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.

Benchmarking aspect-conditioned sentiment classification of Uzbek healthcare feedback with auxiliary medical entity supervision

Botir Elov

elov@navoiy-uni.uz

Alisher Navo'i Tashkent State University of the Uzbek Language and Literature

Ruhillo Alaev

National University of Uzbekistan named after Mirzo Ulugbek

Research Article

Keywords: Aspect-based sentiment analysis, patient feedback, Uzbek language, natural language processing, medical named entity recognition, multi-task learning, BERTbek, XLM-R

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11748609/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

A healthcare comment can express different sentiments about waiting time, staff interactions, and other service aspects. We benchmarked sentiment classification for a supplied aspect in Uzbek online healthcare feedback and examined how scores differed with auxiliary medical entity supervision.

Methods

The main comparison covered 13 configurations under a reported group-stratified five-fold protocol, with text–aspect pairs from the same normalized comment kept in one fold. BERTbek and XLM-R were compared with lexical, subword, and transformer baselines. Two additional review-only controls supported ablations of aspect conditioning, aspect fusion, and sequential and joint medical named entity recognition training. Macro-averaged F1 on six service aspects was the primary endpoint. The corresponding score on the full medical-domain benchmark was secondary.

Results

BERTbek with aspect fusion and joint medical entity supervision had the highest reported macro-averaged F1 on the service subset (84.91) and full benchmark (82.74), on a 0–100 scale. The differences from standard paired-input BERTbek were 0.84 and 1.83 points, respectively. The direct differences from BERTbek with aspect fusion alone were 0.50 and 0.78 points, respectively. The largest descriptive differences from the paired-input baseline among service aspects were for waiting time and price.

Conclusions

The reported point estimates support further evaluation of medical entity supervision for aspect-conditioned sentiment classification in Uzbek healthcare feedback. Matched test predictions are needed to estimate uncertainty. Independent institutional data are needed to assess usefulness in routine review of service feedback.

Background

Free-text feedback allows patients to describe experiences that structured questionnaires may not capture in detail, including how they were treated, how long they waited and where they encountered difficulties obtaining care. For healthcare quality teams, the challenge is to identify recurring concerns in these narratives while preserving the context needed to interpret them. In a scoping review of 52 studies applying natural language processing (NLP) to unstructured patient feedback, sentiment analysis was the most common technique, appearing in 32 studies [1]. The review also found limited evidence linking NLP-derived findings to established quality indicators or changes in practice [1]. Cammel et al. demonstrated one

practical use in two hospitals by combining topic, sentiment and frequency information to help prioritize patient experience issues [2]. This distinction between extracting usable information and measuring improvements in care informs the scope of the present study.

A single polarity label for an entire comment can obscure the service issue that a reviewer needs to see. A comment can praise a clinician's explanation and criticize the length of the wait. One overall label cannot express both judgments. Aspect-based sentiment analysis (ABSA) links sentiment to a particular care dimension and encompasses several tasks, including identifying aspects, classifying sentiment for a given aspect and extracting aspects and opinions jointly [3]. Here, the unit of analysis is a text–aspect pair. The aspect category is supplied to the classifier, which estimates whether the text expresses positive, negative or neutral sentiment toward that category. This formulation evaluates whether a model can distinguish sentiment across known aspects within the same text.

Studies in several languages illustrate why the task and source of feedback matter. Yazdani et al. combined sentiment classification and topic modeling to analyze Persian-language feedback from hospitalized cancer patients at a single institute [4]. Using online reviews, Gui and He paired sentence-level aspect discovery with document-level sentiment supervision [5]. In Norway, Storset et al. developed an aspect-level benchmark from national patient experience surveys covering general practice and specialist mental healthcare [6]. Mæhlum et al. examined the difficulties of assigning sentiment to Norwegian patient comments, including the use of the neutral category [7]. These studies establish precedents for analyzing free-text feedback, while their sampling frames, annotation schemes and prediction targets define different evaluation questions.

For Uzbek, UzABSA provides an aspect-level benchmark based on restaurant reviews [8]. UzMED-ABSA documents a broader taxonomy of healthcare aspects and includes a mixed-sentiment class [9]. UzMedSentiment provides text–aspect pairs with three polarity classes and annotations for negation, speculation, and sarcasm, alongside textual adverse drug reaction (ADR) signals [10]. Elov and Xusainova described a workflow for analyzing online Uzbek patient comments using a multitask approach that included aspect extraction [11]. The present benchmark evaluates polarity classification in online Uzbek medical-domain feedback when an aspect is supplied, with a service-focused evaluation and a second evaluation covering both clinical and service aspects.

The study compares an Uzbek-specific encoder with a multilingual encoder to test how well each performs on the same task. BERTbek was previously evaluated on Uzbek sentiment classification, topic classification and named entity recognition (NER) [12]. XLM-R provides a multilingual comparison, with pretraining across 100 languages including Uzbek [13]. Published performance on general-language tasks does not establish how either encoder will behave when a review mentions several aspects or expresses sentiment indirectly. The comparison is therefore between specific pretrained models under a common evaluation protocol. Their training corpora, tokenizers and model configurations also differ, which limits attribution of any performance difference to pretraining language alone.

Auxiliary medical NER offers task supervision that may help the encoder represent medically relevant spans while learning aspect-conditioned sentiment. An entity mention can help identify what a remark refers to, but the mention itself does not establish sentiment. For example, a treatment may be discussed alongside praise

for a clinician or dissatisfaction with its cost. Multitask studies in ABSA and clinical information extraction provide methodological precedents for sharing representations across related tasks [14, 15]. We use token-level annotations from the Uzbek Medical Entity Benchmark (UzMedNER) [16] to examine the contribution of auxiliary supervision. Comparing aspect-aware modeling alone with sequential NER transfer and joint training addresses whether the additional task improves sentiment classification under this evaluation.

The study benchmarked aspect-conditioned sentiment classification in Uzbek healthcare feedback and examined the contribution of medical entity supervision. It compared BERTbek and XLM-R with lexical, subword, and transformer baselines in a service-focused and a full-benchmark setting. Within each backbone, it examined differences associated with explicit aspect conditioning, aspect fusion, sequential NER transfer, and joint sentiment and NER training. Descriptive comparisons examined service aspects and linguistic cues. The intended application is to help human reviewers find recurring service concerns in comments with known aspect labels.

Methods

Study design and task

This was a retrospective benchmark of aspect-conditioned sentiment classification using existing Uzbek medical-domain feedback. The input was a review text and a supplied aspect category. For each text–aspect pair, the classifier predicted one of three sentiment classes: negative, neutral or positive. A comment could contribute more than one pair when it addressed several aspects. The task concerned polarity toward the supplied category, with aspect identification outside the evaluation design.

There were two evaluation settings. The primary setting comprised rows annotated with one of six service-related aspects: diagnostics (diagnostika), general service (xizmat), clinician attitude or interaction (shifokor-munosabati), infrastructure (infratuzilma), price (narx) and waiting time (kutish-vaqti). Diagnostics was included because it can describe an experience at the interface between service delivery and clinical assessment. The secondary setting used the full medical-domain benchmark, which also covered symptoms, treatment or procedures, medicines and diet. The service subset was defined by its aspect labels. Its results characterize these topics within the corpus of online feedback.

The available records do not establish whether the service scores were obtained by restricting predictions from full-corpus models or by training and tuning separate service-only models.

Data sources and annotation

UzMedSentiment supplied the sentiment examples. The public documentation describes Uzbek medical-domain comments from forums, Telegram, Instagram and web sources. The tabular schema contains text, an aspect category, a three-class sentiment label, a polarity score and source and language metadata, together with annotated textual cues for adverse drug reactions, negation, speculation and sarcasm [10, 17]. The aspect category and sentiment label were used as the conditioning variable and prediction target,

respectively. The cue annotations were used to define diagnostic subsets. The polarity score and other auxiliary metadata were excluded from the sentiment inputs.

The source study reports 4791 pairs in the benchmark and 1897 in the service subset. The full count matches the inventory documented in the April 2026 Mendeley Data record [17]. The November 2026 release documents a distinct inventory of 4527 pairs and 3177 distinct text strings [10]. The frozen experimental file is required to establish which public release corresponds to the reported benchmark.

The training design used token-level medical NER annotations from the Uzbek Medical Entity Benchmark (UzMedNER) resource family [16]. Its tagging scheme used beginning, inside, and outside (BIO) labels for entity spans. UzMedNER served as the auxiliary training resource for sequential transfer and joint training. The available records do not identify the exact auxiliary file or the implemented tag inventory.

Preprocessing and evaluation partitions

The reported preprocessing included Unicode and whitespace normalization, orthographic cleanup, canonicalization of aspect labels, and consistency checks for source and language tags. Data validation included polarity-score range and duplicate checks. Rows with distinct aspects were eligible even when the underlying comment matched. Figure 1 shows the reported workflow. Text–aspect pairs sharing the same normalized comment formed a single group for partitioning.

The reported evaluation design used a common group-stratified five-fold outer partition across configurations. In each iteration, one fold was reserved for testing. An inner development partition drawn from the remaining folds supported tuning and early stopping. The source describes five random seeds. Exact normalization rules, realized group assignments, seed identifiers, and overlap checks between the sentiment and NER sources were unavailable for verification.

Baseline and proposed models

The baselines comprised a majority-class classifier, linear support vector machines (SVMs) with term frequency–inverse document frequency (TF–IDF) features, fastText [18], multilingual Bidirectional Encoder Representations from Transformers (mBERT) [19], XLM-R [13], and BERTbek [12]. Table 1 identifies the 13 main configurations. In PAIR, the aspect category was rendered as readable Uzbek text and paired with the review, following the general practice of conditioning classification on explicit aspect text [20]. The two TEXT-ONLY ablation controls received the review alone. The reported transformer checkpoint names were bert-base-multilingual-cased, xlm-roberta-base, and bertbek-news-big-cased.

Table 1
Model configurations for aspect-conditioned sentiment classification

ID	Model	Input or training configuration
M0	Majority class	Most frequent sentiment class
M1	TF–IDF + linear SVM	Review-only lexical features
M2	TF–IDF + aspect SVM	Lexical features from review and aspect text
M3	fastText-PAIR	Review and aspect text with subword features
M4	mBERT-PAIR	Paired input with multilingual BERT
M5, M6	XLM-R-PAIR and BERTbek-PAIR	Standard paired-input fine-tuning
M7, M8	XLM-R-AF and BERTbek-AF	Aspect fusion with sentiment supervision
M9, M10	XLM-R-AF-NERInit and BERTbek-AF-NERInit	Aspect fusion with sequential NER transfer
M11, M12	XLM-R-AF-MTL and BERTbek-AF-MTL	Aspect fusion with joint sentiment and NER training

Figure 2 summarizes aspect fusion (AF) and the two auxiliary training schemes. The AF design combines a sequence-level representation and an aspect representation with an attention-weighted summary of review tokens. The combined representation is passed to a sentiment head trained with class-weighted cross-entropy. The auxiliary token classifier uses token-level cross-entropy over medical BIO labels. Sentiment inference takes the review and supplied aspect.

Two configurations use medical entity supervision. In sequential transfer (AF-NERInit), NER training produces an encoder initialization for subsequent fine-tuning on aspect-conditioned sentiment. In joint multi-task learning (AF-MTL), the sentiment and NER heads share the encoder, and training alternates sentiment and NER mini-batches. A loss weight governs the contribution of the NER objective in the joint configuration. The AF variant without NER provides the matched reference for each auxiliary-training contrast.

Within each backbone, the ablation contrasts were PAIR versus TEXT-ONLY for explicit aspect conditioning, AF versus PAIR for fusion, AF-NERInit versus AF for sequential transfer and AF-MTL versus AF for joint supervision. The AF-MTL versus AF-NERInit comparison addresses the relative performance of the two auxiliary training strategies.

Table 2 summarizes the transformer search space and training settings reported in the source configuration table. Development data supported hyperparameter selection within each outer fold. The selected run settings and software and checkpoint revisions were unavailable for verification.

Table 2
Transformer search settings in the reported protocol

Hyperparameter	Candidates or protocol setting
Optimizer and schedule	AdamW with linear scheduling and warm-up
Learning rate	2×10^{-5}, 3×10^{-5} or 5×10^{-5}
Batch size	16 or 32
Maximum sequence length	128 or 256 tokens
Dropout	0.10
Epoch budget	Up to 10
Warm-up ratio	0.10
Weight decay	0.01
Auxiliary loss weight	0.1, 0.3 or 0.5
Early-stopping patience	2 (value in the source configuration table)

The reported search space lists candidate settings. Run-level selections and the implemented early-stopping patience could not be verified.

Outcomes and statistical analysis

The primary endpoint was macro-averaged F1 (Macro-F1) on the service subset. Macro-F1 on the full medical-domain benchmark was secondary. Macro-F1 weights the F1 values of the three sentiment classes equally. Support-weighted F1 averages the class-specific scores in proportion to their test support. Accuracy is the percentage of correctly classified pairs.

The source protocol describes summarizing outer-test scores across folds and seeds. This report presents the available aggregate scores and calculates point differences by subtracting values displayed for the same evaluation setting. The exact aggregation rule could not be verified. The results are descriptive point estimates.

Diagnostic comparisons used the reported scores by service aspect and by textual ADR, negation, speculation, or sarcasm cue. Cue subsets could overlap. Source-specific summaries covered forum entries, Telegram, and pooled Instagram and web entries. These summaries describe strata within the same benchmark.

Reproducibility

Reproduction requires the frozen data files, preprocessing and split manifests, model and software revisions, run settings, and matched out-of-fold (OOF) predictions. A prediction record should identify the model, test fold, seed, text–aspect pair, normalized-comment group, reference sentiment, and predicted sentiment. Availability of these materials is described in the Declarations.

ChatGPT (OpenAI) assisted with manuscript and figure preparation. The scope is described in the AI use declaration.

Results

Evaluation corpus and partitions

Results are based on aggregate tables from the source study. Absolute differences were recalculated from the reported point estimates.

The reported benchmark comprised 4791 text–aspect pairs, matching the inventory described in the April repository record [17]. The six service categories accounted for 1897 pairs (39.60%, Table 3). The full set included 1970 negative, 1783 neutral, and 1038 positive pairs. Forum entries accounted for 3688 pairs (76.98%), followed by Telegram (927 pairs, 19.35%) and combined Instagram and web entries (176 pairs, 3.67%). A comment could contribute more than one pair when it addressed multiple aspects.

Table 3
Reported benchmark composition

Inventory item	Text–aspect pairs	Share of full set (%)
Full benchmark	4791	100.00
Service subset	1897	39.60
Negative	1970	41.12
Neutral	1783	37.22
Positive	1038	21.67
Forum	3688	76.98
Telegram	927	19.35
Instagram and web entries	176	3.67

Shares use the reported denominator of 4791 pairs. The service subset overlaps the class and source categories. Shares were rounded independently.

The class and source counts each sum to 4791. The ten aspect counts in the source inventory sum to 4788, a discrepancy of three pairs. Counts of distinct normalized-comment groups and their distribution across folds were unavailable.

Primary and secondary model comparisons

BERTbek-AF-MTL (M12) had the highest reported Macro-F1 in the primary service-focused setting at 84.91, with a weighted F1 of 86.83 and accuracy of 87.19% (Table 4). Compared with BERTbek-PAIR (M6), the strongest standard paired-input transformer, the absolute differences were + 0.84 points for Macro-F1, + 1.05 for weighted F1 and + 1.07 percentage points for accuracy. The strongest lexical or subword baseline by

service Macro-F1 was fastText-PAIR (M3) at 77.11. The difference between M12 and M3 was 7.80 Macro-F1 points.

Table 4
Reported performance in the service subset and full benchmark

Model ID	Service Macro-F1	Service Wt. F1	Service Accuracy (%)	Full Macro-F1	Full Wt. F1	Full Accuracy (%)
M0	19.30	22.84	39.96	19.43	23.96	41.12
M1	75.61	76.61	76.65	71.29	70.39	70.49
M2	76.64	77.65	77.65	75.95	75.13	75.13
M3	77.11	78.02	78.08	74.61	74.92	74.88
M4	81.92	83.63	84.11	78.34	81.02	81.47
M5	83.61	85.17	85.59	79.88	82.47	82.96
M6	84.07	85.78	86.12	80.91	83.44	83.92
M7	84.19	85.82	86.21	80.96	83.28	83.69
M8	84.41	86.17	86.49	81.96	84.27	84.71
M9	84.44	86.03	86.37	81.34	83.71	84.12
M10	84.68	86.53	86.87	82.41	84.78	85.17
M11	84.62	86.36	86.71	81.77	84.19	84.58
M12	84.91	86.83	87.19	82.74	85.19	85.65

Model IDs are defined in Table 1. F1 scores use a 0–100 scale. Wt. F1, support-weighted F1. The service-subset M0 scores are shown as reported. Their aggregation could not be verified.

On the secondary full medical-domain benchmark, M12 ranked first by reported Macro-F1 at 82.74, with a weighted F1 of 85.19 and accuracy of 85.65%. M6 scored 80.91, 83.44 and 83.92%, respectively, giving differences of + 1.83 Macro-F1 points, + 1.75 weighted F1 points and + 1.73 percentage points in accuracy. The strongest lexical baseline was the aspect-conditioned SVM (M2) at 75.95 Macro-F1. Its gap to M12 was 6.79 points.

For PAIR, BERTbek had a higher reported Macro-F1 than XLM-R by 0.46 points on service aspects and 1.03 points on the full benchmark. In the matched AF-MTL configuration, the differences were 0.29 and 0.97 points. A constant negative prediction applied to the reported full-benchmark label distribution gives 41.12% accuracy, 19.43 Macro-F1, and 23.96 weighted F1, which reproduce the displayed full-benchmark M0 entries.

Ablation and auxiliary entity supervision

Table 5 gives the scores and contrasts for explicit aspect conditioning, aspect fusion, and the two auxiliary training variants. Figure 3 displays the point differences. In BERTbek, PAIR exceeded TEXT-ONLY by 0.52 points on service aspects and 1.73 points on the full benchmark. AF exceeded PAIR by 0.34 and 1.05 points, respectively. Relative to AF, sequential NER transfer added 0.27 and 0.45 points. The AF-MTL versus AF differences were 0.50 and 0.78 points.

Table 5
Ablation Macro-F1 scores and direct contrasts

Configuration or contrast	Service BERTbek	Service XLM-R	Full BERTbek	Full XLM-R
TEXT-ONLY	83.55	83.02	79.18	78.61
PAIR	84.07	83.61	80.91	79.88
AF	84.41	84.19	81.96	80.96
AF-NERInit	84.68	84.44	82.41	81.34
AF-MTL	84.91	84.62	82.74	81.77
PAIR − TEXT-ONLY	+ 0.52	+ 0.59	+ 1.73	+ 1.27
AF − PAIR	+ 0.34	+ 0.58	+ 1.05	+ 1.08
AF-NERInit − AF	+ 0.27	+ 0.25	+ 0.45	+ 0.38
AF-MTL − AF	+ 0.50	+ 0.43	+ 0.78	+ 0.81
AF-MTL − AF-NERInit	+ 0.23	+ 0.18	+ 0.33	+ 0.43

Unsigned values are reported scores on a 0–100 scale. Signed values are differences in Macro-F1 points calculated from the displayed scores. TEXT-ONLY denotes the review-only ablation controls.

For XLM-R, the changes on the service subset and full benchmark, in that order, were + 0.59 and + 1.27 for PAIR versus TEXT-ONLY and + 0.58 and + 1.08 for AF versus PAIR. Relative to AF, AF-NERInit added 0.25 and 0.38 points, while AF-MTL added 0.43 and 0.81 points. AF-MTL also exceeded AF-NERInit by 0.23 points on service aspects and 0.33 points on the full benchmark for BERTbek, and by 0.18 and 0.43 points, respectively, for XLM-R. These last contrasts compare the two auxiliary-training regimes.

Diagnostic and error analyses

M12 had a higher reported Macro-F1 than M6 for each of the six service aspects (Table 6). The largest differences were for waiting time (68.57 to 76.34, + 7.77 points, 77 pairs) and price (74.18 to 80.62, + 6.44 points, 104 pairs). The lowest M12 estimate was for waiting time, and the highest was for infrastructure (88.74). The reported support is the number of text–aspect pairs.

Table 6
Reported Macro-F1 by service aspect

Service aspect	Pairs	M6	M12	Δ (points)
Diagnostics	626	81.22	83.48	+ 2.26
General service	469	85.17	87.26	+ 2.09
Clinician attitude or interaction	424	78.44	82.91	+ 4.47
Infrastructure	197	86.03	88.74	+ 2.71
Price	104	74.18	80.62	+ 6.44
Waiting time	77	68.57	76.34	+ 7.77

M6, BERTbek-PAIR. M12, BERTbek-AF-MTL. Δ is M12 minus M6 on the 0–100 scale. Supports count text–aspect pairs.

Among cue-defined subsets, the largest differences were observed for ADR signals (+ 6.45 points, 158 pairs) and negation (+ 6.35 points, 816 pairs). Sarcasm had the lowest M12 point estimate among the four cue-defined subsets at 63.94, based on 48 pairs. The difference was + 5.67 points for sarcasm and + 4.39 for speculation (Table 7). The source-stratified differences ranged from + 2.19 points for the pooled Instagram and web entries to + 2.62 points for Telegram. These source comparisons describe diagnostic strata of the same benchmark. Cue categories can overlap and their counts should therefore be interpreted separately.

Table 7
Reported Macro-F1 by annotated cue and recorded source

Diagnostic subset	Pairs	M6	M12	Δ (points)
ADR signal present	158	71.84	78.29	+ 6.45
Negation present	816	72.31	78.66	+ 6.35
Speculation present	1321	75.42	79.81	+ 4.39
Sarcasm present	48	58.27	63.94	+ 5.67
Forum	3688	81.14	83.56	+ 2.42
Telegram	927	82.47	85.09	+ 2.62
Instagram and web entries	176	80.88	83.07	+ 2.19

ADR, adverse drug reaction signal in text. The pooled source subset comprises Instagram and web entries. Cue-positive subsets can overlap. Δ is M12 minus M6 in Macro-F1 points.

Discussion

Principal findings and relation to prior work

BERTbek with aspect fusion and joint medical named entity recognition (NER) training had the highest reported Macro-F1 among the evaluated configurations in both settings. On the primary service subset, its score was 84.91, a difference of 0.84 points from BERTbek-PAIR. On the full benchmark, the corresponding difference was 1.83 points (Table 4). These comparisons combine the addition of aspect fusion with auxiliary training. The direct comparison for the auxiliary task is AF-MTL versus AF within the same backbone. In that comparison, the reported differences for BERTbek were 0.50 points on the service subset and 0.78 points on the full benchmark. The corresponding differences for XLM-R were 0.43 and 0.81 points (Table 5). Both backbones showed the same direction of change in the aggregate estimates. The smaller differences between joint training and sequential NER transfer leave the choice between these training strategies open pending paired estimates and comparable training budgets.

The TEXT-ONLY, PAIR and AF comparisons provide context for the size of the auxiliary-task differences. The largest matched improvement on the full benchmark occurred when the aspect was supplied explicitly. In the service subset, the observed changes from aspect conditioning, fusion and auxiliary supervision were smaller. One explanation is that specifying the aspect helps the model connect evaluative language with the intended target in comments that address multiple topics. Auxiliary NER may further help the encoder retain domain terms when a comment mentions a service issue alongside a symptom, medicine or procedure. These mechanisms are hypotheses. They could be tested with matched training budgets, a control for the additional supervision and paired examples where the same text conveys different sentiments toward two supplied aspects. Prior work has used shared representations in multitask ABSA and clinical information extraction [14, 15]. The present comparison examines a different auxiliary signal, namely medical entity labels from a separate annotation resource used to support sentiment classification.

BERTbek had higher reported Macro-F1 than XLM-R in the matched configurations. This comparison extends the evidence about BERTbek in Uzbek by considering healthcare feedback and a supplied-aspect sentiment task [12, 13]. Differences in pretraining corpora, tokenization and other checkpoint properties could all contribute to the observed ordering. Related Uzbek work has documented an ABSA benchmark for restaurant reviews, a healthcare annotation resource with a broader aspect inventory and a workflow for analyzing patient comments that includes aspect extraction [8, 9, 11]. The contribution here is to examine sentiment for known healthcare aspects and to compare two ways of using medical entity supervision within that task.

The healthcare feedback literature also provides comparators for the scope of these findings. Storset et al. annotated Norwegian survey comments for aspects and sentiment in general practice and specialist mental healthcare, while Yazdani et al. combined sentiment analysis and topic modeling for Persian-language feedback from one cancer institute [4, 6]. Gui and He studied online healthcare reviews using sentence-level aspect discovery and document-level sentiment supervision [5]. These studies show how provenance, annotation unit and model inputs shape the evaluation question. Their outcomes help situate the Uzbek benchmark, whose input consists of an online comment and a supplied category. Mæhlum et al. also documented challenges in assigning neutral sentiment in Norwegian patient feedback [7]. Class-level outputs are needed to determine how neutral sentiment contributes to the error patterns observed here.

Implications for healthcare feedback analysis

Aspect-level output could help a reviewer distinguish feedback about waiting, costs and other features of the same service encounter. For a workflow that already tags comments by topic, a sentiment classifier could organize candidate comments for review and direct attention to recurring issues within a category. Cammel et al. combined topics, sentiment and frequency to identify candidate priorities in two hospitals [2]. A related evaluation in Uzbek could examine whether summaries by aspect agree with service teams' assessments of the original comments. If a workflow begins with unlabeled comments, aspect detection is an additional component whose errors must be measured jointly with sentiment classification.

Waiting time and price showed the largest observed advantages for AF-MTL over PAIR among the six service aspects. They are useful priorities for further error review because the estimates were based on 77 and 104 text–aspect pairs, respectively, and waiting time still had the lowest reported AF-MTL score among the six aspects. A reviewer study could examine missed complaints and false alarms for these categories, measure agreement between reviewers and assess the time required to review a batch of comments. Such evidence would connect technical performance to a defined quality-improvement task. A review of the broader literature found limited evidence linking automated feedback analyses to routine practice or established quality measures [1].

The diagnostic patterns also merit focused review. The ADR and negation subsets showed the largest cue-specific point differences, and sarcasm remained the lowest-scoring cue subset, with 48 reported pairs. A review of matched errors could determine whether mistakes occur when negation changes the evaluation of a service or when an apparently positive phrase conveys criticism. An ADR label in this study represents a signal annotated in text. Its operational interpretation would require separate clinical assessment and a validation design specific to that purpose.

Limitations and generalizability

First, interpretation depends on the analytic corpus and the training scope. The reported total of 4791 pairs differs by three from the sum of the ten aspect counts. Reconciliation requires the frozen experimental release, eligible rows, and class and group counts by fold. Separate service-only training and selection of service predictions from full-benchmark models address different estimands. The available records do not identify which procedure produced the reported service scores. The realized partitions and overlap with the auxiliary NER resource are needed to assess the protection against leakage.

Second, the available aggregate results support descriptive comparisons. Aligned predictions would allow comparison by pair, comment group, fold, and seed and would support uncertainty estimation [21]. Training partitions can overlap across folds, and repeated seeds can evaluate the same comments. A paired comment-group bootstrap on fixed out-of-fold predictions could quantify variation across evaluated comments conditional on the fitted models and realized splits. Dispersion across seeds would characterize a separate source of variation. Rare or overlapping diagnostic subsets also require counts by comment group and class.

Third, the source and task limit transfer. Forum entries contributed approximately 77% of the reported pairs, and the service subset was defined by aspect labels within the same online corpus. Source-stratified scores describe internal diagnostic subsets. A separate holdout by source or collection period would test transfer to

new feedback, and an independent sample from a healthcare institution would address relevance to that institution's workflow. Future evaluations should record the collection setting and sampling denominator, check annotation agreement and quantify performance for each class and service aspect. An institution-based evaluation could also compare the priorities obtained from model-assisted review with those identified through existing surveys or complaint records.

Conclusions

BERTbek with aspect fusion and joint medical named entity recognition had the highest reported Macro-F1 for supplied-aspect sentiment classification in Uzbek online healthcare feedback. The highest point estimates were 84.91 on the primary service subset and 82.74 on the full benchmark. Relative to aspect fusion alone, joint training was associated with small positive differences for both BERTbek and XLM-R. These comparisons identify medical entity supervision as a candidate for further evaluation.

A potential application is to organize comments with known aspect labels for review by healthcare service teams. Waiting time and price were priorities for targeted error analysis. Before assessing operational utility, future work should reproduce the paired comparisons from matched out-of-fold predictions and validate performance on independently collected feedback.

Abbreviations

ABSA Aspect-based sentiment analysis

ADR Adverse drug reaction

AF Aspect fusion

BERT Bidirectional Encoder Representations from Transformers

BIO Beginning, inside and outside

mBERT Multilingual BERT

MTL Multi-task learning

NER Named entity recognition

OOF Out-of-fold

SVM Support vector machine

TF–IDF Term frequency–inverse document frequency

XLM-R XLM-RoBERTa

Declarations

Ethics approval and consent to participate

The study used existing text benchmark datasets and involved no direct recruitment of or interaction with participants. The institutional determination regarding ethics review and consent for secondary use has not been verified.

Consent for publication

Not applicable. The study reports aggregate results without individual-level material.

Availability of data and materials

Public UzMedSentiment records are available from Mendeley Data at https://doi.org/10.17632/6yrswp74pd.2 [17] and https://doi.org/10.17632/sc3ft63ncb.1 [10]. The public UzMedNER record is available at https://doi.org/10.17632/77jwmshbcp.1 [16]. The public data records identify candidate sources. The exact experiment files and availability of the evaluation code, split manifest, and prediction outputs could not be confirmed.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no external funding.

Authors' contributions

Individual author roles and approval of this manuscript version are pending author confirmation.

Acknowledgements

An earlier version of this work is available as an SSRN preprint [22].

AI use declaration

ChatGPT (OpenAI) assisted with structuring and editing the manuscript, checking references and arithmetic, and preparing code for Figures 1–3 from the supplied study protocol and reported results. The authors are responsible for verifying the text, figures, and scientific interpretation.

References

1. Feizollah A, Lin CY, O'Malley L, Thompson W, Listl S, Byrne M. The use of natural language processing to interpret unstructured patient feedback on health services: scoping review. J Med Internet Res. 2025;27:e72853. 10.2196/72853.
2. Cammel SA, De Vos MS, van Soest D, Hettne KM, Boer F, Steyerberg EW, et al. How to automatically turn patient experience free-text responses into actionable insights: a natural language programming (NLP) approach. BMC Med Inf Decis Mak. 2020;20:97. 10.1186/s12911-020-1104-5.
3. Hua YC, Denny P, Wicker J, Taskova K. A systematic review of aspect-based sentiment analysis: domains, methods, and trends. Artif Intell Rev. 2024;57:296. 10.1007/s10462-024-10906-z.
4. Yazdani A, Shamloo M, Khaki M, Nahvijou A. Use of sentiment analysis for capturing hospitalized cancer patients' experience from free-text comments in the Persian language. BMC Med Inf Decis Mak. 2023;23:275. 10.1186/s12911-023-02358-2.
5. Gui L, He Y. Understanding patient reviews with minimum supervision. Artif Intell Med. 2021;120:102160. 10.1016/j.artmed.2021.102160.
6. Storset LC, Jelin E, Norman RM, Bjertnæs ØA, Øvrelid L, Velldal E. From pain to praise: aspect-based sentiment analysis for Norwegian patient feedback. In: Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026). Rabat: Association for Computational Linguistics; 2026. pp. 191–202. 10.18653/v1/2026.healing-1.16
7. Mæhlum P, Samuel D, Norman RM, Jelin E, Bjertnæs ØA, Øvrelid L et al. It's difficult to be neutral – human and LLM-based sentiment annotation of patient comments. In: Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) at LREC-COLING 2024. Torino: ELRA and ICCL; 2024. pp. 8–19. https://aclanthology.org/2024.cl4health-1.2/. Accessed 15 Nov 2026.
8. Matlatipov SG, Rajabov J, Kuriyozov E, Aripov M. UzABSA: aspect-based sentiment analysis for the Uzbek language. In: Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages at LREC-COLING 2024. Torino: ELRA and ICCL; 2024. pp. 394–403. https://aclanthology.org/2024.sigul-1.47/. Accessed 15 Nov 2026.
9. Elov B, Alaev R. UzMED-ABSA: An Uzbek Medical Domain Dataset for Aspect-Based Sentiment Analysis [dataset]. Version 1. Mendeley Data; 2026. https://doi.org/10.17632/v43nkjzw39.1.
10. Elov B, Alaev R. UzMedSentiment: An Uzbek Medical-Domain Dataset for Aspect-Based Sentiment Analysis [dataset]. Version 1. Mendeley Data; 2026. https://doi.org/10.17632/sc3ft63ncb.1.
11. Elov BB, Xusainova ZY. Bemor izohlarida sentiment tahlil uchun BPMN axborot modeli [A BPMN information model for sentiment analysis of patient comments]. TATU Xabarlari. 2026;1(77):21–41. Uzbek. https://tsuull.uz/sites/default/files/bemor_izohlarida_sentiment_tahlil_uchun_bpmn_axborot_modeli.pdf. Accessed 15 Nov 2026.
12. Kuriyozov E, Vilares D, Gómez-Rodríguez C. BERTbek: a pretrained language model for Uzbek. In: Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages at LREC-COLING 2024. Torino: ELRA and ICCL; 2024. pp. 33–44. https://aclanthology.org/2024.sigul-1.5/. Accessed 15 Nov 2026.

13. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F et al. Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. pp. 8440–51. 10.18653/v1/2020.acl-main.747
14. He R, Lee WS, Ng HT, Dahlmeier D. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics; 2019. pp. 504–15. 10.18653/v1/P19-1048
15. Mulyar A, Uzuner O, McInnes B. MT-clinical BERT: scaling clinical information extraction with multitask learning. J Am Med Inf Assoc. 2021;28(10):2108–15. 10.1093/jamia/ocab126.
16. Elov B, Alaev R. Uzbek Medical Entity Benchmark [dataset]. Version 1. Mendeley Data; 2026. https://doi.org/10.17632/77jwmshbcp.1.
17. Elov B, Alaev R. An Uzbek Medical-Domain Dataset for Aspect-Based Sentiment Analysis [dataset]. Version 2. Mendeley Data; 2026. https://doi.org/10.17632/6yrswp74pd.2
18. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Volume 2, Short Papers. Valencia: Association for Computational Linguistics; 2017. pp. 427–31. https://aclanthology.org/E17-2068/. Accessed 15 Nov 2026.
19. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Volume 1, Long and Short Papers. Minneapolis: Association for Computational Linguistics; 2019. pp. 4171–86. 10.18653/v1/N19-1423
20. Sun C, Huang L, Qiu X. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Volume 1, Long and Short Papers. Minneapolis: Association for Computational Linguistics; 2019. pp. 380–5. 10.18653/v1/N19-1035
21. Dror R, Baumer G, Shlomov S, Reichart R. The hitchhiker's guide to testing statistical significance in natural language processing. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Volume 1, Long Papers. Melbourne: Association for Computational Linguistics; 2018. pp. 1383–92. 10.18653/v1/P18-1128
22. Elov B, Alayev R. Aspect-based sentiment analysis of hospital service reviews using BERTbek and XLM-R [preprint]. SSRN. 2026. https://doi.org/10.2139/ssrn.6570837.

Figures

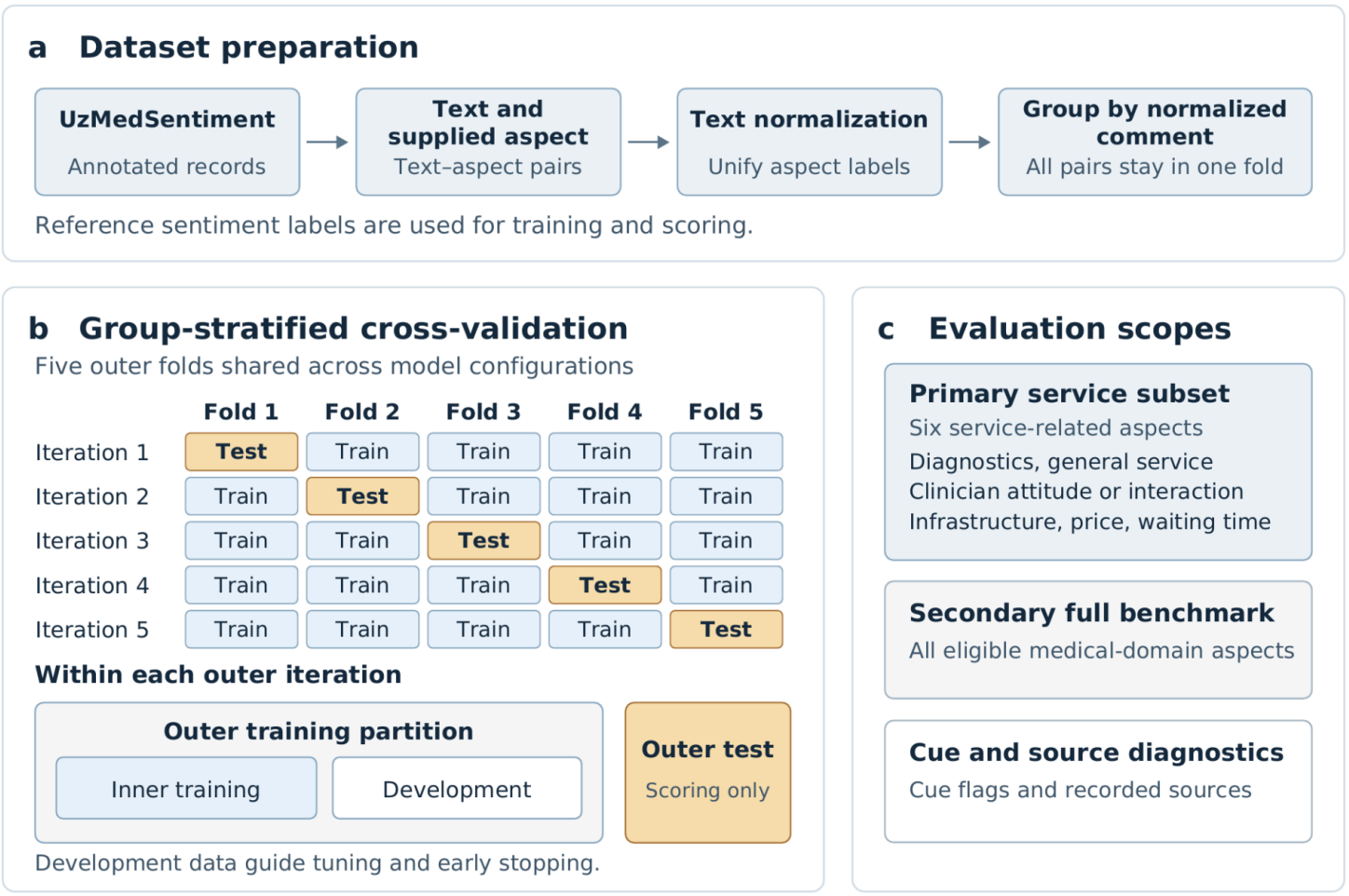


Figure 1

Dataset preparation and group-stratified evaluation workflow

(a) The reported preprocessing groups text–aspect pairs by normalized comment. (b) Five outer folds rotate as held-out test sets, with development data drawn from each outer training partition. (c) The evaluation covers six service-related aspects and the full medical-domain benchmark. Cue and source strata are used for descriptive diagnostics. ChatGPT (OpenAI) assisted with drawing the reported workflow.

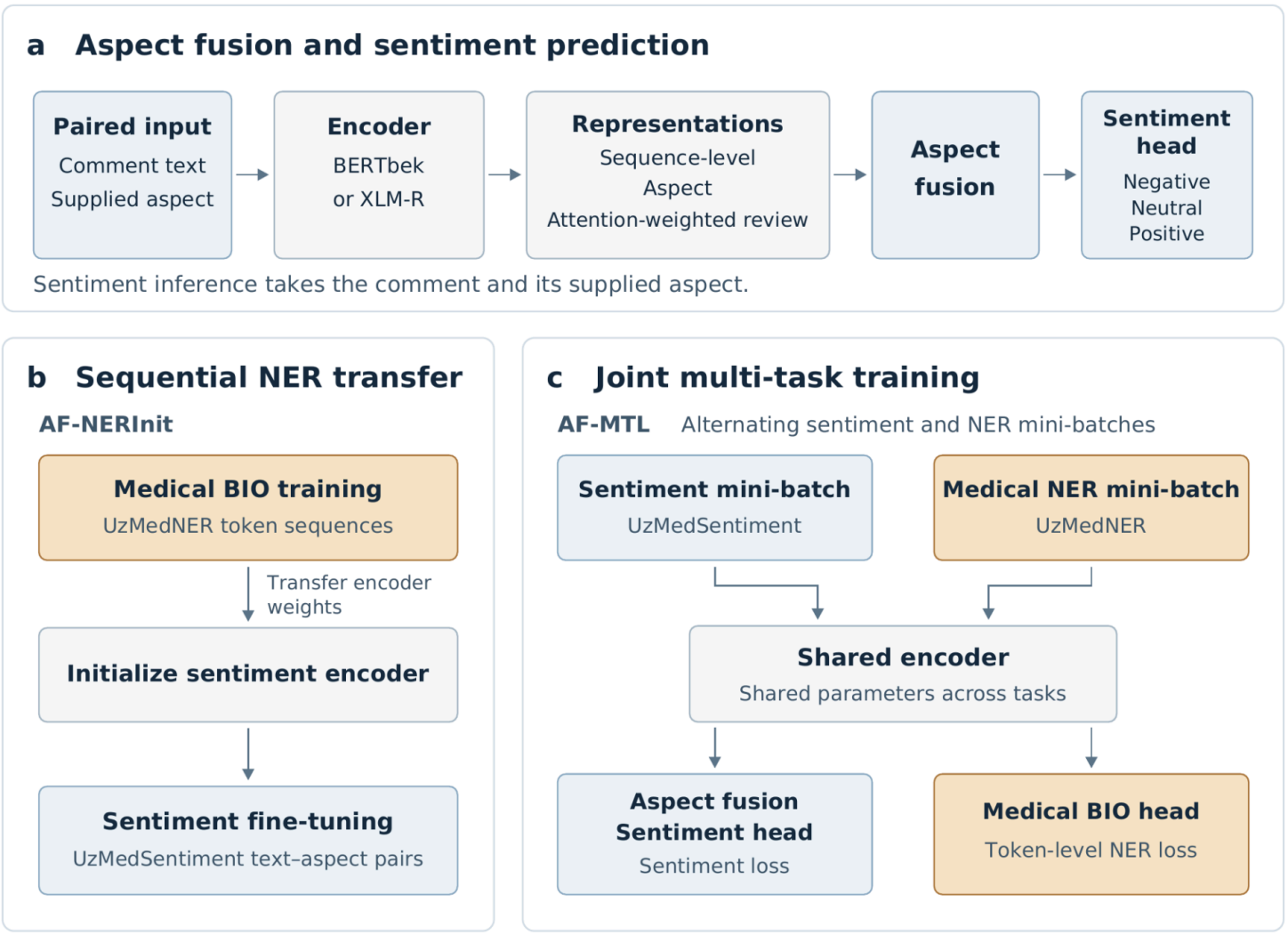


Figure 2

Aspect fusion and auxiliary medical entity supervision

(a) The supplied aspect and review are encoded for three-class sentiment prediction with aspect fusion. (b) Sequential transfer initializes the sentiment encoder with weights from medical named entity recognition (NER) training. (c) Joint training alternates sentiment and NER mini-batches through a shared encoder with separate task heads. ChatGPT (OpenAI) assisted with drawing the model schematic from the reported design.

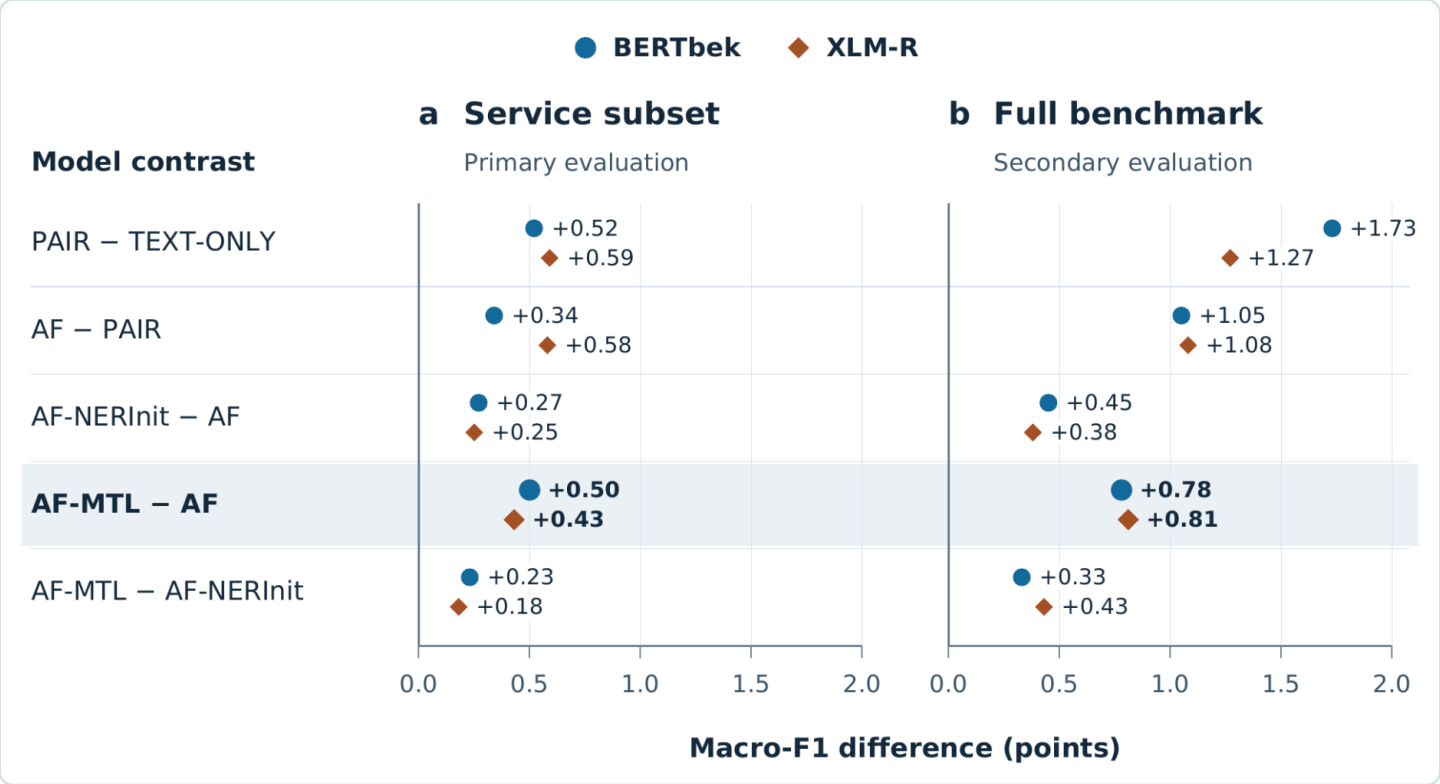


Figure 3

Reported Macro-F1 contrasts across evaluation settings

(a) Service subset. (b) Full benchmark. Each point is the Macro-F1 score of the first configuration minus that of the second for the same backbone and setting. Differences are in points on a 0–100 scale. Circles denote BERTbek, and diamonds denote XLM-R. Shading highlights AF-MTL minus AF, the direct contrast of joint NER training with sentiment-only aspect fusion. Table 5 gives the source scores. ChatGPT (OpenAI) assisted with plotting these reported contrasts.