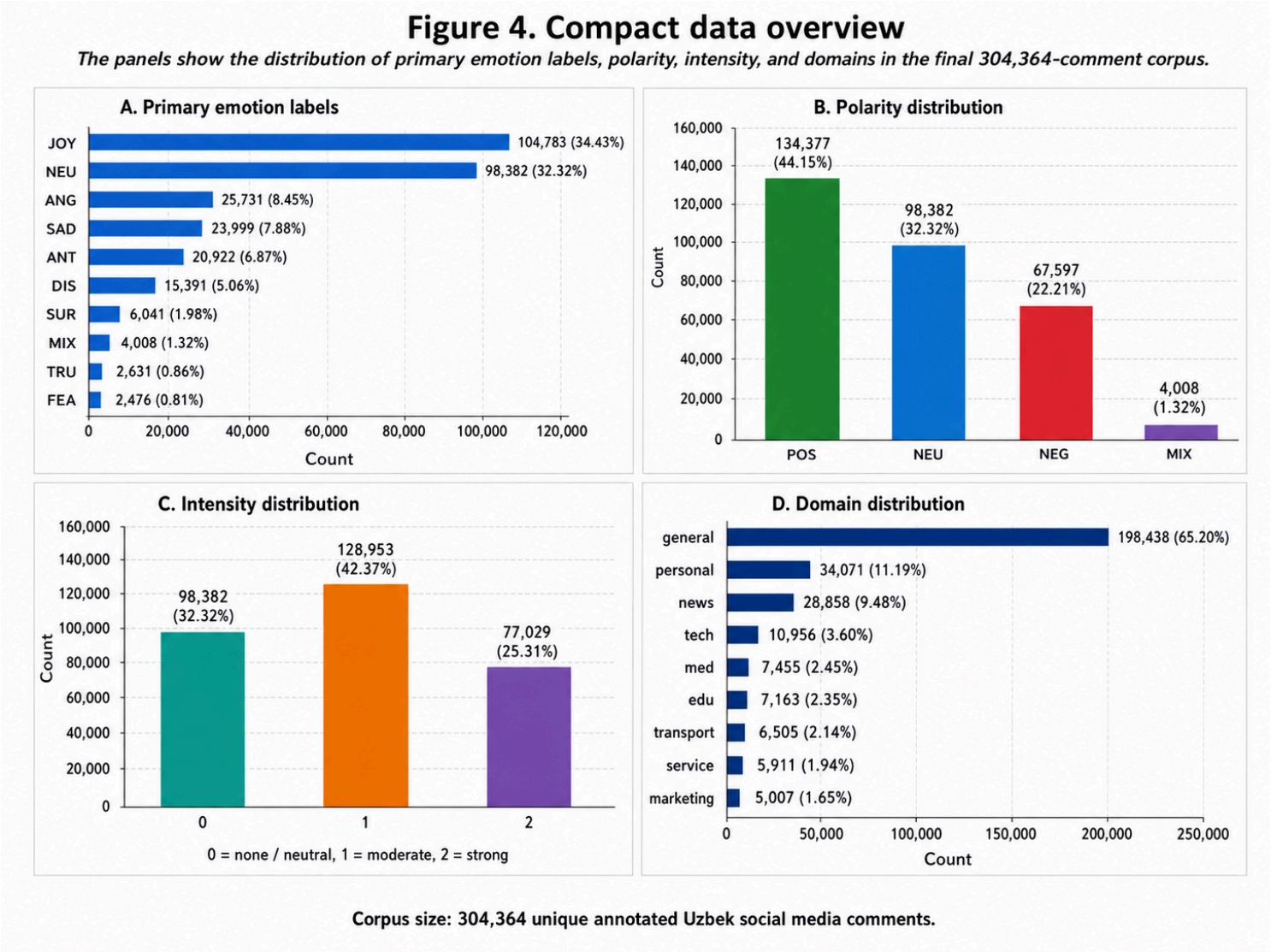


Figure 4

Compact data overview. The panels show the distribution of primary emotion labels, polarity, intensity, and domains in the final 304,364-comment corpus.

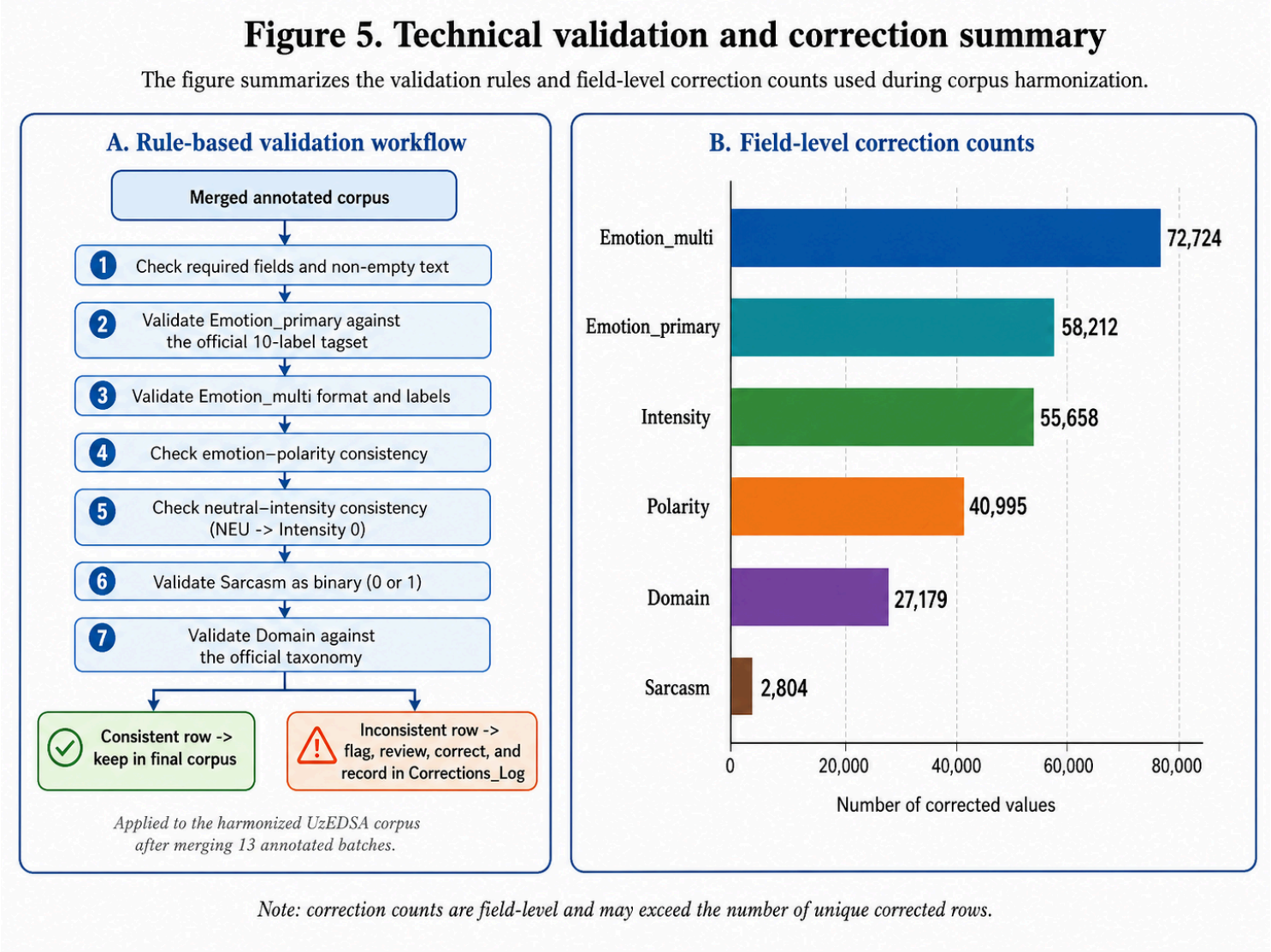


Figure 5

Technical validation and correction summary. The figure summarizes the validation rules and field-level correction counts used during corpus harmonization.