KeAi
CHINESE ROOTS
GLOBAL IMPACT

International Journal of Innovation Studies



LARGE-SCALE ASPECT-BASED SENTIMENT ANALYSIS OF HOSPITAL PATIENT REVIEWS USING AN ATTENTION-ENHANCED BIDIRECTIONAL LONG SHORT-TERM MEMORY NETWORK

[1]S. Mathi, [2]Dr. V. Asaithambi

[1]Research Scholar, PG and Research Department of Computer Science,
Government Arts College (Autonomous), Nandanam, Chennai-35, Tamil Nadu, India

[2]Associate Professor, PG and Research Department of Computer Science,
Government Arts College (Autonomous), Nandanam, Chennai-35, Tamil Nadu, India

[1]mathischolar2024@gmail.com, [2]dr.v.asaithambi@gmail.com

Abstract

Patient reviews on public portals provide much more detail of hospital care than the satisfaction surveys do, and a single polar score eradicates the information about which services satisfy or disappoint patients. Aspect-based sentiment analysis (ABSA) is a promising way to actionable health-care analytics as it associates opinions with named service aspects. The typical number of annotated items in existing hospital studies is a limited number of thousands, omitting the neutral class or the least common features, or requiring intensive computation, which are usually required. This paper addresses the challenges associated with large-scale ABSA on the hospital reviews corpus consisting of 87,310 cleaned reviews (from 105,636 raw reviews) each labeled with a three-class sentiment class and a multi-label annotation of the 10 service aspects and a no-aspect. We design a two-branch bidirectional long short-term memory (BiLSTM) model that consists of an attention-pooled sentiment classifier trained using class-balanced weight and a dedicated aspect detector trained using positive weight damped and validated decision thresholds. The opinion profile for aspects is derived from the predicted sentiments by taking the average of the detected sentiments. The feature-selection stage consists of vocabulary selection, sequence masking and TF-IDF feature limits. On a held-out test set of 8,763 reviews, it achieves 95.77% accuracy with a Matthews correlation coefficient of 0.833 and a macro-F1 score of 0.880 outpacing a unidirectional LSTM, a plain BiLSTM, and two TF-IDF linear baselines. The dedicated aspect detector achieves 0.962 micro-F1 and 0.937 macro-F1, while auxiliary heads within the sentiment networks only achieve 0.790 and 0.327. The greatest negative shares are for cost (40.6%), waiting time (32.2%) and test procedures (30.3%). The network contains approximately three million parameters and trains in about 6 minutes on one CPU core. Constraints are the one run training and only supervision at the review level. The results validate use of lightweight recurrent models as a viable foundation for continuous monitoring of hospital feedback; the Python code released enables the model to be replicated on other review collections.

Keywords: aspect-based sentiment analysis; bidirectional LSTM; attention mechanism; hospital reviews; patient experience; multi-label classification; class imbalance; healthcare quality



1. Introduction

Patients' reviews about hospitals are a rolling diary of hospital care as it really is. For one group of hospitals, map platforms, social networks and dedicated portals accumulate thousands of comments each month – and these numbers have been too high to read by hand for some time. Saying that, automated text analysis is a practical necessity and sentiment analysis has been applied in healthcare over more than 10 years (Greaves et al., 2013; Rahim et al., 2021) [19, 30].

In recent years, the research direction has shifted from global-polarity-based sentiment analysis to aspect-based sentiment analysis (ABSA) in which a system must determine which aspect is mentioned in the text in addition to the attitude expressed toward each aspect (Zhang et al., 2023) [37]. “the surgeon was wonderful but the billing took 2 hours” is a combination of two opposing ideas that a hospital manager has to respond to if he wants to be able to do so. There are three main streams in the recent press.

Work published in 2024. Perikos and Diamantopoulos (2024) [28] optimized BERT, RoBERTa, DistilBERT and XLNet for ABSA and enhanced the best model (RoBERTa) by using token-level explanation techniques. AlNasser and AlMuhaideb (2024) [5] constructed an Arabic hospital-review corpus, which is achieved by combining MARBERT with a SVM to achieve higher accuracy of 92.14% compared to GPT-4 on aspect and polarity tasks. To address class imbalance problem in aspect-level sentiment classification, Ganganwar and Rajalakshmi (2024) [17] generated synthetic training data, while Azman Busst et al. (2024) [7] used an ensemble of BiLSTM models and a rule-based aspect–sentiment mapper that achieved F1 scores above 92% on SemEval benchmarks.

Work published in 2025. The diversity of healthcare ABSA was quite deepened. Alamoudi and Solaiman (2025) [2] combined pseudo-labelling with partial human input to produce an explainable healthcare dataset in Arabic, while Li et al. (2025a, 2025b) [25, 26] applied ChatGPT pseudo-labelling to incorporate ABSA templates and chain-of-thought prompts into the model for classification of Chinese hospital reviews and their ICD-11 disease group linkage. Prabhune et al. (2025) [29] introduced a keyword layer to RoBERTa for Indian hospital reviews and achieved an F1 score of 0.923, and Singh et al. (2025) [34] suggested an architecture of BiLSTM-GRU model with embedding NER-BERT for healthcare decision making. In the non-medical context, Alhusseini and Feizi-Derakhshi (2025) [4] demonstrated that a convolutional network (96.41%) can perform at or above the accuracy of a BiLSTM (93.12%) in short balanced reviews; a reminder that the architecture is dependent on the data.

Work published in 2026. We are able to find only the latest preprint that extends ABSA to the low-resource language, Uzbek, using entity-aware multi-task learning, with 84.91 macro-F1 on the hospital-service subset (Elov and Alayev, 2026) [14]. The direction of the trend is clear: bigger pre-trained encoders, language specific resources and prompt-based large language models rule the current healthcare ABSA, typically measured on small and carefully annotated resource sets.

1.1 Research gap



These studies when read together indicate four gaps to motivate the present work. First, the size of the evaluation is small. While hospital specific studies have focused on a few thousand items (e.g. 4,862 in Prabhune et al., 2025, or 6,000 sentences in Alamoudi and Solaiman, 2025), real review streams consist of tens, if not hundreds, of thousands of short reviews. Secondly, class imbalance is generally not addressed in a systematic way [21]. Fahmi et al. (2023) [15] report that the absence of neutral labels was one reason for the sub-optimal neutral-class result of their Indonesian BERT-LSTM model, and Prabhune et al. (2025) [29] did not even have any neutral reviews as they argued it was very sparse. Rarity is even less of a concern, like safety or cost. Third, transformers, or LLM pipelines, are computational burdensome and, in the case of proprietary models, privacy and cost burdens that restrict their deployment in resource-constrained hospitals; a few smaller, compact recurrent models with attention have been less studied at scale in this context. Fourth, unlike the methods based on explicit sentiment labels of the SemEval-style corpora, many hospital corpora already contain a review-level sentiment label, along with a set of aspect tags without providing polarity for each aspect; hence, methods designed for the SemEval-style corpora cannot be directly applied to these corpora and an explicit aggregation strategy must be developed. We believe that none of the studies that we have read here has integrated all of these features with a corpus of this size.

1.2 Objectives

The goals of this study are: (i) to create an attention-enhanced BiLSTM model to predict three-class sentiment and eleven aspect indicators over a large dataset of over 87,000 hospital reviews; (ii) to solve the extreme sentiment class imbalance (86.1% positive) and aspect class imbalance (618 safety mentions, 47,045 doctor mentions) using principled weighting and per-aspect thresholds; (iii) to quantify the impact of each component (direction of reading, attention pooling, auxiliary tasks, dedicated aspect detection) through controlled comparison with principled models; and (iv) to translate the predictions into an aspect-level opinion profile to be shared with hospital managers.

1.3 Major contributions

- A large-scale study involving 87,310 cleaned reviews over 161 hospitals with a detailed description of the label and aspect structure of the hospital reviews ABSA dataset.
- Two-branch sentiment classification using a BiLSTM with a weighted pooling mechanism, using balanced weights for the two classes, and a multi-label aspect detector using a damped weighted pooling mechanism with validation-tuned thresholds, followed by aspect-level aggregation.
- Short mathematical analysis of the balanced weighting, of masked attention pooling and of the relationship between positive weights and optimal decision thresholds.
- A reproducible comparison with unidirectional LSTM, a plain BiLSTM, a multi-task BiLSTM and the plain linear models provided by TF-IDF with bootstrap confidence intervals, McNemar tests, confusion matrices, and a component-by-component impact analysis.



1.4 Organisation of the remaining sections

In section 2, twenty recent studies (from 2021 to 2026) are reviewed and the methodology, results and limitations of these studies are recorded. Section 3 introduces the proposed strategy: the architecture, the data set, and the entire preprocessing, feature selection, classification, setting of the parameters, and coding of the strategy. Section 4 presents the results of the environment of implementation, metrics of evaluation, the impact of each methodological component, a discussion, and the practical and theoretical implications. Finally, advantages and disadvantages, and future directions, are given in Section 5.

2. Literature Review

Twenty studies were reviewed in this section, between 2021 and 2026. Table 1 provides the authors, the proposed methodology, the main result and the limitation that is relevant for this work for each of them. Limitations are expressed by the original authors as they expressed them; otherwise they are our own interpretation in the context of the goals of this paper and expressed as gaps rather than defects.

2.1 Patient-review analysis before the large-language-model era (2021–2023)

The first attempts at hospital reviews used classical learning and topic models. Rahim et al. (2021) [30] applied sentiment and service-quality classifiers on Malaysian Facebook reviews and concluded that a support vector machine is the best choice in cross validation. Because there are few aspect-level labels, Gui and He (2021) [20] combined unsupervised aspect discovery with supervised document sentiment and achieved 98.2% accuracy on the Yelp's healthcare reviews sentiment dataset. In India, Bansal and Kumar (2022) [9] extracted attributes from the reviews of hospitals and in Saudi Arabia, Alhazzani et al. (2023) [3] used predefined categories of patient experience from the reviews. Fahmi et al. 2023 [15] applied the BERT-LSTM on Indonesian reviews and obtained F1 = 0.597, primarily due to the neutral class. In other applications outside of health care, Chen et al. (2021) [12] demonstrated the effectiveness of a BiLSTM with interactive attention over simpler benchmark models. Moderate size of the corpus and the emphasis on either topics or polarity but not both at the same time and with the help of controlling imbalance are common features of these papers.

2.2 Transformers and prompt-based large language models (2024–2026)The focus was then moved to pre-trained encoders and prompting from 2024 on. Perikos and Diamantopoulos (2024) [28] fine-tuned four transformers and set RoBERTa as the best one, as well as introduce explanation techniques. AlNasser et al., 2024 [5] showed that Arabic fine-tuned models or embedding-based models can outperform GPT-4 in a hospital environment. Alamoudi and Solaiman (2025) [2] reduced the cost of the annotation by having ChatGPT label sentences under human supervision, while Li et al. (2025a, 2025b) [25, 26] classified patient needs without training any model. Prabhune et al. (2025) [29] used RoBERTa and added rules to boost rare aspects in Indian reviews, while Elov and Alayev (2026) [14] supervised an aspect-conditioned encoder with medical entities in the case of Uzbek. These systems get high scores on their own data, but cannot be directly compared with one another: they are different in language, in granularity of labels, polarity scheme and corpus size.



2.3 Recurrent models, attention and imbalance handling

Recurrent encoders are still appealing when there are restrictions on computing resources. Azman Busst et al. (2024) [7] created BiLSTM ensembles which achieved an F1 over 92% on SemEval data, while Singh et al. (2025) [34] combined NER-BERT embeddings with an LSTM-GRU layer for healthcare. Alhusseini and Feizi-Derakhshi [4] warn that a CNN may be better than a BiLSTM on short, balanced text. Ganganwar and Rajalakshmi (2024) [17] demonstrated that synthetic data can outperform weighted oversampling and class-weighted loss for aspect classification using BERT. Last but not least, Basha et al. (2024) [10] extract topics from hospital-website reviews, and Alsaleh et al. (2025) [6] classify app reviews using a hybrid BERT-BiLSTM-CNN model. Data scarcity, domain transfer and robustness are all cited as unresolved issues in the surveys conducted by Zhang et al. 2023 [37] that are evident in hospital data.

Table 1. Summary of twenty recent studies (2021–2026): proposed methodology, reported result and limitation.

Study	Proposed methodology	Reported result	Limitation
Rahim et al. (2021) [30] Healthcare 9(12):1679	Supervised ML sentiment analyzer and SERVQUAL quality classifier trained on manually annotated Facebook reviews of Malaysian public hospitals; several classical learners compared by 5-fold cross-validation.	SVM consistently outperformed the other learners in 5-fold cross-validation, both for the quality dimensions and for sentiment.	Classical sparse features; sentiment and service-quality dimensions are modelled separately; no aspect-level polarity.
Gui and He (2021) [20] Artif Intell Med 120:102160	Joint framework: unsupervised sentence-level aspect extraction plus supervised document-level sentiment classification, trained on healthcare reviews without aspect annotation.	98.2% sentiment accuracy on Yelp healthcare reviews; aspects and their polarity distribution recovered without aspect labels.	Aspects are discovered rather than mapped to a fixed management taxonomy; no minority-class metrics reported in the abstract.
Chen et al. (2021) [12] J Front Comput Sci Technol 15(3)	Bi-IAN: BiLSTM encodes target and context separately, and an interactive attention network learns their mutual attention vectors.	Better accuracy and F1 than benchmark models on SemEval-2014 Task 4 and Chinese review data.	Requires annotated aspect terms; benchmark data are not hospital reviews.



Study	Proposed methodology	Reported result	Limitation
Bansal and Kumar (2022) [9] New Gener Comput 40(4)	ABSA built on attribute extraction from roughly 30,000 Indian hospital reviews (as summarised by Alamoudi and Solaiman, 2025) [2]; aspects include doctors, staff, facilities and cost.	Showed that aspect-level granularity yields actionable insight for hospital management.	Pipeline depends on extracted attributes; imbalance-aware multi-label learning not reported.
Alhazzani et al. (2023) [3] Appl Sci 13(18):10305	PX_BERT and related deep models classify Saudi-dialect patient-experience comments into predefined categories (about 19,000 labelled comments, as summarised by Alamoudi and Solaiman, 2025) [2].	Accurate multi-label topic classification of patient comments.	No sentiment labelling; dataset internal to the Ministry of Health and not public.
Fahmi et al. (2023) [15] UGM repository	Multiclass-multioutput BERT-LSTM predicting three aspects (services, human resources, facilities) with four sentiment states for Indonesian hospital reviews.	F1 = 0.597.	Weak neutral-class recognition because neutral labels are scarce (reported by the authors).
Zhang et al. (2023) [37] IEEE TKDE 35(11)	Survey that unifies single and compound ABSA tasks in one taxonomy and reviews pre-trained and generative methods.	No experiments; identifies limited labelled data, domain transfer and robustness as open problems.	No hospital-specific evaluation.
Ganganwar and Rajalakshmi (2024) [17] J Inf Telecommun 8(2)	Synthetic-data generation to balance classes for BERT-based aspect sentiment classification on SemEval-2014/15/16 restaurant data.	Synthetic balancing beat weighted oversampling and class-weighted cross-entropy.	Sentence-level benchmark labels; restaurant domain; extra generation cost.



Study	Proposed methodology	Reported result	Limitation
Azman Busst et al. (2024) [7] HighTechInnov J 5(1)	Ensemble of BiLSTM models extracting aspects and sentiments, merged by a rule-based aspect–sentiment mapper.	F1 = 92.98% (SemEval-2014 Restaurant, correct features) and 95.54% (SemEval-2016 Laptop, mapper features).	Small benchmarks; ensemble raises cost; no healthcare data.
Perikos and Diamantopoulos (2024) [28] Big Data CognComput 8(11):141	Fine-tuned BERT, RoBERTa, DistilBERT and XLNet on MAMS, SemEval and Naver data, then applied explainability techniques to token-level decisions.	RoBERTa best: 89.16% accuracy (MAMS and SemEval) and 97.62% (Naver).	General-domain data; large pre-trained models with high compute demand.
AlNasser and AlMuhaideb (2024) [5] Big Data CognComput 8(11):156	HEAR Arabic hospital-review corpus; BERT embeddings with classical classifiers and fine-tuned Arabic BERT models compared with GPT-4.	Joint MARBERT + SVM model best, 92.14% accuracy, ahead of GPT-4.	Arabic-specific; transformer-based cost; imbalance handling not the focus.
Basha et al. (2024) [10] IJ-AI 13(3):3456	Attribute extraction and aspect-term selection from hospital-website reviews using a modified LDA with TF-IDF (M-LDA TF-IDF).	Coherence score 0.478, better than other topic models compared.	Topic-level stage only; no sentiment classifier in the reported result.
Alsaleh et al. (2025) [6] CMC 82(1) (as summarised by Alhusseini and Feizi-Derakhshi, 2025) [4]	Hybrid BERT + BiLSTM + CNN with an ABSA framework for app-review classification.	About 94% accuracy.	App-review domain; not healthcare.



Study	Proposed methodology	Reported result	Limitation
Alamoudi and Solaiman (2025) [2] IDEAS 2025 (Springer)	EHSAN: ChatGPT pseudo-labelling with tiered human review for an explainable Arabic healthcare ABSA dataset (6,000 sentences); AraBERT and DistilBERT fine-tuned.	AraBERT sentiment accuracy 0.93, macro-F1 0.84; aspect macro-F1 0.66 (17 classes) and 0.78 (6 classes); little loss with ChatGPT-only labels.	Single data source and Saudi dialects (reported by the authors); small corpus.
Li et al. (2025a) [25] PLoS ONE 20(3):e0313442	ABSA templates with chain-of-thought prompts embedded in ChatGPT-4o to sort Chinese patient reviews into three need categories.	Weighted precision 0.944 versus 0.890 for direct prompting.	Proprietary LLM; only three coarse categories; cost and privacy concerns.
Li et al. (2025b) [26] J Med Internet Res 27:e66344	ChatGPT with chain-of-thought ABSA prompts plus the ICD-11 API to map dissatisfaction onto disease categories; ChatGPT-3.5, ChatGPT-4o and Qwen-7b compared.	ChatGPT-3.5 gave the best balance of stability, expense and runtime.	Three coarse aspects; dependence on API-served models.
Prabhune et al. (2025) [29] J Health Res 39(6)	RoBERTa fine-tuning plus a keyword-based aspect layer on 4,862 Google Maps reviews of 25 Indian hospitals; six aspects; 670 external reviews.	F1 = 0.923, accuracy 92.67%; better than lexicon and monolingual transformer baselines.	Neutral class dropped (binary polarity) because of sparsity; relatively small corpus.
Alhusseini and Feizi-Derakhshi (2025) [4] arXiv:2509.19346	CNN versus BiLSTM on oversampled, balanced reviews of two LLM apps with TextBlob-derived labels.	CNN 96.41% versus BiLSTM 93.12% on a balanced 1,700-review test set.	Lexicon labels and oversampling; not aspect-level; not healthcare.
Singh et al. (2025) [34] IEEE Access 13:47459	OptiASAR: BiLSTM-GRU sequence model with NER-BERT embeddings for healthcare ABSA.	Authors report significant improvements for healthcare reviews.	Hybrid architecture with a heavy pre-trained component; scale of evaluation not stated in the abstract.



Study	Proposed methodology	Reported result	Limitation
Elov and Alayev (2026) [14] SSRN preprint	BERTbek and XLM-R with Entity-Aware Aspect Fusion and joint medical-NER multi-task learning for Uzbek hospital reviews.	Hospital subset: macro-F1 84.91, weighted-F1 86.83, accuracy 87.19; largest gains on waiting time (+7.77) and price (+6.44).	Preprint, not peer reviewed; low-resource single language.

2.4 Synthesis

Table 1 gives rise to three observations. First, the reported scores vary between F1 = 0.597 (Fahmi et al., 2023) [15] and 98.2% accuracy (Gui and He, 2021) [20] – but it is not possible to compare the performance of the two models with each other because the labels, languages, class counts and test sets are different. If they were the same, the scores would be comparable; we only compare our model against the baselines re-implemented on one corpus. Second, only few studies measure greater than around five thousand items and only a few reports per-aspect precision, recall and thresholds for rare aspects. Third, the neutral class is consistently the weakest class (Fahmi et al., 2023) [15] or it is eliminated (Prabhune et al., 2025) [29]. The framework that follows maintains all three sentiment classes, examines 87310 reviews and provides a report for each of the aspects separately.

3. Proposed Method

The proposed method considers the hospital reviews as short sequences and processes them sequentially in two aligned branches, having the same encoder design. One branch identifies the sentiment (out of three) of each review, and the other branch identifies the aspects (out of eleven) indicated in each review. Finally, the two outputs are aggregated to form an aspect level opinion profile. The general architecture of the system (Section 3.1), the dataset (Section 3.2) and the full workflow (feature selection, classification, code and parameter settings) (Section 3.3) are presented.

3.1 Overall architecture

Fig. 1 shows the end-to-end flowchart and Fig. 2 the network architecture. To be prepared for this shared pipeline, cleaned reviews go through a preparation process where a vocabulary is chosen from the training partition only and each review is transformed into a fixed-length integer sequence with a padding mask. The first branch (A) is a sentiment model, which consists of an embedding layer, a bidirectional LSTM taking the reviews as input from left to right, and from right to left, an additive attention layer that projects the word states into a review vector, a dense layer, and a softmax output over negative, neutral, and positive classes. In our experiments the multi-task design with an auxiliary sigmoid head for the aspects achieved the best sentiment accuracy when attention pooling was employed, so we chose to include it in the same dense layer. The second branch (B) is a stand-alone

aspect detector which has the same encoder and eleven sigmoid outputs but is trained independently using a positively weighted binary cross-entropy, and a threshold value for each aspect is tuned separately. Lastly, the aggregation module calculates what percentage is predicted negative, neutral and positive out of all the reviews in which an aspect is found.

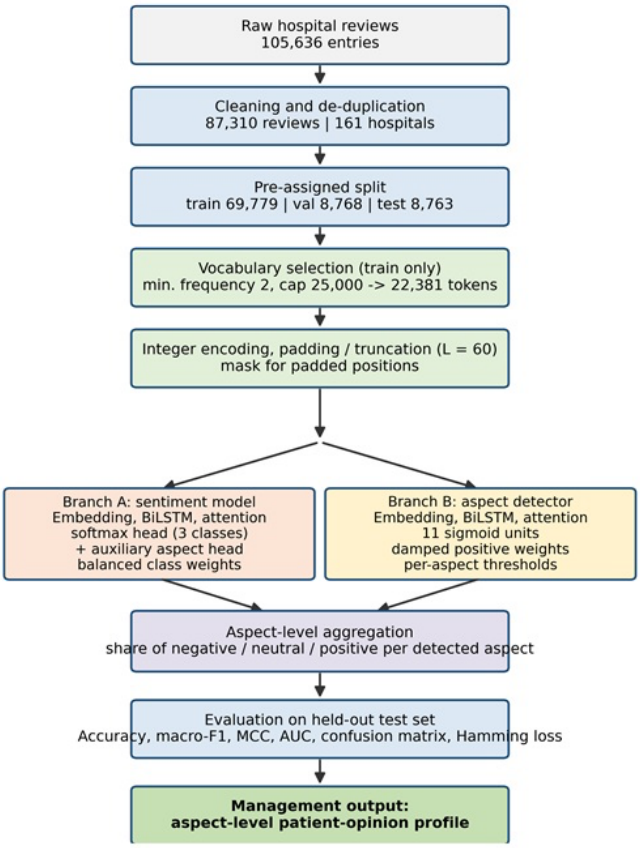


Fig. 1. Flowchart of the proposed large-scale aspect-based sentiment analysis pipeline, from the raw corpus to the aspect-level opinion profile.

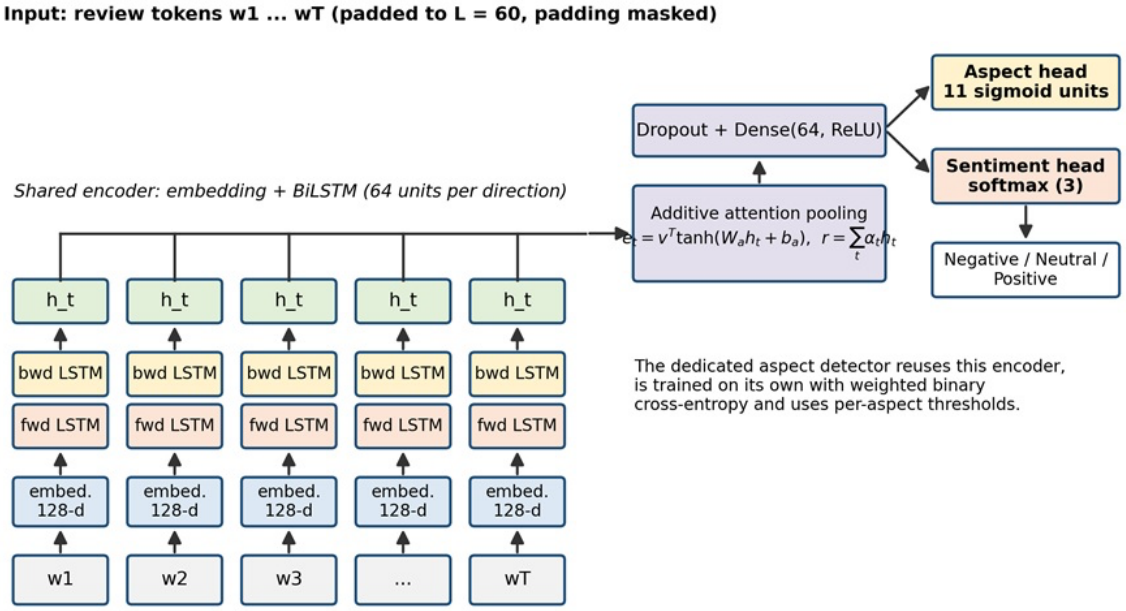


Fig. 2. Architecture of the attention-enhanced BiLSTM. The shared encoder (embedding, forward and backward LSTM, attention pooling, dense layer) feeds a sentiment head and an aspect head; the dedicated aspect detector reuses the encoder design.

The rationale behind each architectural decision is as follows: Bidirectional reading is performed as the meaning of an opinion word can be contingent on the negation or contrast that follows it (e.g., "not at all helpful") [33]. Attention pooling allows the network to focus on the opinion and aspect words in very short texts [8] and the weights of the attention can be analyzed by experts. As confirmed in Section 4.3, two separate branches are used since the sentiment task has a clear majority class (positive) and the aspect task is multi-label with rare classes.

3.2 Dataset description

The experiments employ the hospital reviews ABSA corpus: a public corpus of patient reviews of hospitals. The initial collection had a sample of 105,636 reviews. Following cleaning and duplicate removal, there were 87,310 reviews (a decrease of 18,326, 17.3%) of 161 hospitals (18 to 962 reviews per hospital, with a mean of 542.3). Only the post-processed data is provided in the released file, the raw collection and the exact deduplication rule is not provided, and the raw count is taken from the documentation provided with the corpus. The cleaned text column is used as the model input as the column containing the original text is already normalised (lower case, many function words removed). Table 2 summarises the 22 columns.

Table 2. Structure of the preprocessed dataset (22 columns, 87,310 rows).

Column(s)	Type	Description
review_id	Integer	Unique identifier of the review



Column(s)	Type	Description
hospital	Text	Name of the reviewed hospital (161 distinct values)
original_text, clean_text	Text	Normalised text before and after final cleaning; clean_text is the model input
tokens, token_count	List, integer	Token list and its length (1 to 375 tokens)
aspects, aspect_count	List, integer	Names of mentioned aspects and their number (0 to 9)
sentiment, label_id	Text, integer	Negative (0), neutral (1) or positive (2)
split	Text	Pre-assigned partition: train, val or test
asp_Doctor … asp_No_aspect	Binary (11 columns)	Multi-label indicators for ten aspects and the no-aspect category

Sentiment is strongly skewed: 75,158 reviews (86.1%) are positive, 6,196 (7.1%) negative and 5,956 (6.8%) neutral. The corpus is pre-partitioned into 69,779 training, 8,768 validation and 8,763 test reviews, and the class proportions are almost equal (Table 3, Fig. 3). We confirmed that there is no string of cleaned text data in the training and test partitions, so there is no exact-match leakage. There are 5,841 strings that were cleaned, and that appear more than once within the whole corpus; most of them are very short and generic (such as a one-word praise).

Table 3. Number of reviews per sentiment class in each partition.

Partition	Negative	Neutral	Positive	Total
Train	4,993	4,744	60,042	69,779
Validation	581	617	7,570	8,768
Test	622	595	7,546	8,763
All	6,196	5,956	75,158	87,310

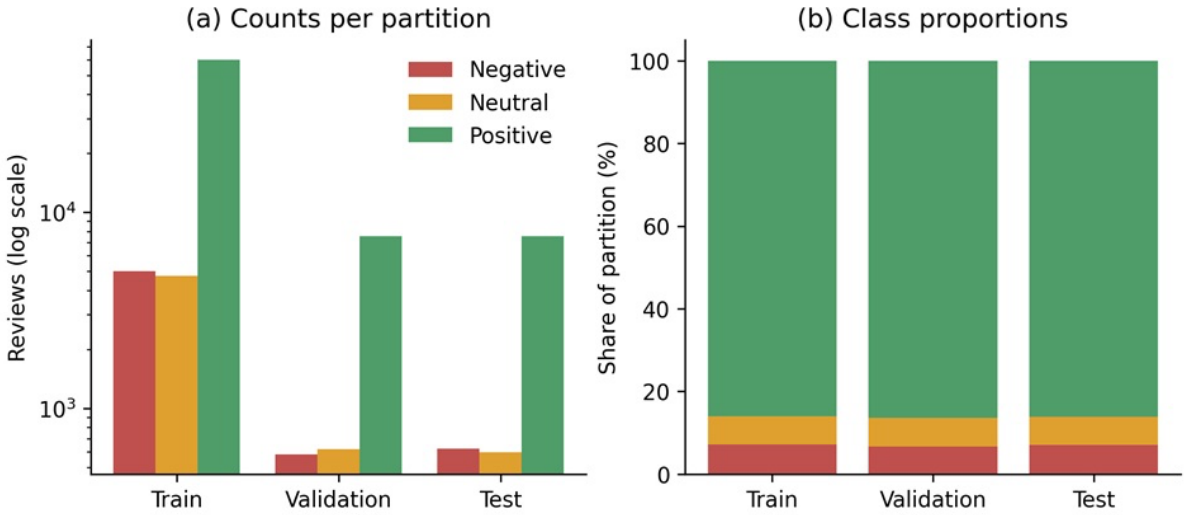


Fig. 3. Sentiment distribution in the training, validation and test partitions: (a) counts on a logarithmic scale, (b) class proportions.

The aspect annotation is a multi-label one. The average number of aspects mentioned in a review is 1.45 things, with 41.4% of reviews containing two or more aspects and 23.4% of reviews not containing any aspects and are tagged as such. The dominant occupations are doctor and nurse, and the rare ones are safety, receptionist and cost (Table 4, Fig. 4). The final column of Table 4 illustrates the proportion of negative reviews that are reported for each aspect, and it is already suggested that cost, waiting time and test procedures are the most criticized topics (Fig. 5).

Table 4. Aspect frequencies in the labelled corpus and negative-review share per aspect.

Aspect	Reviews	Share of corpus (%)	Negative share (%)
Doctor	47,045	53.88	6.40
Nurse	35,926	41.15	6.27
Parking	16,617	19.03	6.56
Waiting time	6,872	7.87	28.03
Test and examination	5,130	5.88	23.57
Food	4,546	5.21	7.39
Cleanliness	4,655	5.33	6.51
Cost	2,770	3.17	35.42
Receptionist	2,631	3.01	21.32
Safety	618	0.71	20.39
No aspect	20,455	23.43	5.89

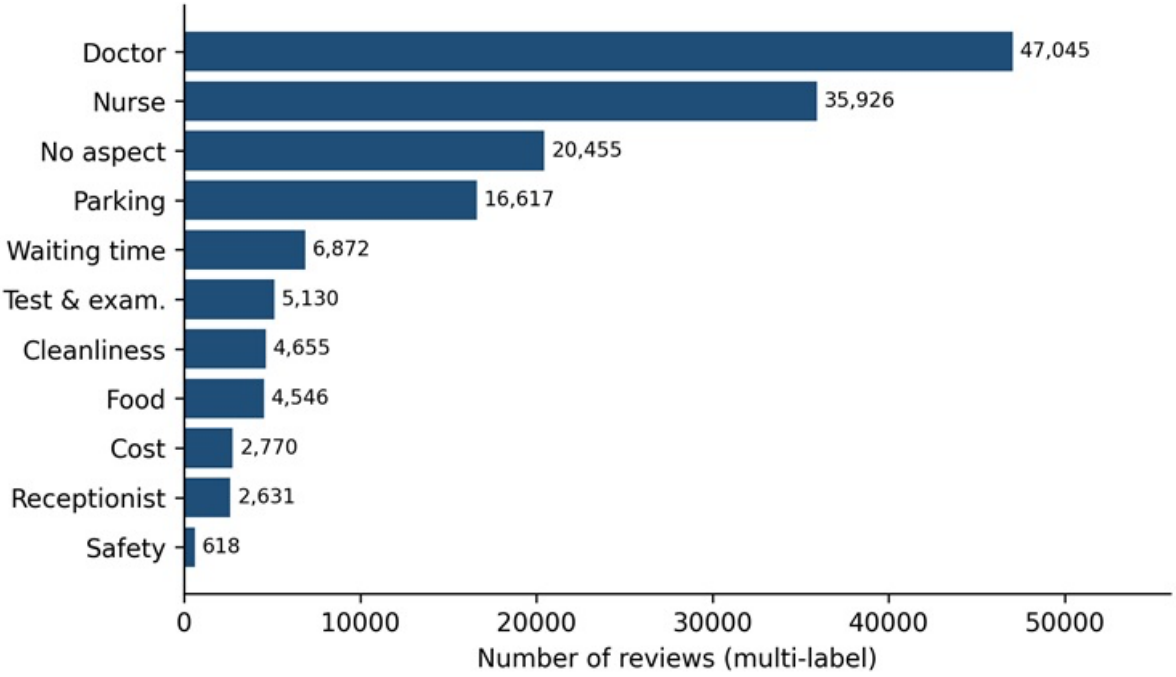


Fig. 4. Aspect frequency in the corpus (multi-label counts).

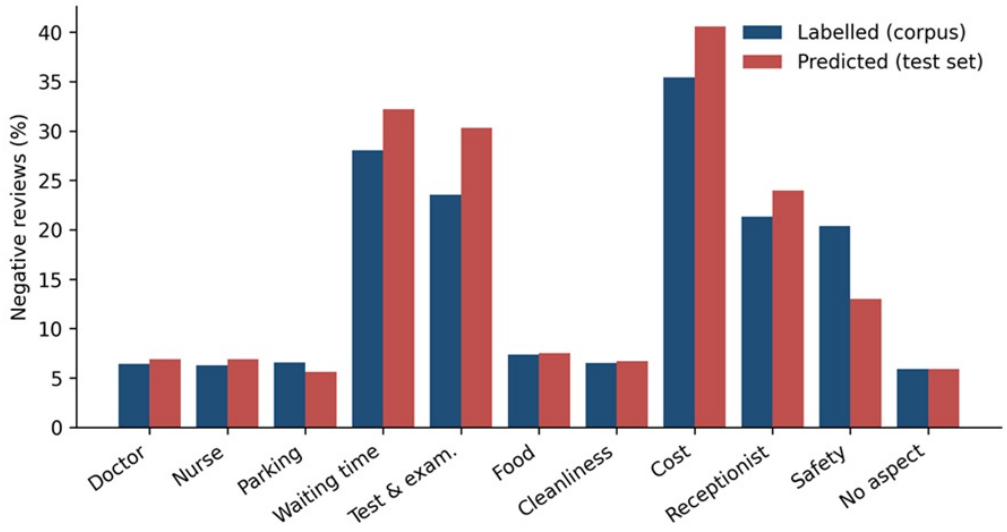


Fig. 5. Share of negative reviews per aspect: labelled share in the corpus and share predicted by the proposed pipeline on the test set



The length of reviews is short, with mean tokens of 15.1, and median 9, but with a long tail to 375 tokens. 97.1% of the reviews are not truncated when the maximum sequence length is 60 tokens. The comments are produced in an informal style in English but may also include code-mixed terms; there is no documentation of the protocol used for the annotation in the released file, which we revisit in Section 4.4.

3.3 Workflow: preprocessing, feature selection and classification

3.3.1 Preprocessing and sequence preparation

Whitespace is used to split the cleaned text. Each review is assigned an integer ID and then chopped to size L = 60 for truncation or zero padded. Padded positions are marked using a boolean mask which is not used by the recurrent layers or attention layer. The pipeline does not stem or remove stop words, as the corpus has been normalised.

3.3.2 Feature selection

The input is a raw text, so the vocabulary and length of the input are the two features that need to be controlled for feature selection. The vocabulary is created from the training partition only, so that information from the validation and test partitions will not contaminate the model. Two identifiers are reserved (padding and unknown) and words occurring with at least two occurrences are retained up to a limit of 25,000, resulting in 22,381 entries. For words outside of the frequency limit, words are mapped to the unknown identifier, which discards noisy hapax and makes the embedding table smaller.

For the linear baselines the corresponding step is a TF-IDF representation of unigrams and bigrams with a minimum document frequency of two and at most 200,000 features (Salton and Buckley, 1988) [32]. The second, positional selection is the length of the sequence, which is only 60 of the 2.9% of reviews that have a length longer than 60.

3.3.3 Network and classification

Let a review be the sequence $w_1,\ldots,w_T$ with $T \leq 60$. An embedding matrix E that is learned from scratch maps each identifier to a d-dimensional vector (d = 128):

$$\boldsymbol{x}_t = E[w_t] \in \mathbb{R}^d, \quad (1)$$

Where

w_t The word or token at position t in the review.

t The position of a token in the input sequence.

T The total number of tokens in the review, where $T \leq 60$.

E The trainable embedding matrix that maps token identifiers to dense vector representations.

$E[w_t]$ The embedding vector corresponds to token w_t.

$\boldsymbol{x}_t$ The vector representation of the token at position t.

d The dimensionality of the embedding vector, set to 128.

$\mathbb{R}^d$ The d-dimensional real-valued vector space.

$\in$ Indicates that the embedding vector belongs to the specified vector space.



A long short-term memory cell with forget gate (Hochreiter and Schmidhuber, 1997; Gers et al., 2000) [18, 22] with u = 64 hidden units updates its state as follows, where σ is the logistic function and ⊙ denotes element-wise multiplication:

$$\begin{aligned} \boldsymbol{i}_t &= \sigma(W_i \boldsymbol{x}_t + U_i \boldsymbol{h}_{t-1} + \boldsymbol{b}_i), \\ \boldsymbol{f}_t &= \sigma\big(W_f \boldsymbol{x}_t + U_f \boldsymbol{h}_{t-1} + \boldsymbol{b}_f\big), \\ \boldsymbol{o}_t &= \sigma(W_o \boldsymbol{x}_t + U_o \boldsymbol{h}_{t-1} + \boldsymbol{b}_o). \end{aligned} \quad (2)$$

$\boldsymbol{i}_t$	Input gate that controls how much new information enters the cell state.
$\boldsymbol{f}_t$	Forget gate that controls how much previous cell-state information is retained.
$\boldsymbol{o}_t$	Output gate that controls how much cell-state information contributes to the hidden state.
σ	Logistic sigmoid activation function, which maps values to the interval $(0,1)$.
W_i, W_f, W_o	Trainable weight matrices associated with the input vectors for the three gates.
U_i, U_f, U_o	Trainable recurrent weight matrices associated with the previous hidden state.
$\boldsymbol{x}_t$	Input embedding vector at the current time step.
$\boldsymbol{h}_{t-1}$	Hidden-state vector from the previous time step.
$\boldsymbol{b}_i, \boldsymbol{b}_f, \boldsymbol{b}_o$	Trainable bias vectors associated with the input, forget, and output gates.
$t-1$	The previous time step in the sequence.

$$\begin{aligned} \tilde{\boldsymbol{c}}_t &= \tanh(W_c \boldsymbol{x}_t + U_c \boldsymbol{h}_{t-1} + \boldsymbol{b}_c), \\ \boldsymbol{c}_t &= \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t, \\ \boldsymbol{h}_t &= \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t). \end{aligned} \quad (3)$$

$\tilde{\boldsymbol{c}}_t$	Candidate cell-state vector containing newly computed information.
$\boldsymbol{c}_t$	Updated cell-state vector at time step t.
$\boldsymbol{c}_{t-1}$	Cell-state vector from the previous time step.
$\boldsymbol{h}_t$	Current hidden-state vector.
W_c	Trainable input weight matrix for calculating the candidate cell state.
U_c	Trainable recurrent weight matrix for calculating the candidate cell state.
$\boldsymbol{b}_c$	Bias vector used in the candidate cell-state calculation.
tanh	Hyperbolic tangent activation function, which maps values to $(-1,1)$.



$\odot$	Element-wise multiplication of two vectors.
$+$	Element-wise addition of compatible vectors.

The bidirectional layer runs one such cell in each direction and concatenates the outputs at every position (Schuster and Paliwal, 1997) [33]:

$$\boldsymbol{h}_t = \left[\overrightarrow{\boldsymbol{h}}_t \,;\, \overleftarrow{\boldsymbol{h}}_t\right] \in \mathbb{R}^{2u}, \qquad u = 64 \tag{4}$$

$\overrightarrow{\boldsymbol{h}}_t$	Hidden-state vector generated by processing the sequence from left to right.
$\overleftarrow{\boldsymbol{h}}_t$	Hidden-state vector generated by processing the sequence from right to left.
$[\,;\,]$	Concatenation operator that joins two vectors.
u	Number of hidden units in each directional LSTM, set to 64.
$2u$	Combined output dimension of the two directions, equal to 128.
$\mathbb{R}^{2u}$	The 128-dimensional real-valued output space.

Additive attention scores each valid position, normalises the scores over the valid set V = {t : w_t is not padding} and forms the review vector r (Bahdanau et al., 2015) [8]:

$$e_t = \boldsymbol{v}^{\top} tanh(W_a \boldsymbol{h}_t + \boldsymbol{b}_a) \;\; \alpha_t = \frac{exp(e_t)}{\sum_{s \in V} exp(e_s)} \;\; \boldsymbol{r} = \sum_{t \in V} \alpha_t \, \boldsymbol{h}_t \tag{5}$$

e_t	Unnormalised attention score assigned to token position t.
$\boldsymbol{v}$	Trainable attention vector used to calculate the scalar attention score.
$\boldsymbol{v}^{\top}$	Transpose of the attention vector.
W_a	Trainable weight matrix in the attention transformation.
$\boldsymbol{b}_a$	Trainable bias vector for the attention transformation.
α_t	Normalised attention weight for token position t.
$\exp(\cdot)$	Exponential function used to calculate positive attention scores.
V	Set of valid token positions, excluding padding positions.
s	Index used to sum attention scores over valid token positions.
$\boldsymbol{r}$	Final review-level vector obtained by aggregating contextual token representations.

After dropout, a dense layer with ReLU activation of width 64 produces z. The sentiment head is a softmax over K = 3 classes and the aspect head consists of J = 11 independent sigmoid units:

$$\boldsymbol{p} = softmax(W_s \boldsymbol{z} + \boldsymbol{b}_s) \;\; q_j = \sigma\left(\boldsymbol{w}_j^{\top} \boldsymbol{z} + \beta_j\right), \quad j = 1, \ldots, 11 \; K = 3, \qquad J = 11 \tag{6}$$



$\boldsymbol{z}$ Feature vector produced by the dense layer after dropout and ReLU activation.
W_s Trainable weight matrix for sentiment classification.
$\boldsymbol{b}_s$ Bias vector for the sentiment classification layer.
$\boldsymbol{p}$ Vector of predicted probabilities for the sentiment classes.
softmax Activation function that converts sentiment logits into probabilities summing to one.
K Number of sentiment classes, set to three: negative, neutral, and positive.
q_j Predicted probability that aspect j is present in the review.
$\boldsymbol{w}_j$ Trainable weight vector associated with aspect j.
β_j Bias parameter for the j-th aspect classifier.
J Total number of aspects, set to 11.
j Index identifying an individual aspect.

3.3.4 Loss functions, class weighting and thresholds

To counter the sentiment imbalance, the loss is a categorical cross-entropy with balanced class weights [21], where N_c is the number of training reviews of class c and N the total:

$$w_c = \frac{N}{K\,N_c}\mathrm{N} \quad L_s = -\frac{1}{N}\sum_{i=1}^{N} w_{y_i} \log p_{i,y_i} \tag{7}$$

w_c Class-balancing weight assigned to sentiment class c.
N Total number of training reviews.
K Total number of sentiment classes, equal to three.
N_c Number of training reviews belonging to class c.
L_s Sentiment classification loss.
i Index of a training review.
y_i True sentiment label of review i.
p_{i,y_i} Predicted probability assigned to the true class of review i.
log Natural logarithm.
$\sum_{i=1}^{N}$ Summation over all training reviews.

For the training partition this gives 4.658 (negative), 4.903 (neutral) and 0.387 (positive). In the dedicated detector the positive term of the binary cross-entropy is multiplied by a damped weight that grows with the rarity π_j of aspect j:

$$\omega_j = \min\left\{8,\ \max\left\{1,\ \sqrt{\frac{1-\pi_j}{\pi_j}}\right\}\right\} \tag{8}$$

ω_j Imbalance weight assigned to the positive examples of aspect j.
π_j Proportion of training reviews in which aspect j is present.
$1-\pi_j$ Proportion of training reviews in which aspect j is absent.



$\sqrt{\cdot}$ Square-root function used to dampen the effect of the imbalance ratio.
$\max\{1,\cdot\}$ Ensures that the calculated weight is at least one.
$\min\{8,\cdot\}$ Limits the maximum weight to eight.

$$L_a = -\frac{1}{NJ}\sum_{i=1}^{N}\sum_{j=1}^{J}[\,\omega_j y_{ij}\log q_{ij} + (1-y_{ij})\log(1-q_{ij})] \quad (9)$$

L_a Aspect classification loss.
N Number of training reviews.
J Number of aspects, equal to 11.
i Index of a review.
j Index of an aspect.
y_{ij} Ground-truth binary label indicating whether aspect j is present in review i.
q_{ij} Predicted probability that aspect j is present in review i.
ω_j Positive-class weight for aspect j.
$1 - y_{ij}$ Indicator of the negative class when the target is binary.
$1 - q_{ij}$ Predicted probability that aspect j is absent.
NJ Product of the number of reviews and the number of aspects.

In the multi-task sentiment model the two losses are summed with equal weight (λ = 1) and the aspect loss is unweighted. After training the dedicated detector, a threshold τ_j is selected for each aspect on the validation partition by maximising its F1 score over the grid {0.10, 0.15, …, 0.90}; aspect j is assigned to a review when $q_{ij} \geq \tau_j$.

$$\tau_j \in \{0.10, 0.15, \ldots, 0.90\} \quad \tau_j^* = \underset{\tau\in\{0.10,0.15,\ldots,0.90\}}{\operatorname{argmax}} F1_j(\tau) \quad \hat{a}_{ij} = \begin{cases} 1, & q_{ij} \geq \tau_j^*, \\ 0, & q_{ij} < \tau_j^*. \end{cases}$$

τ_j Candidate decision threshold for aspect j.
τ_j^* Selected threshold that maximises the validation F1 score.
argmax Operator that returns the threshold producing the highest F1 score.
$F1_j(\tau)$ F1 score for aspect j at candidate threshold τ.
q_{ij} Predicted probability of aspect j for review i.
$\hat{a}_{ij}$ Predicted binary label for aspect j in review i.
1 Indicates that the aspect is predicted to be present.
0 Indicates that the aspect is predicted to be absent.

3.3.5 Aspect-level aggregation

Since supervision is attached to the whole review, aspect-level sentiment is estimated by aggregation. If R_j is the set of reviews in which aspect j is detected, the profile is

$$s_j(c) = \frac{1}{|R_j|}\sum_{i \in R_j} \mathbb{1}[\hat{y}_i = c], \qquad c \in \{\text{negative,neutral,positive}\} \tag{10}$$

$s_j(c)$ Proportion of reviews associated with aspect j that are predicted to have sentiment class c.

R_j Set of reviews in which aspect j is detected.

$|R_j|$ Number of reviews in the set R_j.

$i \in R_j$ Indicates that review i belongs to the set of reviews associated with aspect j.

$\mathbb{1}[\cdot]$ Indicator function that equals one when the condition is true and zero otherwise.

$\hat{y}_i$ Predicted sentiment class of review i.

c Sentiment class being considered.

The vector s_j should be read as the sentiment of reviews that talk about an aspect, not as sentiment expressed strictly towards that aspect.

3.3.6 Mathematical properties

Three short results justify design choices. **Balanced weighting:** with $w_c = N/(K \cdot N_c)$ the weighted empirical risk equals the unweighted mean of the K per-class average losses, because $N_c \cdot w_c = N/K$ for every class; a class holding 7% of the data therefore contributes exactly as much as the class holding 86%. **Masked attention:** the weights α_t are non-negative and sum to one over V, so r is a convex combination of the valid states and every coordinate lies between the minimum and maximum of that coordinate across valid positions; subtracting a large constant M from padded scores drives their weights to zero, so the representation is independent of the amount of padding. **Weights and thresholds:** for a positive weight ω on the positive term, the pointwise minimiser of the expected loss is $p^* = \omega\pi/(\omega\pi + 1 - \pi)$, which is at least τ exactly when $\pi \geq \tau/(\tau + \omega(1 - \tau))$. With $\omega > 1$ and $\tau = 0.5$ the implied posterior threshold $1/(1 + \omega)$ is below one half, so a weighted model over-predicts the aspect; raising τ_j above 0.5 restores the balance. This agrees with the validation-selected thresholds, which are at least 0.70 for nine of the eleven aspects (Table 11).

3.3.7 Algorithm

Algorithm 1: BiLSTM-Based Sentiment and Aspect Analysis

Steps
Begin 1. Load the review corpus and predefined training, validation, and test partitions. 2. Build the vocabulary using only the training texts.



Steps

Encode all reviews into token sequences.
Pad or truncate each sequence to length L = 60.

3. Calculate the class weights w_c from the training sentiment labels.
Calculate the damped positive-class weights ω_j
from the training aspect labels.

4. Train and evaluate the following four network configurations:

a. Neutral model (without sentiment supervision).
b. Sentiment model (with sentiment supervision).
c. Aspect model (without sentiment supervision).
d. BiLSTM model (with sentiment and auxiliary
aspect-classification heads).

5. Train the dedicated aspect detector using weighted binary cross-entropy loss.
For each aspect j = 1, 2, ..., J:
Select the threshold τ_j on the validation set
by maximising the aspect-specific F1 score.

6. Use the trained sentiment model to predict the sentiment label ŷ_i for each test review.

Use the dedicated aspect detector to identify the
aspects present in each test review.

7. For each aspect j = 1, 2, ..., J:
Identify the set R_j of test reviews in which
aspect j is detected.
For each sentiment class c:
Compute the aspect-level sentiment profile
$s_j(c) = (1 / |R_j|) \Sigma(i \in R_j) 1[\hat{y}_i = c]$

8. Evaluate the test predictions using:
a. Accuracy
b. Macro-F1 score
c. Matthews Correlation Coefficient (MCC)
d. ROC-AUC



Steps
e. Hamming loss f. Confusion matrices 9. Report the comparative performance of the four network configurations and the aspect detector. End

3.3.8 Complexity

The bidirectional LSTM layer has 98,816 parameters, which is 4 times an LSTM forward propagation (u(d + u) + u) = 49,408. The embedding table has 22,381 × 128 = 2,864,768 parameters, the dense layer 8,256, the attention projection 8,256 plus 64 for the scoring vector, the sentiment head 195 and the aspect head 715, which sums to 2,981,070 parameters and agrees with the count reported by the framework. The computation of the embedding is light, with about 96% of the parameters stored in the embedding table, and the time per review is O(T·u(d + u)) per direction plus O(T·u) for attention.

3.3.9 Parameter settings

Table 5. Hyper-parameters and settings used in the experiments.

Component	Setting
Sequence length / vocabulary	L = 60; 22,381 entries (minimum frequency 2, cap 25,000)
Embedding	128 dimensions, trained from scratch, spatial dropout 0.25
Recurrent layer	64 units per direction (LSTM and BiLSTM variants)
Attention / dense layers	Additive attention with 64 units; dense layer of 64 units; dropout 0.4 and 0.3
Optimiser	Adam, learning rate 0.002, halved on plateau (minimum 0.0001)
Batch size / epochs	256; at most 12 (sentiment) and 14 (aspect); early stopping on validation loss with restoration of best weights
Class handling	Balanced class weights (sentiment); damped positive weights, maximum 8 (aspect detector)
Thresholds	One per aspect, grid 0.10–0.90 step 0.05, maximising validation F1
Linear baselines	TF-IDF unigrams and bigrams (min. document frequency 2, up to 200,000 features); logistic regression (C = 4) and linear SVM (C = 0.5), both with balanced class weights

Component	Setting
Random seed / hardware	42; one CPU core, 3 GB memory; TensorFlow 2.21 and scikit-learn

Results and Discussion

4.1 Implementation details

The experiments were conducted on a single central-processing-unit core, 3 GB of memory and no graphics accelerator, all of which is quite typical for a hospital's analytics team. The neural networks were built using TensorFlow 2.21 (Abadi et al., 2016) [1] and trained using the Adam optimiser (Kingma and Ba, 2015) [24] and dropout regularisation (Srivastava et al., 2014) [36]; the linear baselines, metrics and bootstrap routines come from scikit-learn (Pedregosa et al., 2011) [27]. Matplotlib and Seaborn were used to draw the figures. All scripts are plain Python files that have been developed and run in Visual Studio Code. Each model was trained on 69,779 training reviews, tuned/early-stopped on 8,768 validation reviews and evaluated after training once on 8,763 test reviews. Only a single trial was made with random seed 42 for each neural model. Six sentiment systems are compared: TF-IDF (logistic regression), TF-IDF (linear SVM), a unidirectional LSTM, a plain BiLSTM and a BiLSTM with an auxiliary aspect head (BiLSTM-MT), and the proposed attention-pooled multi-task BiLSTM (BiLSTM-Att-MT).

4.2 Evaluation metrics

The word accuracy is misleading when applied to a very positive corpus (86.1% of positive reviews) as in He and Garcia, 2009 [21]. Thus, we report class-wise precision (P), recall (R) and F1 score, macro-averaging, weighted F1 score, Matthews correlation coefficient (MCC; Chicco and Jurman, 2020) [13] and the confusion matrix (Sokolova and Lapalme, 2009) [35]. If there are true positives TPc, false positives FPc and false negatives FNc in the class c, then:

$$P_c = \frac{TP_c}{TP_c+FP_c} \qquad R_c = \frac{TP_c}{TP_c+FN_c} \qquad F1_c = \frac{2P_cR_c}{P_c+R_c} \tag{11}$$

P_c Precision for sentiment class c.

R_c Recall for sentiment class c.

$F1_c$ Harmonic mean of precision and recall for class c.

TP_c True positives: reviews correctly predicted as class c.

FP_c False positives: reviews incorrectly predicted as class c.

FN_c False negatives: reviews belonging to class c but predicted as another class.

2 Constant used in the F1 calculation to obtain the harmonic mean.

$$Macro - F1 = \frac{1}{K}\sum_{c=1}^{K} F\,1_c \quad Accuracy = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\,[\hat{y}_i = y_i] \tag{12}$$



$\operatorname{Macro\text{-}F1}$ Arithmetic mean of the F1 scores across all sentiment classes.

K Number of sentiment classes, equal to three.

$F1_c$ F1 score for class c.

$Accuracy$ Proportion of reviews whose predicted labels match their true labels.

N Total number of evaluated reviews.

$\hat{y}_i$ Predicted sentiment label for review i.

y_i Ground-truth sentiment label for review i.

$\mathbb{1}[\hat{y}_i = y_i]$ Equals one for a correct prediction and zero for an incorrect prediction.

$$MCC = \frac{cn - \sum_{k=1}^{K} p_k t_k}{\sqrt{(n^2 - \sum_{k=1}^{K} p_k^2)(n^2 - \sum_{k=1}^{K} t_k^2)}} \quad (13)$$

MCC Matthews correlation coefficient, a measure of agreement between predicted and actual class labels.

c Number of correctly classified samples across all classes.

n Total number of samples evaluated.

K Number of classes.

k Index of a class in the multiclass calculation.

p_k Number of samples predicted to belong to class k.

t_k Number of samples that belong to class k.

$\sum_{k=1}^{K} p_k\, t_k$ Sum of the products of predicted and actual class counts.

$\sum_{k=1}^{K} p_k^2$ Sum of the squared predicted class counts.

$\sum_{k=1}^{K} t_k^2$ Sum of the squared actual class counts.

$\sqrt{\cdot}$ Square root used to normalise the numerator.



The trace of the confusion matrix c, the number of examples n, and the numbers of predicted examples of class k (pk) and true examples of class k (tk) are used in Eq. (13). We report the micro-averaged, macro-averaged and example-averaged F1 and the Hamming loss (the percentage of incorrect aspect indicators out of all N·J aspect indicators). Uncertainty is given using percentile bootstrap intervals (1,000 resamples) and paired McNemar tests with continuity correction.

4.3 Impact of each methodological component on the results

4.3.1 Overall sentiment performance

Table 6 and Fig. 6 give the test results. The proposed BiLSTM-Att-MT obtains the highest accuracy (95.77%), macro-F1 (0.8798), macro-precision (0.8730), weighted F1 (0.9582) and MCC (0.8328). This TF-IDF logistic regression, while performing well on the minority classes, has the lowest macro F1, while linear SVM is a good baseline with 93.81% accuracy.

Table 6. Sentiment classification results on the test set (8,763 reviews).

Model	Acc.	Macro-P	Macro-R	Macro-F1	Weighted F1	MCC	AUC
TF-IDF + logistic regression	0.9118	0.7471	0.8646	0.7925	0.9190	0.7057	0.9658
TF-IDF + linear SVM	0.9381	0.8173	0.8385	0.8276	0.9389	0.7569	n/a
LSTM (unidirectional)	0.9457	0.8228	0.8885	0.8528	0.9477	0.7981	0.9785
BiLSTM	0.9498	0.8409	0.8933	0.8644	0.9514	0.8106	0.9793
BiLSTM multi-task	0.9385	0.8082	0.8988	0.8458	0.9420	0.7824	0.9782
BiLSTM + attention multi-task (proposed)	0.9577	0.8730	0.8880	0.8798	0.9582	0.8328	0.9779

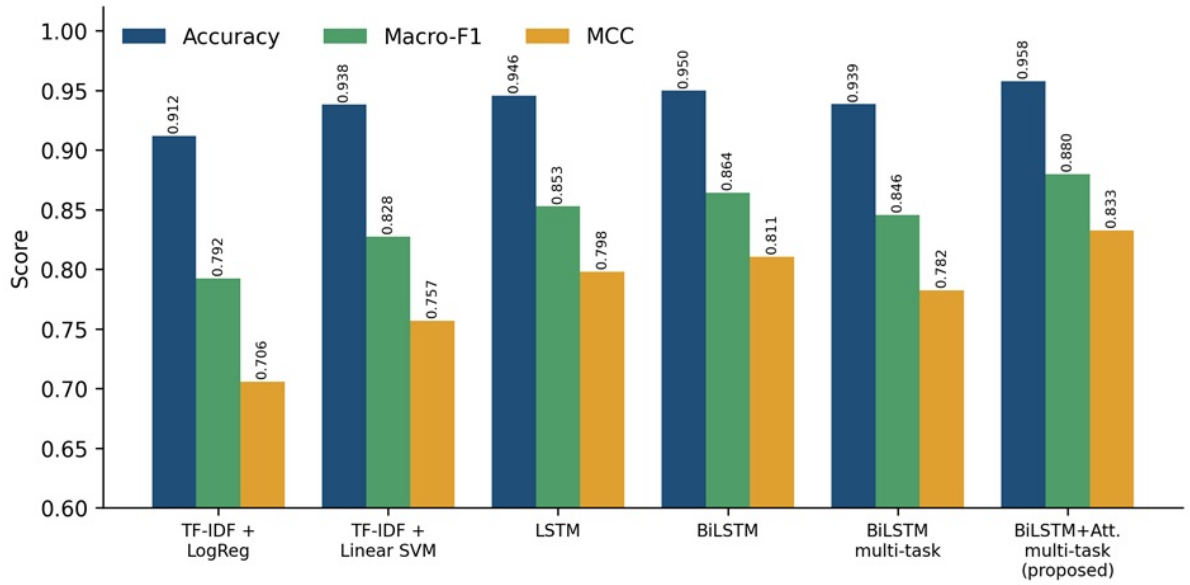


Fig. 6. Accuracy, macro-F1 and Matthews correlation coefficient of the six sentiment systems on the test set.

4.3.2 Contribution of each component

The comparison between the models is presented as an ablation chain in Table 7, as the models vary in only one component at a time. This means that by switching from the best linear model to a recurrent network, the accuracy increases by 0.76 points and the macro-F1 score by 2.52 points, indicating that the words' order contains information not provided by bags of n-grams. The accuracy gain for reading the review twice will be 0.41 points and Macro-F1 1.16 points. As multi-task learning is not necessarily beneficial, attaching an unweighted auxiliary aspect head to the BiLSTM decreases both metrics, as observed in the literature that the benefit of an auxiliary task relies on the similarity between the auxiliary task and the main task (Caruana, 1997; Ruder, 2017) [11, 31]. The combination of the same auxiliary head with attention pooling further improves the accuracy by 1.92 points and macro F1 by 3.40 points over the multi-task BiLSTM model, while the full model is 0.79 and 1.54 points better than the plain BiLSTM model. A reasonable explanation, which we cannot rule out with our experiments, is that the shared representation helps to highlight aspect-bearing words that are also informative for polarity. To test this, a model with attention but no auxiliary head would be necessary.



Table 7. Impact of each methodological component (changes in percentage points on the test set).

Step in the methodology	Comparison	Δ Accuracy	Δ Macro-F1
Sequence modelling instead of n-gram features	SVM → LSTM	+0.76	+2.52
Bidirectional reading	LSTM → BiLSTM	+0.41	+1.16
Unweighted auxiliary aspect head	BiLSTM → BiLSTM-MT	−1.13	−1.86
Attention pooling on the multi-task model	BiLSTM-MT → BiLSTM-Att-MT	+1.92	+3.40
Full proposed model versus plain BiLSTM	BiLSTM → BiLSTM-Att-MT	+0.79	+1.54
Dedicated weighted aspect detector (micro-F1 / macro-F1)	auxiliary head → detector	+17.26 (micro-F1)	+60.98 (aspect macro-F1)

In our experiments, the effect of balanced class weighting is not separated from the rest, and is corroborated analytically (Section 3.3.6) and indirectly by the excellent recall for the minority classes (Section 4.3.3), but one would have to train without class weighting to measure it directly. The same applies to the choice of words and word ordering.

4.3.3 Confusion matrices and per-class behaviour

The confusion matrix of the proposed model is given in Fig. 7. Of 622 negative reviews, 522 (83.92%) are recovered, and 505 of 595 neutral reviews (84.87%) are correct. The majority error is the classification of minority class as positive class (70 negative and 66 neutral reviews are labelled as positive) and the majority class as negative class (139 positive reviews are labelled as negative). The confusion of the matrices corresponding to the plain BiLSTM and the linear SVM are shown in Fig. 8, where the linear SVM confuses 129 of the 595 neutral reviews with the positive class, compared to 11.1% for the proposed model.

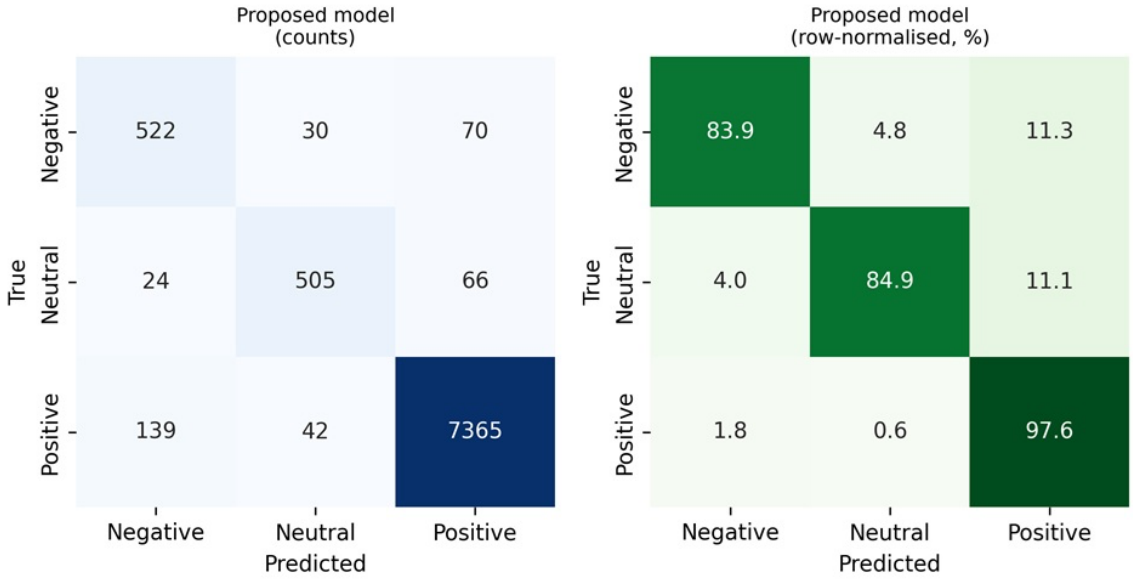


Fig. 7. Confusion matrix of the proposed model on the test set: counts (left) and row-normalised percentages (right).

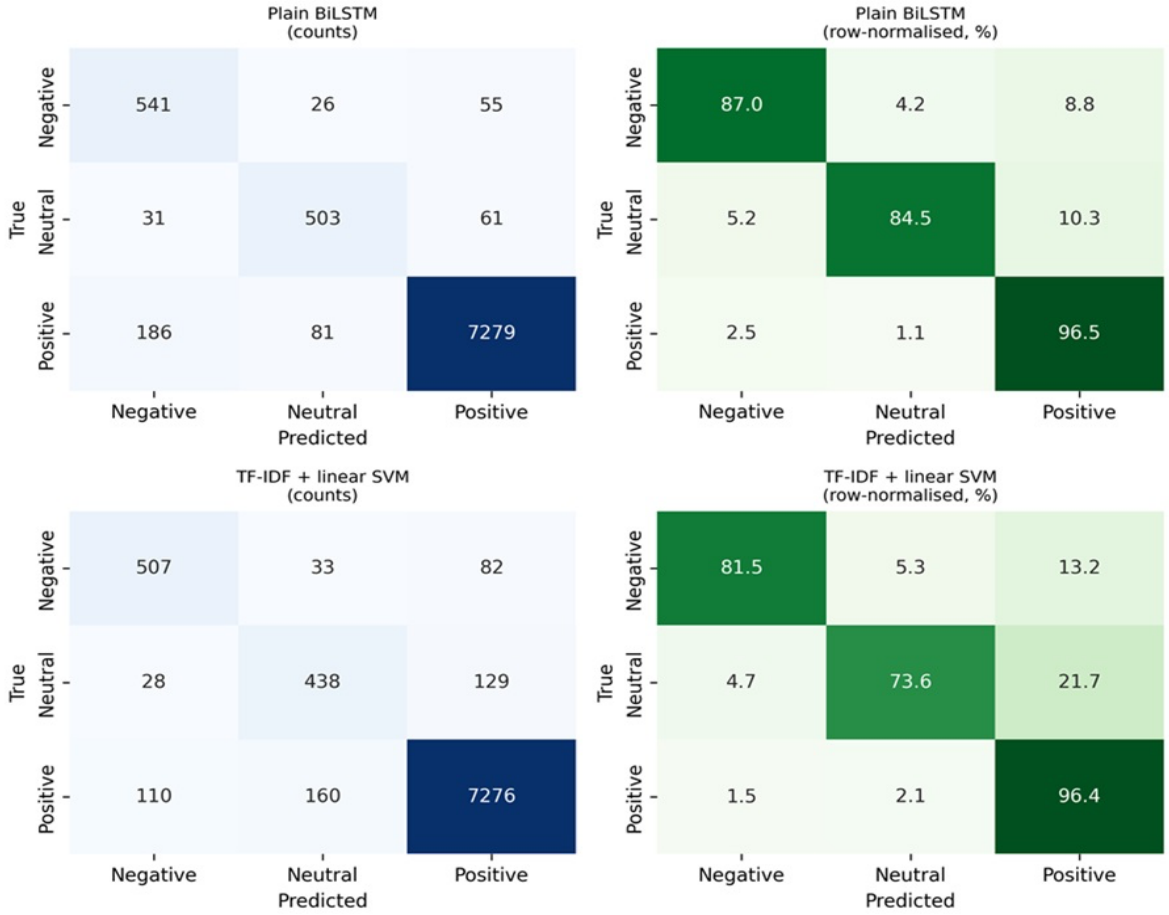


Fig. 8. Confusion matrices of the plain BiLSTM (top) and the TF-IDF linear SVM (bottom) on the test set.

Table 8. Per-class precision, recall and F1 of the proposed sentiment model.

Class	Precision	Recall	F1	Support
Negative	0.7620	0.8392	0.7988	622
Neutral	0.8752	0.8487	0.8618	595
Positive	0.9819	0.9760	0.9789	7,546

The positive class is recognised almost perfectly (F1 = 0.979). Precision (0.762) is less than recall (0.839) for the negative class, as the model is eager to classify negatives and there are many positives so there are a lot of false alarms if a small proportion are misclassified. The confusion matrices of the three models are used to calculate per-class F1 in Fig. 9. The proposed model benefits more at the neutral class (0.862 compared with 0.715 for the SVM and 0.835 for the plain BiLSTM). The advantage of the proposed model over the three models is more evident in the neutral and positive classes, as well as in the number of false alarms in the negative class, with the proposed model performing better than the SVM (0.800) and the plain BiLSTM (0.784) classes; it is not so clear in negative reviews.

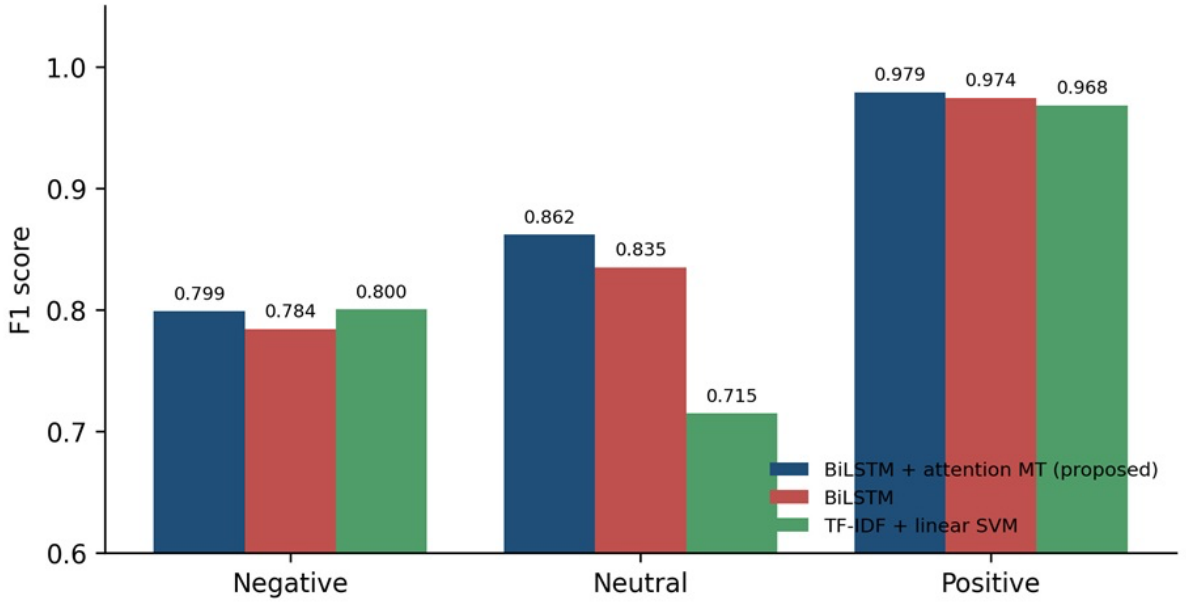


Fig. 9. Per-class F1 score of the proposed model, the plain BiLSTM and the linear SVM (computed from the test-set confusion matrices).

4.3.4 Confidence intervals and significance

Bootstrap 95% intervals are displayed in Table 9, and paired McNemar tests in Table 10. The intervals for the proposed model and the multi-task model without attention don't overlap and the proposed model is significantly better than both of the other neural variants on this test set. The difference to the plain BiLSTM is statistically significant ($p \approx 10^{-5}$) but small in size.

Table 9. Bootstrap 95% confidence intervals on the test set (1,000 resamples).

Model	Accuracy interval	Macro-F1 interval
LSTM (unidirectional)	[0.9408, 0.9504]	[0.8396, 0.8648]
BiLSTM	[0.9452, 0.9544]	[0.8523, 0.8770]
BiLSTM multi-task	[0.9336, 0.9431]	[0.8331, 0.8568]



Model	Accuracy interval	Macro-F1 interval
BiLSTM + attention multi-task (proposed)	[0.9532, 0.9620]	[0.8674, 0.8916]

Table 10. Paired McNemar tests on test-set predictions (first model of each pair is the better one).

Comparison	First only correct	Second only correct	χ^2	p-value
Att-MT vs BiLSTM	153	84	19.51	1.0e-5
BiLSTM vs LSTM	93	57	8.17	0.004
Att-MT vs LSTM	182	77	41.76	1.0e-10
Att-MT vs BiLSTM-MT	240	72	89.39	< 1e-10

These tests only show sampling variation of the finite set of tests and not variation from random initialisation. For the LSTM and BiLSTM, the earlier iteration achieved 0.9464 and 0.9469 accuracy, respectively, whereas the current version achieved 0.9457 and 0.9498 accuracy, respectively, indicating that the accuracy difference between the two runs is approximately 0.3 percentage points. Everything in between is the advantage that the bidirectional model has over the unidirectional model (0.41 points) and is just suggestive, but the higher contribution from attention is more likely to be real. More experiments with multiple seed need to be done to resolve this. The macro area under the curve of the proposed model (0.9779) is slightly lower than that of the LSTM (0.9785) and BiLSTM (0.9793), indicating that the attention model performs better at fine-grained level of decisions (as observed from F1 and MCC) rather than at the global level based on scores.

4.3.5 Aspect detection

Table 11 and Fig. 10 make a comparison of aspect detectors. The auxiliary aspect is embedded within the multi-task sentiment networks trained using an unweighted loss and a fixed threshold of 0.5, and captures the common aspects but makes very few positive predictions for six rare aspects (test and examination, food, cleanliness, cost, receptionist and safety) with a macro-F1 of only 0.33. The micro-F1 value of the TF-IDF one-versus-rest logistic regression is 0.945. The best on all four metrics, the dedicated detector, trained with damped positive weights and per-aspect thresholds, achieves micro-F1 0.962, macro-F1 0.937 and a Hamming loss of 0.0113. There is an element of weighting and also of searching for a threshold, but this was not divided up experimentally.

Table 11. Aspect detection results on the test set (upper part) and per-aspect results of the dedicated detector (lower part).

Model	Micro-F1	Macro-F1	Example-F1	Hamming loss
TF-IDF + logistic regression (one-vs-rest)	0.9445	0.8770	0.9352	0.0164



Model	Micro-F1	Macro-F1	Example-F1	Hamming loss
BiLSTM multi-task, auxiliary head	0.7901	0.3407	0.7837	0.0587
BiLSTM + attention multi-task, auxiliary head	0.7898	0.3273	0.7754	0.0595
Dedicated weighted BiLSTM + attention detector	0.9624	0.9371	0.9634	0.0113

Aspect	Precision	Recall	F1	Threshold τ
Doctor	0.979	0.959	0.969	0.40
Nurse	0.997	0.955	0.976	0.80
Parking	0.990	0.929	0.958	0.90
Waiting time	0.993	0.920	0.956	0.75
Test and examination	0.989	0.885	0.934	0.85
Food	0.966	0.914	0.939	0.75
Cleanliness	0.986	0.886	0.933	0.80
Cost	0.988	0.860	0.920	0.85
Receptionist	0.983	0.895	0.937	0.90
Safety	0.935	0.741	0.827	0.70
No aspect	0.942	0.980	0.961	0.55

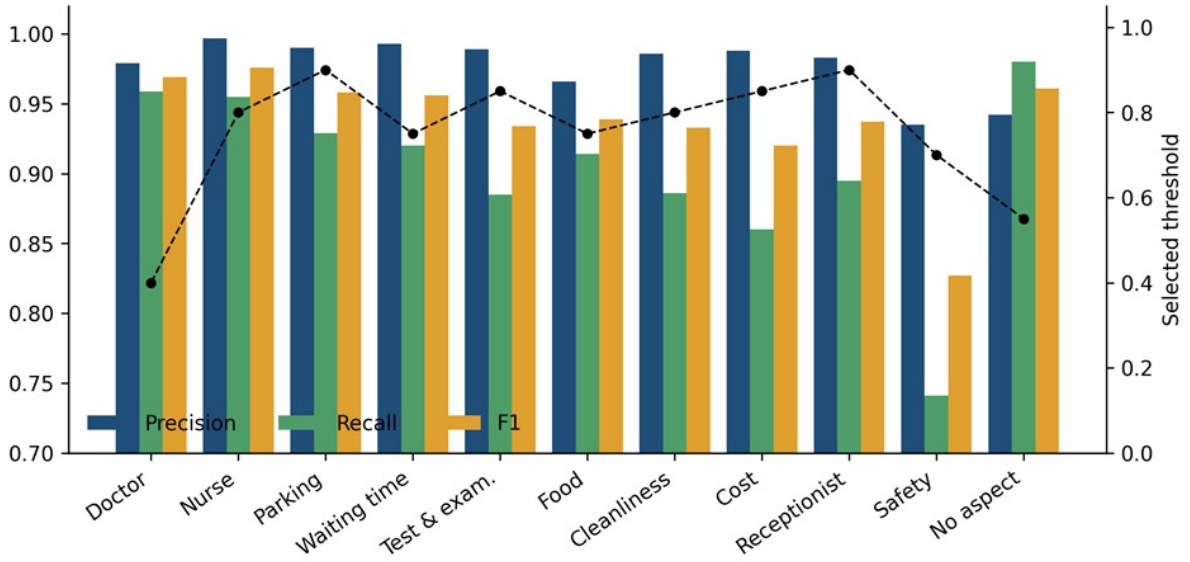


Fig. 10. Precision, recall and F1 per aspect for the dedicated detector, with the validation-selected decision thresholds (dashed line, right axis).

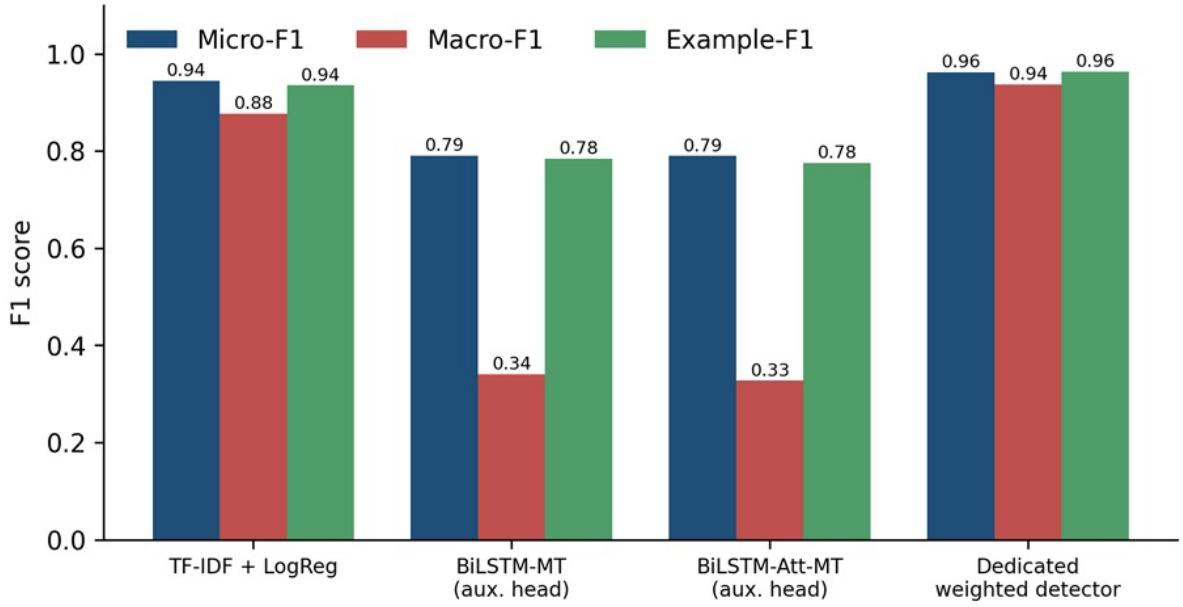


Fig. 11. Micro-, macro- and example-averaged F1 of the four aspect detectors.

Common aspects like doctor, nurse and parking have an F1 greater than 0.95. Safety is the weakest (F1 = 0.827): this has only 618 positive examples in the whole corpus; cost is the next weakest. Despite this, the F1 of the safety is still much higher than that from the linear baseline, showing the advantage of positive weighting and threshold tuning for rare aspects. Precision is greater than 0.93 for all aspects, and recall is limited, unsurprisingly, as thresholds are increased to compensate for the positive weights (Section 3.3.6).

4.3.6 Aspect-level sentiment profile

Applying Eq. (10) to test set yields the profile in table 12 and Fig. 12. The percentage of reviews which are predicted to be negative is greatest in the cost area (40.6%), followed by



waiting time (32.2%), test and examination (30.3%) and receptionist (24.0%). Negative reviews for doctors and nurses, parking, cleanliness and food only occur in around 6 to 8% of cases. The same ordering is observed in the labelled data of Table 4, which suggests that the pipeline can generate a pipeline with the same structure as the annotation. This is a narrow margin between predicted (13.0%) and labelled (20.4%) safety and should not be overinterpreted due to few test reviews.

Table 12. Predicted sentiment shares (%) among test reviews in which each aspect is detected.

Aspect	Negative	Neutral	Positive	Labelled negative share
Doctor	6.9	4.5	88.6	6.4
Nurse	6.9	2.5	90.6	6.3
Parking	5.6	1.5	92.9	6.6
Waiting time	32.2	4.8	63.0	28.0
Test and examination	30.3	3.9	65.8	23.6
Food	7.5	2.7	89.7	7.4
Cleanliness	6.7	0.7	92.6	6.5
Cost	40.6	4.8	54.6	35.4
Receptionist	24.0	5.6	70.4	21.3
Safety	13.0	0.0	87.0	20.4
No aspect	5.9	12.4	81.7	5.9

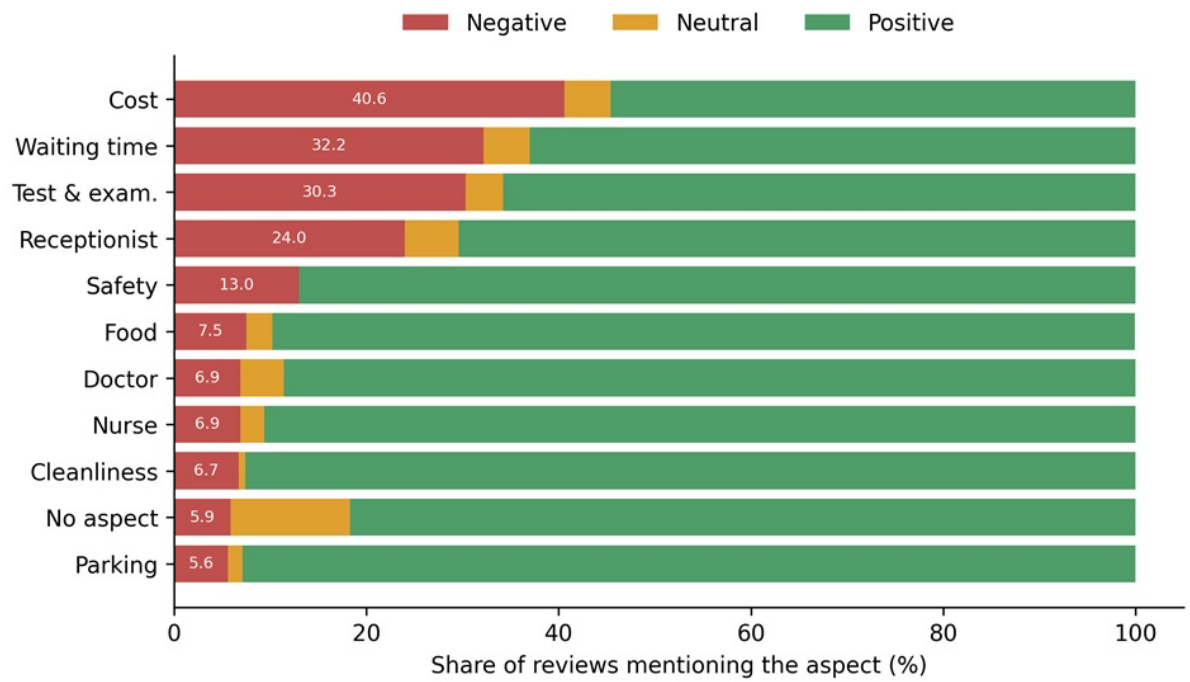


Fig. 12. Predicted sentiment profile per aspect on the test set (aspects sorted by negative share).

4.3.7 Review length, errors and efficiency

Table 13 and Fig. 13 show a high accuracy for the short reviews and the lowest accuracy for the longest reviews. In a review of more than 25 tokens, opinions will be given to multiple aspects and a single review-level label will not be able to represent them, so there will be more chances to make errors. The examination of misclassified items in a qualitative fashion reveals two types of errors: genuine errors, such as a negative comment about waiting, but predicted as neutral; and items whose label seems to contradict the text, e.g., a text starting "worst experience" and labelled as positive. Since the annotation protocol is not documented, it is unclear if such cases result from labelling based on ratings, disagreement amongst the annotators, or content that our reading fails to cover, but it appears that some of the remaining error is label noise.

Table 13. Accuracy of the proposed model by review length.

Length bucket (tokens)	Reviews	Accuracy
1–3	1,207	96.85%
4–6	1,967	97.10%
7–12	2,381	95.80%
13–25	1,963	96.18%
> 25	1,245	91.89%

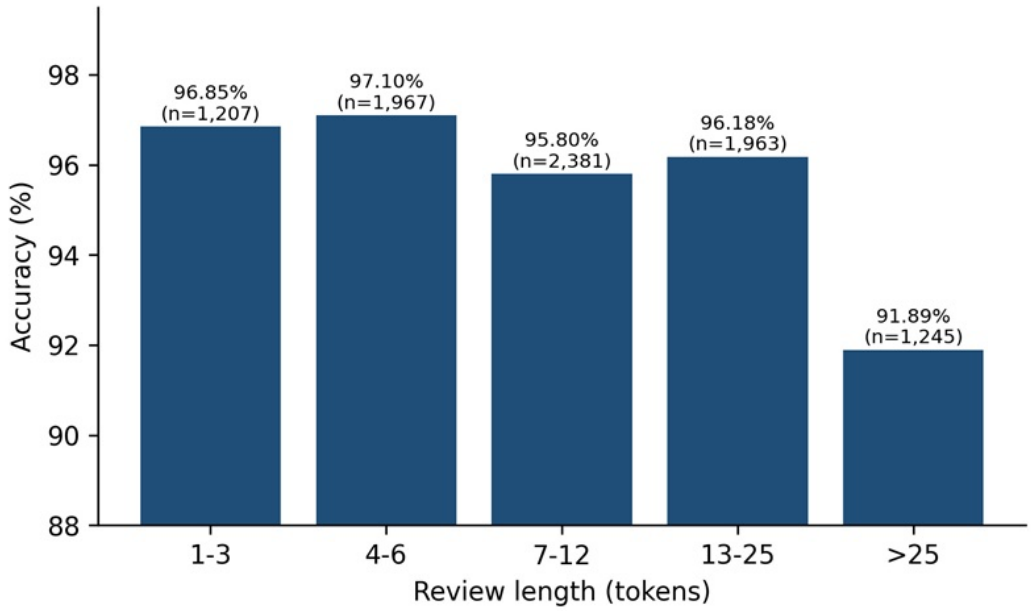


Fig. 13. Accuracy of the proposed model by review length.

In the 161 hospitals, the proportion of negative reviews varies from 0.53% to 43.26% (mean 8.31%). Hospitals are not included here for ethical reasons and those hospitals with fewer than 18 reviews have unstable percentages, however the spread of the percentages illustrates that hospital-specific dashboards are possible, providing there are confidence intervals that accompany each percentage. Regarding efficiency (Table 14, Fig. 14), the linear models train in seconds, while the proposed model requires around 6 minutes on one core, and with 2,981,070 parameters, requires only around 12 MB of weights for 32-bit precision. The bidirectional variants run two times to three times slower than the unidirectional LSTM due to the computation of two directions.

Table 14. Parameters, training epochs and training time on a single CPU core.

Model	Parameters	Epochs	Time (s)
TF-IDF + logistic regression	sparse	–	7
TF-IDF + linear SVM	sparse	–	1
LSTM (unidirectional)	2,918,531	3	126
BiLSTM	2,972,035	3	251
BiLSTM multi-task	2,972,750	4	332
BiLSTM + attention multi-task (proposed)	2,981,070	4	333

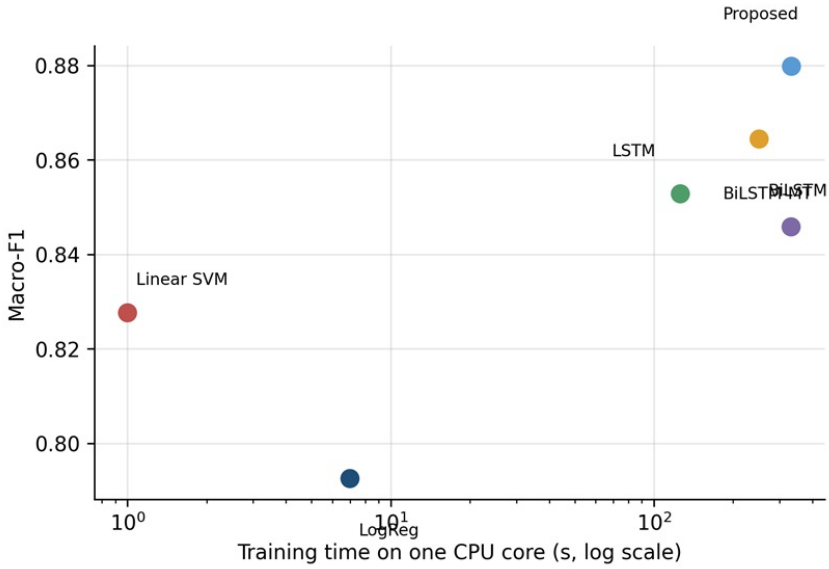


Fig. 14. Macro-F1 versus training time (logarithmic axis) for the six sentiment systems.

4.4 Discussion

The results confirm three key messages: First, on large and highly imbalanced hospital data, a compact BiLSTM (with attention pooling and balanced class weights) achieves a macro-F1 of 0.88 with no use of pre-trained knowledge, on a single CPU core. Second, the sentiment/tasks and aspect/tasks should not have a single loss: an unweighted auxiliary head suffers from rare aspects, while a dedicated, positively-weighted detector with appropriately tuned thresholds does not (macro-F1 0.937). Third, the aspect-level profile turns the predictions into something an additional manager can use, namely: dissatisfaction focuses on cost, waiting time, test procedures and reception, and less on clinical staff.

The study is compared to recent work in table 15. The scores presented in this table cannot be used for direct comparison as there are differences in corpus, language, label scheme and test protocol, and the table is a contextual map, not a ranking. It does demonstrate that our evaluation corpus is more than ten times larger than the size stated in the table for the hospital corpora, that we keep all three classes of sentiments and that our aspect detector is evaluated separately per aspect, even rare ones.

Table 15. Positioning of the present study against recent healthcare ABSA studies (contextual comparison only; data and protocols differ).

Study	**Data scale / language**	**Method**	**Headline result**
Prabhune et al. (2025) [29]	4,862 reviews, English, 25 Indian hospitals	RoBERTa + keyword layer; binary polarity	F1 0.923; accuracy 92.67%
Alamoudi and Solaiman (2025) [2]	6,000 sentences, Arabic	AraBERT, ChatGPT pseudo-labels	Sentiment macro-F1 0.84; accuracy 0.93



Study	Data scale / language	Method	Headline result
AlNasser and AlMuhaideb (2024) [5]	HEAR corpus, Arabic	MARBERT + SVM	Accuracy 92.14%
Fahmi et al. (2023) [15]	Indonesian hospital reviews	BERT-LSTM, 3 aspects × 4 states	F1 0.597
Elov and Alayev (2026) [14]	Uzbek hospital subset	BERTbek / XLM-R with entity fusion	Macro-F1 84.91; accuracy 87.19
This study	87,310 reviews, English, 161 hospitals	Attention BiLSTM + dedicated aspect detector	Accuracy 95.77%; macro-F1 0.880; aspect micro-F1 0.962

Threats to validity. (i) Labels are whole reviews, so the aspect profile reflects the sentiment of the reviews, one would expect that aspect to be mentioned in the review, rather than the sentiment towards the aspect; In a review for which the doctor is praised, and the billing is criticized, then both aspects have the same polarity. (iii) The neural models were trained only once, thus small differences cannot be concluded. (iii) The annotation protocol is not documented; the labels appearing to have a low aspect appear inconsistent with their text, if the aspect labels resulted from some keyword-based process, a high aspect score is partly a reflection of the ability to recreate that process. (iv) English texts about Indian hospitals, transfer to other regions, languages or platforms is not tested. (v) There was no assessment of a transformer, and there may be models that get a higher rating at a higher price. (vi) The fixed split is not tested on an unseen set of hospitals.

4.5 Implications

4.5.1 Practical implications

- **Targeted quality improvement.** The highest negative shares are "bill transparency," "queue management," and "test scheduling," indicating that there are additional areas of effort that do not require clinical staff but can be applied to these areas to improve the cost, waiting time and test procedures and reception.
- **Low-cost deployment.** This model, which trains in minutes on a single core, can be deployed at a hospital without graphics hardware or paid LLM interfaces, and patient text never leaves the hospital, alleviating privacy concerns.
- **Monitoring dashboards.** Negative shares can be updated weekly per hospital, or per aspect, if appropriate – and intervals should be displayed in conjunction with percentages, in case it is only a small hospital.
- **Human-in-the-loop review.** The negative class does not have as high of precision as recall, so all alerts in this class should be sent to staff for follow-up and confirmation before action, and apply to long reviews as well, where accuracy is lower.



4.5.2 Theoretical implications

- **Imbalance handled at the right level.** The analysis demonstrates that balanced weights equalise the contribution of the classes and that positive weights also lead to changing the optimal output probability, thus threshold tuning is a necessary companion; without thresholds a positive weight would over-predict rare aspects.
- **Task coupling is conditional.** The multi-task head was found to be beneficial for the plain BiLSTM and beneficial once attention was added, suggesting that the usefulness of auxiliary tasks is not intrinsic to the tasks themselves but rather relies on the architecture that implements them [11, 31].
- **Aggregation as a bridge.** The aspect-profile formulation provides a principled approach to assigning review-level labels to provide the insight into aspects, and it makes explicit what is being measured and what is not – aspect-level polarity annotation would be a direct measurement.
- **Scale matters for model choice.** Our results demonstrate that a small recurrent model can be competitive with large, pre-trained encoders, a result which confirms, but does not challenge, the transformer advantage found in benchmark studies, on large short-text and domain-specific corpora.

5. Conclusion

In this paper, we addressed the large-scale aspect-based sentiment analysis of hospital reviews using the BiLSTM model with an attention mechanism. The method comprises two branches: (1) a sentiment classifier that uses bidirectional encoding, uses masked additive attention, and applies balanced class weights; and (2) a dedicated multi-label aspect detector, which is based on damped positive weights and validation-tuned thresholds. The predictions are merged to form an aspect-level profile. Feature selection is a special vocabulary (only for training), a minimum-frequency rule, a fixed padded length with short mathematical results about weighting, attention and thresholds.

On 87,310 cleaned reviews of 161 hospitals, the sentiment model reached 95.77% accuracy, 0.880 macro-F1 and 0.833 MCC on the held-out test set, ahead of the LSTM (94.57%), the plain BiLSTM (94.98%) and the linear SVM (93.81%). The dedicated aspect detector achieved 0.962 (micro-F1) and 0.937 (macro-F1), making up for the lack of the auxiliary heads on rare aspects. The profile revealed that dissatisfaction is focused on cost (40.6%), waiting time (32.2%), test procedures (30.3%) and reception (24.0%) but not on clinical staff.

Advantages. It is accurate, cheap to use (it takes around 6 minutes of single core training), easily explained with attention weights and explicitly learns the rare classes and aspects.

Disadvantages. The improvement over the plain BiLSTM, as well as the negative class's lower precision than recall, the label noise at the review level and the learning of embeddings from scratch are limiting factors for the achieved accuracy.



Future directions. There are four paths pursued. (1) Multi-seed, cross-validated and hospital-wise split experiments to test the stability of the gains and generalisation to unseen hospitals. (2) Domain-adapted and multilingual pre-trained encoders, and a comparison with prompt-based large language models, for measuring the accuracy–cost trade-off on code-mixed text. (3) Aspect level polarity annotation or weak supervision with span level attention to get the profile for each aspect. (4) Faithful explanation methods, calibrated probabilities, drift monitoring and per-hospital uncertainty estimates for continuous quality monitoring at an even larger scale.

Acknowledgments

The authors gratefully acknowledge the academic and computational support provided by the PG & Research Department of Computer Science, Government Arts College (Autonomous), Nandanam, Chennai, Tamil Nadu, India.

References

[1] Abadi, M., Barham, P., Chen, J., et al. (2016). TensorFlow: a system for large-scale machine learning. Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, 265–283.

[2] Alamoudi, E., & Solaiman, E. (2025). EHSAN: leveraging ChatGPT in a hybrid framework for Arabic aspect-based sentiment analysis in healthcare. Proceedings of the International Conference on Intelligent Data Engineering and Automated Systems (IDEAS 2025), Springer. arXiv:2508.02574.

[3] Alhazzani, N. Z., Al-Turaiki, I. M., &Alkhodair, S. A. (2023). Text classification of patient experience comments in Saudi dialect using deep learning techniques. Applied Sciences, 13(18), 10305.

[4] Alhusseini, M. M., & Feizi-Derakhshi, M.-R. (2025). Benchmarking ChatGPT and DeepSeek in April 2025: a novel dual-perspective sentiment analysis using lexicon-based and deep learning approaches. arXiv:2509.19346.

[5] AlNasser, S., &AlMuhaideb, S. (2024). Listening to patients: advanced Arabic aspect-based sentiment analysis using transformer models towards better healthcare. Big Data and Cognitive Computing, 8(11), 156.

[6] Alsaleh, N., Alnanih, R., &Alowidi, N. (2025). Hybrid deep learning approach for automating app review classification: advancing usability metrics classification with an aspect-based sentiment analysis framework. Computers, Materials & Continua, 82(1).

[7] Azman Busst, M. B. M., Anbananthen, K. S. M., & Kannan, S. (2024). Aspect-level sentiment analysis through aspect-oriented features. HighTech and Innovation Journal, 5(1), 109–128.

[8] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. International Conference on Learning Representations (ICLR).

[9] Bansal, A., & Kumar, N. (2022). Aspect-based sentiment analysis using attribute extraction of hospital reviews. New Generation Computing, 40(4), 941–960.



[10] Basha, N. T. M., et al. (2024). A framework of attribute extraction and dependable aspect term selection from reviews of hospital websites. IAES International Journal of Artificial Intelligence, 13(3), 3456–3465.
[11] Caruana, R. (1997). Multitask learning. Machine Learning, 28(1), 41–75.
[12] Chen, H., Yang, Y., & Du, S. (2021). Research on aspect-level sentiment analysis of user reviews. Journal of Frontiers of Computer Science and Technology, 15(3), 478–485.
[13] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21, 6.
[14] Elov, B., & Alayev, R. (2026). Aspect-based sentiment analysis of hospital service reviews using BERTbek and XLM-R. SSRN preprint (abstract id 6570837).
[15] Fahmi, M., Yudanto, F., Nazhifah, N., Sari, Y., &Afiahayati. (2023). Deep learning approach for aspect-based sentiment analysis on Indonesian hospitals reviews. Universitas Gadjah Mada institutional repository (eprint 2862).
[16] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.
[17] Ganganwar, V., & Rajalakshmi, R. (2024). Employing synthetic data for addressing the class imbalance in aspect-based sentiment classification. Journal of Information and Telecommunication, 8(2), 167–188.
[18] Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: continual prediction with LSTM. Neural Computation, 12(10), 2451–2471.
[19] Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). Harnessing the cloud of patient experience: using social media to detect poor quality healthcare. BMJ Quality & Safety, 22(3), 251–255.
[20] Gui, L., & He, Y. (2021). Understanding patient reviews with minimum supervision. Artificial Intelligence in Medicine, 120, 102160.
[21] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284.
[22] Hochreiter, S., &Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.
[23] Jain, S., & Wallace, B. C. (2019). Attention is not explanation. Proceedings of NAACL-HLT, 3543–3556.
[24] Kingma, D. P., & Ba, J. (2015). Adam: a method for stochastic optimization. International Conference on Learning Representations (ICLR).
[25] Li, J., Yang, Y., Chen, R., Zheng, D., Pang, P. C. I., Lam, C. K., Wong, D., & Wang, Y. (2025a). Identifying healthcare needs with patient experience reviews using ChatGPT. PLoS ONE, 20(3), e0313442.
[26] Li, J., Yang, Y., Mao, C., Pang, P. C. I., Zhu, Q., Xu, D., & Wang, Y. (2025b). Revealing patient dissatisfaction with health care resource allocation in multiple dimensions using



large language models and the International Classification of Diseases 11th Revision: aspect-based sentiment analysis. Journal of Medical Internet Research, 27, e66344.

[27] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[28] Perikos, I., & Diamantopoulos, A. (2024). Explainable aspect-based sentiment analysis using transformer models. Big Data and Cognitive Computing, 8(11), 141.

[29] Prabhune, A. G., P S, K. P., Hari, V. R. S., Rayampalli, A., Thakur, A. S., & Yohannan, E. M. (2025). A hybrid RoBERTa–rule-based aspect-based sentiment analysis model for hospital patient reviews: development, validation, and case application in India. Journal of Health Research, 39(6).

[30] Rahim, A. I. A., Ibrahim, M. I., Chua, S. L., & Musa, K. I. (2021). Hospital Facebook reviews analysis using a machine learning sentiment analyzer and quality classifier. Healthcare, 9(12), 1679.

[31] Ruder, S. (2017). An overview of multi-task learning in deep neural networks. arXiv:1706.05098.

[32] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513–523.

[33] Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11), 2673–2681.

[34] Singh, D., Barve, S., & Dwivedi, A. K. (2025). OptiASAR: optimized aspect sentiment analysis with BiLSTM-GRU and NER-BERT in healthcare decision-making. IEEE Access, 13, 47459–47474.

[35] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437.

[36] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., &Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929–1958.

[37] Zhang, W., Li, X., Deng, Y., Bing, L., & Lam, W. (2023). A survey on aspect-based sentiment analysis: tasks, methods, and challenges. IEEE Transactions on Knowledge and Data Engineering, 35(11), 11019–11038.