Lexicon-Augmented Explainable Aspect-based Sentiment Analysis for Azerbaijani Finance and Business Text

Yusuf Evren Aykaç[1,2], Merve Özkan[1], and Refik Samet[1]

[1] Ankara University, Ankara, Türkiye
[2] Ankara Yıldırım Beyazıt University, Ankara, Türkiye
yeaykac@aybu.edu.tr

Abstract. Aspect-based sentiment analysis (ABSA) for Azerbaijani has received limited attention, and finance / business text introduces domain terms whose polarity can be context-dependent. We study a Finance / Business Azerbaijani ABSA dataset with ∼27.2K aspect-level instances (≈18 tokens on average) and examine whether the SentiAzNet polarity lexicon can complement a multilingual transformer. Using XLM-RoBERTa as the base encoder, we integrate lexicon cues in two lightweight ways: (1) concatenating token-level polarity features and (2) adding an attention bias towards lexicon-matched tokens. We also define SentiAzNet-Fin through a small set of finance-specific polarity overrides. Lexicon coverage is sparse (3.17% of tokens; ∼38% of instances have ≥1 hit), so the lexicon acts as an optional cue rather than a primary signal. Across 5-fold cross-validation, the attention-bias model with SentiAzNet-Fin reaches 0.769 accuracy and 0.738 macro-F1, improving over vanilla XLM-R by 3.6 macro-F1 points (McNemar $p < 0.001$). For analysis, we compare SHAP attributions, attention saliency, and lexicon hits; agreement between these signals tends to coincide with higher correctness. We further summarize frequent errors such as sarcasm, neutral–negative ambiguity, mixed sentiment, and code-switching.

Keywords: Aspect-based sentiment analysis · XLM-RoBERTa · financial sentiment analysis · sentiment lexicon · explainable AI · SHAP · attention

1 Introduction

Finance and business text is often short and mixed in tone: a single comment may praise *service* but complain about *fees*, or simply report a rate change without an explicit stance. In these settings, sentence-level sentiment is frequently too coarse. Aspect-Based Sentiment Analysis (ABSA) aims to address this by predicting sentiment with respect to a given aspect term or category rather than assigning one label to the whole sentence.

Most ABSA benchmarks and modeling work have focused on high-resource languages, with SemEval shared tasks providing widely used evaluation settings [18, 9, 2]. More recently, multilingual datasets have enabled cross-lingual

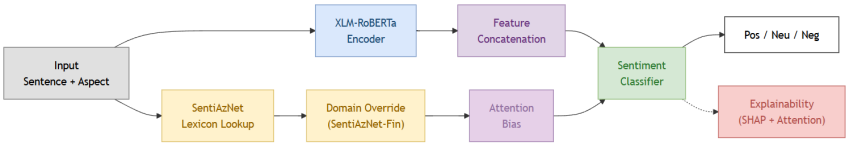

Fig. 1. Overview of the proposed ABSA pipeline with lexicon injection and explainability analysis.

evaluation [23]. Still, for low-resource languages, aspect-level supervision and domain-specific resources are often limited, and cross-lingual transfer does not fully remove data and domain gaps [12, 19].

Azerbaijani is a low-resource Turkic language where existing sentiment work has mostly targeted sentence-level classification in news and social media [10]. In finance and business, the vocabulary is more specialized, and polarity can be context-dependent; similar issues are well known in English financial NLP and motivated domain lexicons such as Loughran–McDonald [14] and domain-adapted encoders such as FinBERT [3]. While multilingual transformers like XLM-R are a strong starting point [7], neutral labels, sarcasm, and occasional code-switching still create practical difficulties [21].

In this paper, we study a Finance/Business Azerbaijani ABSA dataset with ∼27.2k aspect-level instances and ask whether an external polarity lexicon can provide helpful cues. We use SentiAzNet [5] and integrate token-level polarity information into XLM-R in two lightweight ways: feature concatenation and attention bias (Fig. 1). Because lexicon matches are sparse (3.17% of tokens; ∼38% of instances have at least one hit; Table 1), we treat the lexicon as an optional signal rather than something the model can rely on for every instance. We also define a small set of finance-specific polarity overrides (SentiAzNet-Fin) to reduce clear domain mismatches.

Finally, we use explainability analysis mainly to understand model behavior and failure cases. We compute SHAP attributions [15] and inspect attention patterns, while keeping in mind that attention weights are not guaranteed to be faithful explanations by themselves [11, 24].

Contributions. We (1) report dataset statistics and lexicon coverage for Finance/Business Azerbaijani ABSA (Table 1, Fig. 2); (2) evaluate two practical lexicon injection strategies for XLM-R and a small domain-adapted variant (SentiAzNet-Fin); (3) provide ablations and significance tests; and (4) include interpretability and error analyses that highlight where the approach helps and where it remains brittle.

2 Related Work

This section briefly reviews prior work that motivates our study. Recent surveys summarize the broader sentiment-analysis landscape [1] as well as open problems in cross-lingual ABSA [19], which are directly relevant to our setting. We

organize related work around (i) sentiment resources and transfer for Azerbaijani and other low-resource languages, (ii) financial sentiment analysis where domain lexicons are commonly used, and (iii) ABSA benchmarks and explainability methods.

2.1 Azerbaijani Sentiment Resources and Low-Resource Transfer

Azerbaijani sentiment research has mainly focused on sentence-level classification, often on news and social media, and publicly available aspect-level resources are limited [10]. In low-resource settings, multilingual pre-trained language models are commonly used because they can transfer knowledge from high-resource languages. Multilingual BERT (mBERT) [8] and XLM-RoBERTa (XLM-R) [7] are widely used starting points, but transfer quality still depends on domain match and label definitions [12, 19]. Lexicons remain relevant as lightweight priors and as tools for analysis. SentiAzNet, the polarity lexicon for Azerbaijani used in this work, was developed by our research group and has been employed in lexicon-based and hybrid sentiment pipelines [5, 4]; in this paper we build on it but treat it as an external, sparse prior rather than a complete sentiment signal.

2.2 Financial Sentiment Analysis and Domain Lexicons

Financial sentiment is often treated separately from general sentiment because domain terms can behave differently in financial contexts. In English, domain lexicons such as Loughran–McDonald were created to account for this mismatch [14]. In parallel, domain-adapted encoders (e.g., FinBERT) and fine-tuned transformer models have been explored for financial text classification [3, 17]. A recurring theme is that dictionary cues can still complement contextual models, particularly for rare or high-impact domain terms, but the benefit depends on coverage and on how well polarity definitions match the target task.

2.3 ABSA Benchmarks and Explainability

SemEval ABSA shared tasks defined common evaluation settings for aspect terms and sentiment polarity [18, 9, 2]. Multilingual ABSA datasets have since extended evaluation beyond English [23]. Recent work has also summarized cross-lingual ABSA tasks and open challenges (e.g., domain shift and annotation mismatch) [19], and explored data augmentation with large language models to improve cross-lingual ABSA under limited supervision [20]. While LLM-based and instruction-tuned approaches are increasingly competitive for ABSA, they typically rely on large models and substantial compute or API access at inference time. In contrast, our goal is a frugal, deployable alternative that augments a moderately sized multilingual encoder with a sparse lexical prior, which is attractive in low-resource settings where such resources are constrained. For explainability, SHAP is widely used as a post-hoc attribution method [15]. Attention visualizations are also common, but several studies caution that attention

weights do not necessarily reflect causal importance, motivating careful interpretation and comparisons with other signals [11, 24]. Sarcasm remains a common failure mode for sentiment systems because surface polarity cues can be misleading, and this has motivated sarcasm benchmarks and dedicated modeling work [21].

Next, we describe the dataset and evaluation setting used in our experiments.

3 Dataset

Task and instances. We focus on three-class aspect-based sentiment classification in the Finance/Business domain. Each label is attached to an aspect term within a sentence and can be written as $\langle s, a_t, a_c, y \rangle$, where s is the sentence, a_t is the aspect term, a_c is the aspect category, and $y \in \{\text{POSITIVE}, \text{NEUTRAL}, \text{NEGATIVE}\}$. Because a sentence may contain multiple aspects with different polarities, we treat each (s, a_t) pair as a separate *aspect-level instance*.

Data origin and annotation. We use an in-house Finance/Business Azerbaijani ABSA dataset curated for this study. The data is annotated at aspect level with an aspect term, an aspect category, and a three-class sentiment label (POSITIVE, NEUTRAL, NEGATIVE). Annotation was carried out by three native Azerbaijani speakers with a background in linguistics or finance, following written guidelines that defined the aspect categories and provided decision rules for ambiguous and mixed-sentiment cases. Each instance was independently labeled by two annotators, and disagreements were resolved by a third adjudicator. Inter-annotator agreement, measured on a doubly annotated portion comprising 15% of the dataset, was substantial, with a Cohen's κ of 0.772 for sentiment polarity and 0.780 for aspect-category assignment. The dataset is currently in-house; we plan to release it, or a representative subset together with the annotation guidelines, to support reproducibility, subject to source and licensing constraints.

Splits and statistics. We report descriptive statistics for a representative 70/10/20 train/validation/test partition stratified by sentiment class; model performance is reported with stratified 5-fold cross-validation (Section 5). For lexicon statistics, we report sentence-level match per split and token-level coverage over the full dataset. Table 1 summarizes the split sizes and basic statistics, and Fig. 2 shows the label distribution. Instances are short on average ($\approx$ 18 tokens with the XLM-R tokenizer), but the long tail includes sequences above 120 tokens. Token-level SentiAzNet coverage is computed over the full dataset, while experimental results are obtained using 5-fold cross-validation.

Lexicon coverage. To quantify how often external polarity cues are available, we measure overlap with SentiAzNet at token level (*coverage*) and at instance level (*match*: at least one hit). Coverage is sparse in this domain (3.17% of tokens; $\sim$38% of instances contain $\geq$1 hit; Table 1), so lexicon signals are absent for most inputs and are best viewed as optional hints rather than a complete sentiment signal.

Table 1. Statistics of the Finance/Business Azerbaijani ABSA dataset for a representative 70/10/20 split.

Statistic	Train	Validation	Test
Total samples	19016	2716	5434
Positive	8278	1183	2365
Neutral	2843	406	813
Negative	7895	1127	2256
Avg. sentence length (tokens)	18.4	18.2	18.6
Max sentence length (tokens)	128	125	131
Vocabulary size	42850	–	–
SentiAzNet coverage (token %, overall)	3.17	–	–
SentiAzNet match (sentence %)	38.2	37.8	38.5

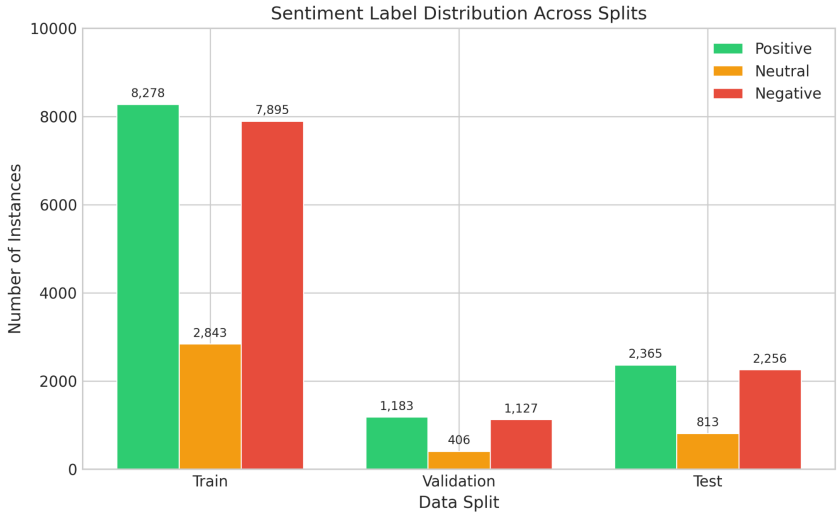

Fig. 2. Distribution of sentiment labels in train/validation/test splits.

Preprocessing. We apply minimal normalization (whitespace and punctuation) and keep stopwords to preserve negation and discourse markers for explainability. Inputs are tokenized with the XLM-R SentencePiece tokenizer. For lexicon lookup, we match surface words and propagate scores to aligned subwords (details in Section 4).

4 Methodology

Figure 1 outlines our approach. For each aspect-level instance, we encode the sentence together with the target aspect using XLM-RoBERTa [7]. In parallel, we compute a token-level polarity signal from the Azerbaijani sentiment lexicon

SentiAzNet [5]. We then inject this signal into the transformer either as an additional input feature (**P1**) or as a soft attention bias (**P2**/**P3**). Finally, we predict one of three sentiment labels (POSITIVE, NEUTRAL, NEGATIVE). We use SHAP and attention patterns mainly for post-hoc analysis (Section 6).

Aspect-conditioned input. We follow a standard paired input format where the sentence and the aspect term are concatenated and separated by special tokens: `<s>` tok(s) `</s>` tok(a_t) `</s>`, where tok($\cdot$) is the XLM-R SentencePiece tokenization. The encoder produces contextual representations for all tokens. We use the pooled representation of the first token as input to a 3-way classification head.

Lexicon signal and domain overrides. SentiAzNet provides prior polarity information for Azerbaijani lexical items. We define a scalar score l_i for each token position i. Because XLM-R operates on subword units, lexicon matching is done at the surface-word level and then propagated to the corresponding subword pieces; tokens without a match receive $l_i = 0$. To better match finance/business usage, we define *SentiAzNet-Fin* as a small list of finance-specific polarity overrides. At lookup time, if a word appears in this list, we replace its base polarity with the finance-specific value, resulting in an updated signal $\tilde{l}_i$. We keep this override set small to limit manual effort and to reduce the risk of overfitting to a few terms.

Lexicon injection via feature concatenation (P1). In feature concatenation, we treat $\tilde{l}_i$ as an additional input feature and append it to each contextual embedding produced by XLM-R. Since this increases the dimensionality by one, we apply a small linear projection to map the augmented representation back to the original hidden size before passing it to subsequent layers and the classifier. This strategy is easy to implement and serves as a lightweight form of late fusion.

Lexicon injection via attention bias (P2/P3). As an alternative, we inject lexicon information inside the self-attention computation, encouraging the model to allocate more attention mass to tokens with stronger polarity cues. Concretely, for scaled dot-product attention [22], we add a lexicon-dependent bias term before the softmax, as shown in Eq. (1):

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + B_{\mathrm{lex}}\right)V. \tag{1}$$

Here B_{lex} denotes the lexicon bias term added to the attention scores. We set $B_{\mathrm{lex}}(i, j) = \alpha \cdot m_j$, where $m_j = |\tilde{l}_j|$ and α is a learnable scalar initialized to a small value ($\alpha_0 = 0.1$). Intuitively, this acts as a soft prior: it can help when lexicon cues are present, but the model can still rely on context for the majority of tokens without lexicon matches.

Training and prediction. All models are trained with the same objective (cross-entropy over three classes) and the same aspect-conditioned input format. We report results under stratified 5-fold cross-validation and discuss ablation settings in Section 6.

Explainability signals. We use SHAP [15] to obtain token-level attributions and extract attention-based saliency from the fine-tuned transformer. Since attention is not guaranteed to be a faithful explanation by itself [11, 24], we treat it as a complementary signal. To summarize agreement, we compare top-k tokens from SHAP and attention with the set of lexicon hits using overlap measures (e.g., Jaccard@k) and relate agreement to prediction reliability (Section 6).

5 Experimental Setup

Models. We compare a lexicon-only baseline (**B1**), classical and neural baselines (**B2**–**B5**), and our lexicon-augmented XLM-R variants (**P1**–**P3**). **B1** is rule-based using only SentiAzNet signals; **B2** is a linear SVM with TF–IDF features; **B3** is a BiLSTM model with FastText embeddings [6]; **B4** fine-tunes mBERT; and **B5** fine-tunes XLM-R without lexicon information. Our proposed models add lexicon cues to XLM-R via feature concatenation (**P1**) or attention bias (**P2**), and **P3** additionally uses the finance-specific polarity overrides (SentiAzNet-Fin).

Evaluation protocol. We use stratified 5-fold cross-validation. In each fold, one split ($\approx 20\%$) is held out for testing. From the remaining data we create a small validation split for early stopping and hyperparameter selection, resulting in an overall train/validation/test ratio of approximately 70/10/20 per fold. We report mean scores across folds.

Metrics. We report Accuracy, macro-averaged Precision/Recall/F1, and Weighted-F1. Macro-F1 is treated as the primary metric because the class distribution is imbalanced and the neutral class is comparatively difficult.

Training details. Transformer-based models are fine-tuned with AdamW [13] and early stopping based on validation macro-F1. Unless stated otherwise, all models share the same preprocessing and tokenization pipeline (Section 3) and the same aspect-conditioned input format (Section 4).

Significance testing. For key comparisons we apply McNemar's test [16] on paired test predictions (correct/incorrect per instance), using continuity correction.

Explainability computation. We compute token-level SHAP attributions and attention-based saliency scores (Section 4). Since SHAP can be computationally expensive, we compute global summaries on a stratified random subset of 550 test instances (sampled to preserve the test-set class distribution) and report local explanations for selected case studies.

Table 2. Main results (5-fold cross-validation).

Model	Acc.	Macro-P	Macro-R	Macro-F1	W-F1
B1: Rule-based (SentiAzNet only)	0.412	0.358	0.341	0.349	0.398
B2: SVM + TF-IDF	0.621	0.583	0.571	0.577	0.618
B3: BiLSTM + FastText	0.658	0.612	0.604	0.608	0.652
B4: mBERT (fine-tuned)	0.714	0.681	0.673	0.677	0.710
B5: XLM-R (vanilla)	0.738	0.706	0.698	0.702	0.734
P1: XLM-R + SentiAzNet (Feature)	0.751	0.722	0.714	0.718	0.748
P2: XLM-R + SentiAzNet (Attn Bias)	0.759	0.731	0.724	0.727	0.756
P3: XLM-R + SentiAzNet-Fin (Attn Bias)	**0.769**	**0.742**	**0.735**	**0.738**	**0.770**

6 Results and Discussion

We begin with a quantitative comparison across baselines and our lexicon-augmented models, and then discuss where the gains come from using class confusions, ablations, and explainability analyses. Finally, we summarize common error patterns that remain challenging in this domain. Some of these remaining difficulties are also discussed in recent surveys of cross-lingual ABSA [19].

6.1 Overall Performance

Table 2 summarizes results averaged over 5-fold cross-validation. Among the transformer baselines, XLM-R (**B5**) is slightly stronger than mBERT (**B4**) in our setting. Adding SentiAzNet cues leads to small but consistent gains: feature concatenation (**P1**) improves macro-F1 over **B5**, while attention bias (**P2**) yields a larger improvement. The best configuration (**P3**), which combines attention bias with the finance-specific overrides (SentiAzNet-Fin), reaches 0.769 accuracy and 0.738 macro-F1. Compared to vanilla XLM-R, this is a +3.6 macro-F1 point gain, and the difference is statistically significant under McNemar's test ($p < 0.001$).

6.2 Confusions Across Classes

Figure 3(a) shows the row-normalized confusion matrix for **P3**. As expected, the neutral class is the most challenging: 61.3% of neutral instances are predicted correctly, while a substantial portion is predicted as negative (25.0%). Positive and negative classes are recognized more reliably (81.2% and 78.1% on the diagonal), but short factual statements and implicit sentiment still cause confusion, especially near the neutral–negative boundary.

Table 3. Ablation study on lexicon injection and training components.

Configuration	Macro-F1	ΔF1 (pts)	Accuracy
XLM-R (base)	0.702	–	0.738
+ SentiAzNet (Feature concat)	0.718	+1.6	0.751
+ SentiAzNet (Attention bias)	0.727	+2.5	0.759
+ Domain adaptation (SentiAzNet-Fin)	0.738	+3.6	0.769
– SentiAzNet (bias only)	0.708	+0.6	0.742
+ SentiAzNet + Focal Loss	0.735	+3.3	0.762
+ SentiAzNet + Weighted CE Loss	0.733	+3.1	0.760

6.3 Ablation Study

Table 3 breaks down which components contribute most. Feature concatenation provides a modest gain over the base encoder (+1.6 macro-F1 points), while attention bias yields a larger gain (+2.5 points). Adding finance-specific polarity overrides (SentiAzNet-Fin) gives an additional improvement, which is consistent with some polarity cues behaving differently in this domain. The "bias only" ablation (keeping the bias mechanism but removing lexicon information) yields only a small gain, suggesting that the lexicon signal itself matters more than the architectural change alone. We also tried class-imbalance-aware objectives (Focal Loss, weighted cross-entropy); in our setting, they provide improvements close to the best configuration but do not clearly surpass the domain-adapted lexicon variant.

6.4 Explainability Signals and Lexicon Alignment

We use explainability mainly to understand model behavior and failure cases, rather than to claim a single "correct" explanation. Figure 3(b) shows a SHAP feature-importance summary for frequently influential tokens in the best model. Many high-importance items are clearly sentiment laden (e.g., *yaxşı* "good", *pis* "bad", *əla* "excellent"), while others reflect common finance/business terms in our corpus (e.g., *komissiya* "fee/commission", *faiz* "interest"). This is broadly consistent with the intuition that opinions in this domain often hinge on service quality, fees, and rate-related language, but SHAP attributions remain context-dependent.

Figure 4 shows a local example. For readability, we visualize token importance over the sentence segment (the aspect segment is omitted in the plot). Attention places most weight on *yaxşıdır* (0.28) and also highlights *komissiya* (0.22) and *çoxdur* (0.20), capturing both the positive and negative cues in the sentence. SHAP attributions align with this picture: *yaxşıdır* contributes positively (+0.34), while *komissiya* and *çoxdur* contribute negatively (-0.28 and -0.22). In our ABSA setting, the final label depends on the target aspect; negative cues are expected to dominate when the aspect relates to fees, while the same sentence could be positive for a service-related aspect.

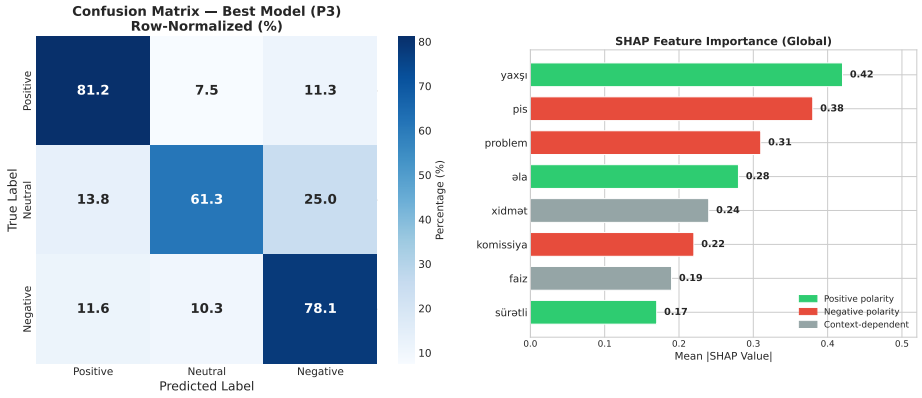

Fig. 3. Global analysis for the best model: (a) row-normalized confusion matrix and (b) SHAP feature importance (mean absolute SHAP) for frequently influential tokens.

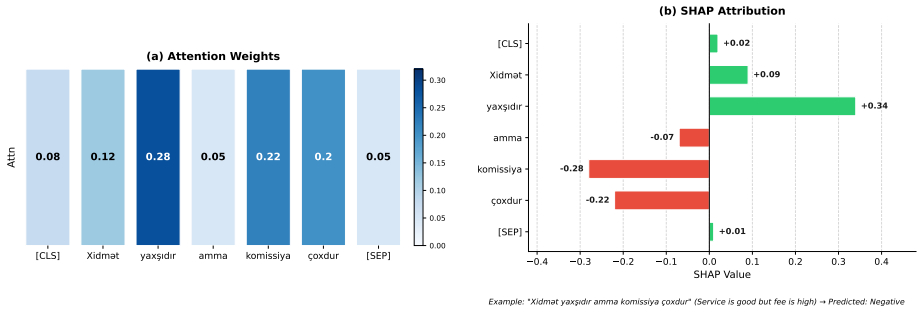

Fig. 4. Local explanation example (sentence segment shown): attention weights (left) and SHAP token-level attributions (right) for "Xidmət yaxşıdır amma komissiya çoxdur" (Service is good but fee is high).

To quantify agreement between explanation signals and lexicon hits, Table 4 reports overlap measures. SHAP and attention show moderate overlap (e.g., Jaccard@5 of 0.42), while overlap with lexicon hits is lower. Given sparse lexicon coverage, this is not surprising: the model often needs to rely on contextual cues beyond lexicon-marked tokens. The coverage buckets in the same table show that macro-F1 and average confidence increase with the number of lexicon hits, which is consistent with lexicon cues being most helpful when they are present. At the same time, most test instances (61.8%) have no lexicon hits, so overall performance still depends heavily on contextual representations learned by the encoder. Because these bucket ratios are computed over the pooled test folds in cross-validation, the implied $\geq$1-hit rate (38.2%) is close but not identical to the sentence-level match for the fixed test split in Table 1 (38.5%).

Table 4. Alignment between SHAP, attention, and lexicon signals under different lexicon coverage levels.

Metric	SHAP↔Attn	SHAP↔Lex	Attn↔Lex	3-way overlap
Jaccard@5	0.42	0.31	0.28	0.19
Jaccard@10	0.51	0.38	0.34	0.24
Kendall's τ	0.47	0.33	0.29	–
Coverage impact (CV test-fold buckets)				
≥3 lexicon hits	Macro-F1: 0.812	Conf.: 0.910	8.3% (CV tests)	
1–2 lexicon hits	Macro-F1: 0.756	Conf.: 0.840	29.9% (CV tests)	
0 lexicon hits	Macro-F1: 0.698	Conf.: 0.730	61.8% (CV tests)	

Table 5. Common misclassification patterns (illustrative examples). Frequencies are computed over the 250 manually inspected misclassified instances that could be assigned to one of these categories.

Error type	Freq.	Example (Az)	True	Pred.
Sarcasm / irony	28.3%	əla bank, hər gün yeni problem tapır	Neg	Pos
Neutral↔Negative	24.1%	Bank faiz dərəcəsini dəyişdi	Neu	Neg
Mixed sentiment	18.7%	Xidmət yaxşıdır amma komissiya çoxdur	Neg	Pos
Implicit sentiment	14.2%	Bu bankdan başqa bank yoxdur ki?	Neg	Neu
Code-switching	8.9%	Çox teşekkür ederim, harika bir hizmet	Pos	Neu
Domain jargon	5.8%	Spread artıb, yield əyrisi düzəlib	Neu	Neg

6.5 Error Analysis

We manually inspected a stratified sample of 250 misclassified instances and grouped them into common patterns (Table 5). Sarcasm/irony is a frequent source of errors, because positive words can be used in clearly negative contexts, which can mislead both lexicon cues and surface-level associations. Neutral–negative ambiguity is also common in finance text: short factual statements (e.g., policy or rate changes) may be annotated as neutral but still carry negative implications for readers. Mixed sentiment and implicit sentiment highlight that cues can be distributed across clauses and may not be explicitly linked to the target aspect. Finally, code-switching and domain-specific jargon remain challenging due to tokenization effects and limited lexicon coverage for borrowed financial terminology.

7 Conclusion

We investigated ABSA for Azerbaijani in the Finance/Business domain and examined whether an external polarity lexicon can serve as useful auxiliary information for multilingual transformers. Using a dataset of approximately 27.2k aspect-level instances, we fine-tuned XLM-RoBERTa and injected SentiAzNet

cues via feature concatenation or attention bias. Although lexicon coverage is sparse, lexicon-aware variants generally improved over the vanilla XLM-R baseline, and the strongest results came from combining attention bias with a small set of finance-specific polarity overrides (SentiAzNet-Fin). We also used SHAP attributions, attention saliency, and lexicon hits to analyze model behavior, and found that agreement between these signals often coincides with higher correctness.

Limitations and future work. Lexicon coverage is low for finance/business text, so improvements are concentrated in instances where sentiment cues are lexicalized and matched by the lexicon. The dataset used in this study is currently in-house, which limits direct external reproduction; to mitigate this we report detailed split statistics and annotation procedures (Section 3) and intend to release the data, or a representative subset with the annotation guidelines, subject to source and licensing constraints. Persistent error patterns include sarcasm/irony, neutral–negative ambiguity, mixed sentiment, and code-switching. Future work may explore expanding domain-specific lexicon coverage (including loanwords and code-switched terms), incorporating morphology-aware modeling, and adding targeted strategies for sarcasm and pragmatic cues. Another direction is to use weak supervision or data augmentation for cross-lingual ABSA, including recent LLM-based augmentation methods [20]. Human-centered evaluation of explanation usefulness would also help assess the practical value of the proposed analysis.

Acknowledgments. This study was funded by TUBITAK (The Scientific and Technological Research Council of Turkey) under Project No. 224N535.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Ali, H.M.U., Farooq, Q., Imran, A., El Hindi, K.: A systematic literature review on sentiment analysis techniques, challenges, and future trends. Knowledge and Information Systems **67**(5), 3967–4034 (2025). https://doi.org/10.1007/s10115-025-02365-x
2. Amendola, M., Cavaliere, D., De Maio, C., Fenza, G., Loia, V.: Towards echo chamber assessment by employing aspect-based sentiment analysis and gdm consensus metrics. Online social networks and media **39**, 100276 (2024)
3. Araci, D.T.: FinBERT: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063 (2019), `https://arxiv.org/abs/1908.10063`
4. Aykaç, Y.E., Özkan, M., Samet, R.: Morpheme-aware hybrid sentiment analysis for azerbaijani: A lexicon–embedding fusion with domain adaptation. IEEE Access **14**, 18828–18856 (2026). https://doi.org/10.1109/ACCESS.2026.3659042
5. Aykaç, Y.E., Samet, R., Pashayev, A., Sabziev, E.: Sentiaznet: A polarity lexicon for azerbaijani. In: 2025 10th International Conference on Computer Science and Engineering (UBMK). pp. 1684–1689. IEEE (2025). https://doi.org/10.1109/UBMK67458.2025.11207041

6. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the association for computational linguistics **5**, 135–146 (2017)
7. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 8440–8451. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.747, https://aclanthology.org/2020.acl-main.747/
8. Devlin, J., Chang, M., 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 (NAACL-HLT). pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423/
9. D'aniello, G., Gaeta, M., La Rocca, I.: Knowmis-absa: an overview and a reference model for applications of sentiment analysis and aspect-based sentiment analysis. Artificial Intelligence Review **55**(7), 5543–5574 (2022)
10. Guliyev, N., Rustamov, S., Rustamov, Z.: Analysis of public sentiment in azerbaijani news and social media. In: Proceedings of the 2024 International Conference on Computing, Machine Learning and Data Science (CMLDS '24). pp. 22:1–22:6. Association for Computing Machinery (2024). https://doi.org/10.1145/3661725.3661749
11. Jain, S., Wallace, B.C.: Attention is not explanation. 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). pp. 3543–3556. Association for Computational Linguistics (2019), https://aclanthology.org/N19-1357/
12. Lamin, N.Z., Aziz, A.A.: Cross-lingual sentiment analysis in low-resource languages: A recent review on tasks, methods and challenges. International Journal of Advanced Computer Science & Applications **16**(11) (2025)
13. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
14. Loughran, T., McDonald, B.: When is a liability not a liability? textual analysis, dictionaries, and 10-ks. The Journal of Finance **66**(1), 35–65 (2011). https://doi.org/10.1111/j.1540-6261.2010.01625.x
15. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. vol. 30, pp. 4765–4774 (2017), https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
16. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika **12**(2), 153–157 (1947). https://doi.org/10.1007/BF02295996
17. Nasiopoulos, D.K., Roumeliotis, K.I., Sakas, D.P., Toudas, K., Reklitis, P.: Financial sentiment analysis and classification: A comparative study of fine-tuned deep learning models. International Journal of Financial Studies **13**(2), 75 (2025). https://doi.org/10.3390/ijfs13020075
18. Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., et al.: Semeval-2016 task 5: Aspect based sentiment analysis. In: Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016). pp. 19–30 (2016)

19. Šmíd, J., Král, P.: Cross-lingual aspect-based sentiment analysis: A survey on tasks, approaches, and challenges. Information Fusion **120**, 103073 (2025). https://doi.org/10.1016/j.inffus.2025.103073
20. Šmíd, J., Přibáň, P., Král, P.: LACA: Improving cross-lingual aspect-based sentiment analysis with LLM data augmentation. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 839–853. Association for Computational Linguistics, Vienna, Austria (2025). https://doi.org/10.18653/v1/2025.acl-long.41, https://aclanthology.org/2025.acl-long.41/
21. Suhartono, D., Wongso, W., Tri Handoyo, A.: Idsarcasm: Benchmarking and evaluating language models for indonesian sarcasm detection. IEEE Access **12**, 87323–87332 (2024). https://doi.org/10.1109/ACCESS.2024.3416955
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30, pp. 5998–6008 (2017), https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
23. Wu, C., Lin, Z., Liu, Y., He, C., Yu, D., Yamada, I., Zhou, J., Miyao, Y.: M-ABSA: A multilingual dataset for aspect-based sentiment analysis. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 2525–2541. Association for Computational Linguistics, Miami, Florida, USA (Nov 2025). https://doi.org/10.18653/v1/2025.emnlp-main.128, https://aclanthology.org/2025.emnlp-main.128
24. Zhu, Z., Chen, H., Ye, X., Lyu, Q., Tan, C., Marasović, A., Wiegreffe, S.: Explanation in the era of large language models. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts). pp. 19–25 (2024)