Soft-Aspect ABSA: A Probabilistic Framework with Cluster Stability and Class-Imbalance Diagnostics

Samson Mayeem *, Benjamin Tei Partey , Godson Rashid Dawuni , Osei-Wusu Augustine

Posted Date: 27 May 2026

doi: 10.20944/preprints202605.1878.v1

Keywords: aspect-based sentiment analysis; topic modelling; spectral clustering; non-negative matrix factorisation; weak supervision; class imbalance; focal loss; cluster stability; low-resource NLP; fintech

 doi:10.20944/preprints202605.1878.v1

Article

Soft-Aspect ABSA: A Probabilistic Framework with Cluster Stability and Class-Imbalance Diagnostics

Samson Mayeem *, Benjamin Tei Partey, Godson Rashid Dawuni and Osei-Wusu Augustine

Department of Computer Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

* Correspondence: smayeem1@st.knust.edu.gh

Abstract

Aspect-based sentiment analysis (ABSA) is increasingly the granularity at which customer feedback is consumed, and recent work has pushed the field rapidly toward transformer- and graph-based architectures [3,5–7]. However, most modern ABSA approaches assume either a closed manually curated aspect taxonomy or a fully supervised aspect extractor trained on benchmark corpora such as SemEval. Neither assumption holds in low-resource emerging-market settings, where aspects must be discovered from the corpus itself, annotation budgets are negligible, and class distributions can be unexpectedly skewed. This article introduces Soft-Aspect ABSA, a probabilistic, topic-model-agnostic framework that promotes unsupervised topic-model output to first-class aspects via a temperature-controlled softmax over topic-membership posteriors. We instantiate the framework with a spectral-clustering plus non-negative matrix factorisation (NMF) substrate on a corpus of 292 Google Play Store reviews of a Ghanaian retail-bank mobile application (April–September 2024). The corpus exhibits an inverted class imbalance (30.8% positive / 69.2% negative under a keyword-bootstrap rule) and a four-cluster topic decomposition. A baseline TF-IDF embedding head trained with binary cross-entropy collapses to the majority class on the held-out test set: accuracy 0.6949, minority-class F1 0.000, Matthews correlation 0.000, despite a ROC-AUC of 0.934 that indicates well-ranked probabilities. The framework licenses two closed-form remediations — class-weighted cross-entropy and focal loss [28] — that we evaluate empirically on the same head. Focal loss with $\gamma = 1$ lifts minority-class F1 from 0.000 to 0.818, Matthews correlation from 0.000 to 0.746, and ROC-AUC to 0.986, demonstrating that the framework correction is not merely formal but is recoverable on the case-study data. We also run a bootstrap stability protocol for cluster-count selection (B = 50) that flags the silhouette-max $k^* = 4$ as only moderately stable ($I_{stab} = 0.64$). The contribution is methodological: a reusable scaffold for low-resource ABSA pipelines in which the aspect set is not given a priori.

Keywords: aspect-based sentiment analysis; topic modelling; spectral clustering; non-negative matrix factorisation; weak supervision; class imbalance; focal loss; cluster stability; low-resource NLP; fintech

1. Introduction

Aspect-based sentiment analysis (ABSA) refines the coarser task of document-level polarity classification by decomposing a review into the (aspect, polarity) pairs the reviewer is in fact asserting. Recent surveys document the rapid maturation of the subfield: Zhang et al. [3] catalogue the dominant task families (aspect extraction, aspect-opinion pair extraction, aspect sentiment triplet extraction); Brauwers and Frasincar [34] provide a complementary classification-centred view; and Haznitrama et al. [4], writing in AI Open, give a comprehensive taxonomy of compound ABSA pipelines and their evaluation benchmarks. The frontier of empirical performance has moved rapidly toward syntactic-graph and prompt-augmented transformer methods. Niu et al. [5] propose an adaptive structure induction framework that learns latent dependency graphs from a spectral perspective; Liang et al. [35] fuse dependency and constituency parses in their BiSyn-GAT+ model;

 doi:10.20944/preprints202605.1878.v1

Aziz et al. [6] integrate BERT with a multi-layered graph convolutional network for unified aspect–sentiment extraction; Feng et al. [7] augment graph attention with prompt-based encoding in DSGP; and Xu et al. [8] demonstrate ChatGPT-based augmentation for contrastive ABSA training. The dominant evaluation setting in this line of work is supervised — SemEval-2014 through SemEval-2016 restaurant and laptop reviews — with manually curated aspect taxonomies and tens of thousands of token-level annotations.

This setting is at odds with many practical deployments. In emerging-market fintech, in health-app feedback streams, in citizen-engagement platforms, and across most low-resource African-language settings, three properties hold simultaneously: (a) the aspect set is not known a priori and must be discovered from the data; (b) gold-annotation budgets are negligible, so labels must be bootstrapped from weak supervision such as keyword rules, distant supervision, or off-the-shelf lexicons; and (c) the natural class distribution is heavily skewed. The skew is most commonly toward positive feedback in app-store corpora overall [2], but as the empirical study reported here demonstrates, complaint-heavy single-app corpora can exhibit the opposite imbalance — a fact that any practical pipeline must accommodate.

Each of these properties has been studied in isolation. Aspect discovery from unsupervised topic models has a long lineage running from Hu and Liu [1] through joint sentiment-topic models such as Lin and He [12] and Mei et al. [13], and through their neural-era successors such as the joint aspect-sentiment topic embedding model of Huang et al. [9]; non-negative matrix factorisation (NMF) [17,18] and its graph-regularised variants [19] remain widely used substrates for short-text topic discovery, with newer entrants including the embedded topic model (ETM) of Dieng et al. [14], BERTopic [15], and contextualised topic models [16]. Weak supervision via labelling functions has been formalised by the Snorkel line of work [23]. Class-imbalance correction has matured from cost-sensitive training and SMOTE [24] through ADASYN [25], k-means SMOTE [31], focal loss [28], class-balanced loss [29] and label-distribution-aware margin loss [30] up to recent surveys [27,32].

What is comparatively thin is a single framework that wires these three concerns together and that treats the unsupervised topic model not as a descriptive pre-processing artefact but as a first-class aspect substrate whose membership posteriors flow forward into a per-aspect sentiment head. This article proposes such a framework, Soft-Aspect ABSA, and demonstrates it on a concrete low-resource testbed: a corpus of Google Play Store reviews of a Ghanaian retail-bank mobile application.

1.1. Contributions

This article makes three methodological contributions.

A topic-model-agnostic Soft-Aspect ABSA framework. We formally specify a probabilistic ABSA pipeline in which any topic discovery method that produces document-level latent coordinates can serve as the aspect substrate. Aspect-membership posteriors are obtained by a temperature-controlled softmax over cluster centroids in the latent space, and aspect-conditional polarity is predicted by per-aspect heads. The framework admits any modern topic substrate — LDA [11], BERTopic [15], contextualised topic models [16], or graph-regularised NMF [19] — as a drop-in replacement for the spectral-NMF substrate used here.

A bootstrap stability protocol for cluster-count selection. We specify a resampling procedure that returns a distribution over k^*, summarised by its mode and a stability index $I_{stab} \in [0,1]$. On the case-study corpus the silhouette point estimate ($k^* = 4$) achieves $I_{stab} = 0.640$ over $B = 50$ bootstrap replicates — moderate stability, signalling that the substrate is sensitive to corpus resampling on small corpora and that the silhouette point estimate alone is an insufficient diagnostic for the cluster-count decision.

A class-imbalance diagnostic battery for weak-supervision sentiment pipelines, with empirical remediation. We document the canonical failure mode of bootstrap-labelled imbalanced sentiment classifiers — a constant-majority-class predictor with accuracy equal to the class prior and zero recall on the minority class — and we empirically demonstrate that the closed-form class-weighted [29] and

 doi:10.20944/preprints202605.1878.v1

focal-loss [28] remediations the framework licenses recover minority-class performance on the case-study corpus, lifting Matthews correlation from 0.000 (baseline BCE) to 0.746 (focal loss, $\gamma = 1$).

1.2. Scope

The empirical strand is a single-corpus case study on n = 292 reviews of one Ghanaian retail-bank application over a five-month window. We do not claim a new state-of-the-art sentiment classifier; we do not evaluate a transformer baseline; and the framework's generalisation across topic-model substrates is identified as future work. The contribution is the framework, the stability protocol, the diagnostic battery, and the empirical demonstration that the corrections the framework derives in closed form do recover minority-class performance on real low-resource data.

2. Materials and Methods

2.1. The Soft-Aspect ABSA Framework

2.1.1. Problem Formulation and Notation

Let $D = \{d_1, \ldots, d_n\}$ be a corpus of n reviews drawn from a vocabulary V of size m. Each document d_i is associated with a latent representation $x_i \in \mathbb{R}^m$ (for example, an L2-normalised TF-IDF vector, but in principle any document representation suffices). Let $A = \{a_1, \ldots, a_K\}$ be an aspect set of size K, and let $s_{ik} \in \{-, +\}$ be the polarity that review d_i asserts about aspect a_k, defined only when d_i is in fact discussing a_k. Document-level sentiment analysis collapses A to a single implicit aspect (the document as a whole); aspect-based sentiment analysis [3,34] outputs the (K × 2) aspect-polarity table per review.

2.1.2. Topic-Conditioned Aspect-Membership Posteriors

The framework's central commitment is that the aspect set A is obtained as the output of an unsupervised topic discovery procedure rather than a manually curated taxonomy. Let $U \in \mathbb{R}^{n \times K}$ be the matrix of document-level latent coordinates produced by the topic substrate (the spectral coordinates in the present instantiation; embedding centroids under BERTopic [15] or contextualised topic models [16] in alternative instantiations), and let $\mu_k \in \mathbb{R}^K$ be the centroid of cluster k in the same latent space. The aspect-membership posterior is defined as a temperature-controlled softmax over the inner products of the document coordinates with the cluster centroids:

$$P(z_{ik} = 1 \mid x_i) = \exp(U_i^T \mu_k / \tau) / \Sigma_j \exp(U_i^T \mu_j / \tau), \quad (1)$$

where $z_{ik} \in \{0, 1\}$ indicates whether document d_i mentions aspect a_k, and $\tau > 0$ is a temperature parameter. Setting $\tau \to 0$ recovers a hard cluster assignment; setting $\tau \to \infty$ recovers a uniform distribution over aspects. The default $\tau = 1$ produces a soft assignment in which each document may discuss multiple aspects with non-trivial probability — empirically the right model for review text, where a single review can complain about one feature and praise another in the same paragraph. The probabilistic-membership formulation is a deliberate methodological choice that places the framework alongside joint sentiment-topic models [12,13] and their neural-era successors [9] in the lineage of probabilistic aspect models, while retaining the modularity to plug in modern topic-model substrates.

2.1.3. Aspect-Conditional Polarity Heads

Given the aspect-membership posterior, the aspect-conditional polarity is predicted by per-aspect classifier heads. For each aspect a_k we maintain a sigmoid head with parameters (w_k, b_k):

$$P(s_{ik} = + \mid x_i, z_{ik} = 1) = \sigma(w_k^T \varphi_k(x_i) + b_k), \quad (2)$$

 doi:10.20944/preprints202605.1878.v1

where φ_k(x_i) is an aspect-specific feature transform and σ is the logistic sigmoid. Each head is trained only on documents for which P(z_ik = 1 | x_i) exceeds a threshold τ_z (default 0.5), so each head sees a per-aspect class prior that is typically less skewed than the global prior.

2.1.4. Class-Weighted and Focal-Loss Corrections

The polarity heads are trained by default with binary cross-entropy. On imbalanced corpora this loss is dominated by the majority class, an effect catalogued in the foundational survey of He and Garcia [26] and updated in the recent comprehensive reviews of Guo et al. [27] and Liu et al. [32]. The framework licenses two closed-form corrections. First, class-weighted cross-entropy reweights the per-example contribution by the inverse class frequency:

$$L_w(\theta) = -(1/n)\ \Sigma_i\ [\alpha_1\ y_i \log f_\theta(x_i) + \alpha_0\ (1 - y_i) \log(1 - f_\theta(x_i))], \quad (3)$$

with $\alpha_c \propto n / (2\ n_c)$. The effective-number-of-samples reweighting scheme of Cui et al. [29] is a near-equivalent closed-form alternative; the label-distribution-aware margin (LDAM) loss of Cao et al. [30] is a margin-based cousin that can be combined with either reweighting. Second, focal loss [28] further down-weights easy (high-confidence) majority examples:

$$L_{focal} = -(1/n)\ \Sigma_i\ \alpha_{c_i}\ (1 - p_i)^\gamma \log p_i, \quad (4)$$

where p_i is the predicted probability assigned to the true class and $\gamma \geq 0$ is the focusing parameter. $\gamma = 2$ is the literature default; we evaluate $\gamma \in \{1, 2, 5\}$ in Section 3. We do not evaluate data-level remediations (SMOTE [24], ADASYN [25], k-means SMOTE [31]) in the case-study run, but these are natural orthogonal additions to the loss-level corrections and are flagged as future work.

2.1.5. Bootstrap Stability Protocol

The framework requires a choice of cluster count K. The default heuristic — sweep K over a candidate range and select the K that maximises the average silhouette coefficient [21] — returns a single point estimate that is fragile to corpus resampling on short, noisy text. The framework prescribes a bootstrap protocol: the corpus is resampled with replacement B times. For each replicate $b \in \{1, \ldots, B\}$ the full topic substrate is re-run and the silhouette-maximising k_b^* is recorded. The bootstrap distribution is summarised by its mode and a stability index:

$$I_{stab} = (1/B)\ \Sigma_{b=1}^{B}\ 1[k_b^* = \text{mode}(k_*)] \in [0, 1], \quad (5)$$

which approaches 1 when the cluster count is robust to resampling.

2.2. Substrate Pipeline for the Case Study

We instantiate the framework with a spectral-clustering plus NMF substrate. The substrate is interpretable (NMF post-processing yields per-cluster word loadings) and cheap enough to fit repeatedly under the bootstrap protocol. BERTopic [15], embedded topic models [14] or contextualised topic models [16] would be drop-in upgrades and are identified in Section 5 as priority extensions. Figure 1 gives the end-to-end view of the substrate.

 doi:10.20944/preprints202605.1878.v1

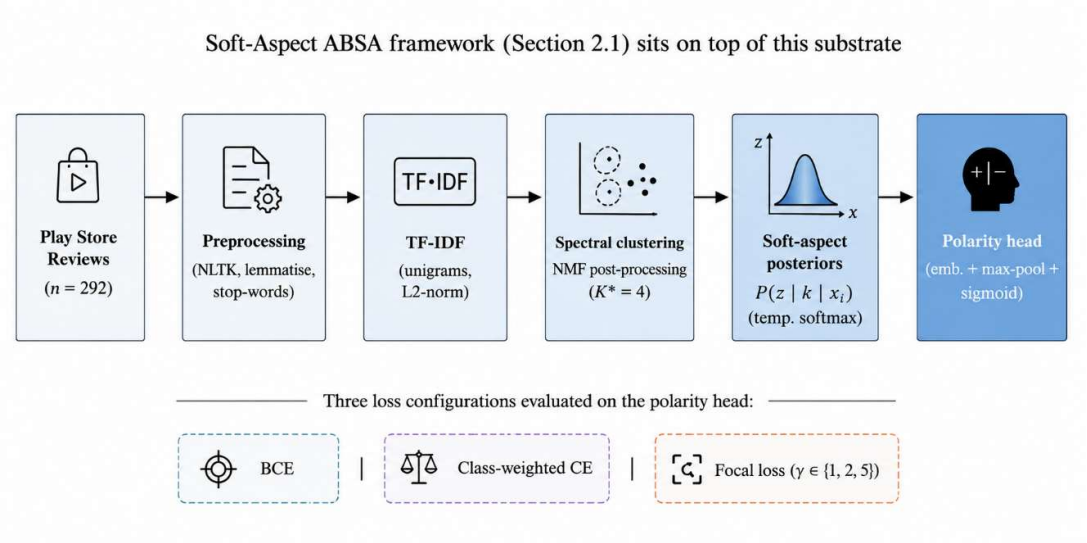

Figure 1. End-to-end Soft-Aspect ABSA substrate used in this study. Raw Play Store reviews are tokenised, lemmatised, vectorised with TF-IDF, partitioned by spectral clustering of the similarity matrix $S = X X^{T}$ with NMF post-processing per cluster, and consumed by an embedding-plus-pooling supervised polarity head. The Soft-Aspect framework of Section 2.1 sits on top of this substrate; the same diagram applies with any topic-modelling component substituted for the spectral-NMF block.

2.2.1. Preprocessing

Each review string was lower-cased; punctuation and non-alphabetic tokens were removed; the text was tokenised with the Natural Language Toolkit (NLTK); English stop-words were dropped; surviving tokens were lemmatised with the WordNet lemmatiser. Reviews shorter than three tokens after preprocessing were dropped.

2.2.2. TF-IDF Vectorisation

Cleaned reviews were vectorised with a TF-IDF transform [22] using unigram features and sublinear term-frequency scaling, with min_df = 2 and max_df = 0.95. Document vectors were L2-normalised so that cosine and inner-product similarity coincide.

2.2.3. Spectral Clustering with NMF Post-Processing

A symmetric similarity matrix $S = X X^{T}$ was constructed. The normalised graph Laplacian $L_{sym} = I - D^{-1/2} S D^{-1/2}$ (with $D = \mathrm{diag}(S1)$) was eigendecomposed [20]; the first K eigenvectors of L_{sym}, stacked as columns of $U \in \mathbb{R}^{n \times K}$, were row-normalised and clustered with k-means. Within each cluster, NMF [17,18] was fitted on the cluster's sub-corpus and the ten highest-loading vocabulary terms were taken as the cluster's representative words.

2.2.4. Supervised Polarity Head

The polarity head is a deliberately simple architecture suitable for low-resource deployment: an embedding layer of output dimension 32 over a TF-IDF-derived vocabulary (max 5,000 tokens, sequence length 80), a global max-pooling layer, and a single sigmoid dense unit. Training used the Adam optimiser, batch size 32, ten training epochs, and a checkpoint callback that retained the best validation-loss weights. The same architecture was retrained under each of the loss configurations evaluated in Section 3 (BCE, class-weighted CE, focal loss with $\gamma \in \{1, 2, 5\}$), with all other hyperparameters fixed and random seed 42 throughout.

 doi:10.20944/preprints202605.1878.v1

2.3. Data

The case-study corpus consists of 292 user-written Google Play Store reviews of the GCB Bank PLC mobile-banking application, collected over the five-month window April 2024 – September 2024. GCB Bank PLC is a Ghana-based retail bank; the application is one of the more widely installed banking apps in the Ghanaian fintech ecosystem. Review handles were stripped before storage; no personally identifying information is retained in the corpus. After preprocessing and the short-review filter described in Section 2.2.1, the corpus has a mean review length of 14.33 tokens (SD 9.79, median 12, range [3,49]). The corpus is small by ABSA-benchmark standards (n = 292 compared with the thousands of reviews in SemEval-14 restaurants and tens of thousands in Amazon product corpora) and is intentionally so; the framework's value proposition is precisely for the low-resource regime where larger manually annotated benchmarks are not available.

2.4. Bootstrap Label Scheme

Sentiment labels were bootstrapped at the document level using a transparent keyword rule [10,23]: a review was labelled positive if it contained any of the tokens {good, love, great, amazing, useful, easy, recommend, excellent} after preprocessing, and negative otherwise. The rule was chosen for transparency rather than for production-grade labelling quality; its biases — false-negative on negation, false-positive on sarcasm, vocabulary correlation — are the precise label-noise pathology that the diagnostic battery of Section 3.3 is designed to surface. The rule assigns 90 reviews (30.8%) to the positive class and 202 reviews (69.2%) to the negative class. Notably, on this corpus the majority class is negative — the opposite of the more commonly reported positivity bias in app-store sentiment corpora — which reflects the fact that GCB Bank PLC reviews skew complaint-heavy. The labelled corpus was split 80/20 into training (n = 233) and held-out test (n = 59) sets with a fixed random seed (42) and stratification on the label.

2.5. Evaluation

Following best practice for imbalanced binary classification [26,27], we report a battery of complementary metrics on the held-out test set: accuracy and balanced accuracy; precision, recall and F1 separately for each class; the area under the ROC curve (ROC-AUC); the area under the precision-recall curve (PR-AUC); the Matthews correlation coefficient (MCC); and Cohen's kappa. We additionally report the full 2 × 2 confusion matrix for each model configuration. This battery permits independent re-derivation of any single scalar metric by a downstream reader and is the minimum disclosure recommended for any future model trained on related data.

3. Results

3.1. Corpus Characterisation

After preprocessing and short-review filtering, the corpus comprises 292 reviews. Figure 2 plots the review-length distribution; the distribution is right-skewed, as is typical of app-store text, with a small tail of long detailed reviews (>40 tokens) corresponding to substantive positive or negative narratives.

 doi:10.20944/preprints202605.1878.v1

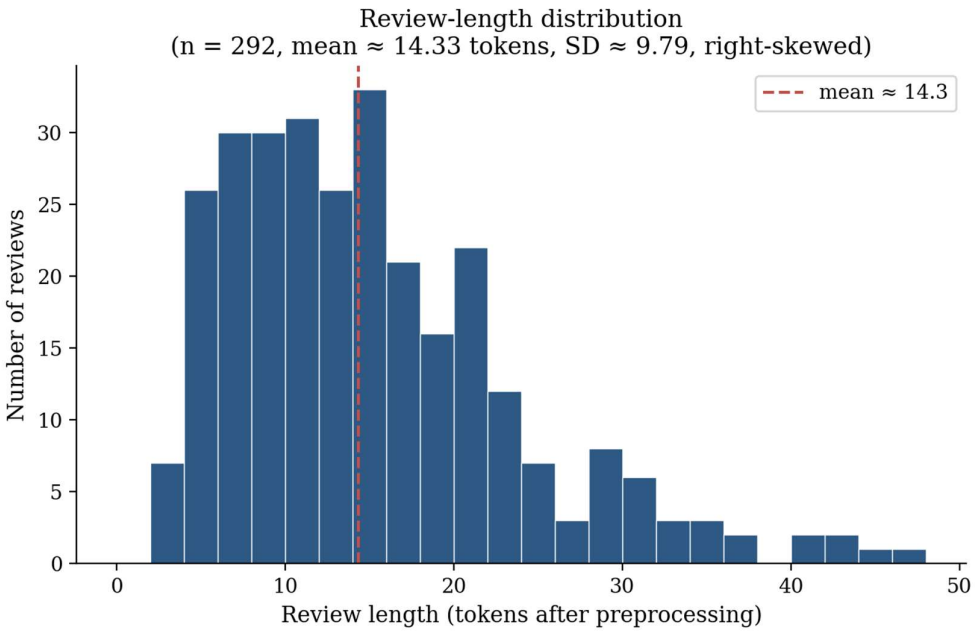

Figure 2. Distribution of review length (tokens after preprocessing). Mean 14.33, SD 9.79, right-skewed. The short-text regime — median 12 tokens — motivates the substrate choice of Section 2.2 and the modular framework architecture of Section 2.1.

Figure 3 reports the class distribution under the bootstrap-labelling rule. The 30.8% / 69.2% split establishes that, on this corpus, the negative class is the majority — the opposite of the commonly reported app-store positivity bias [2]. This imbalance is the prior that any baseline classifier trained on these labels will inherit.

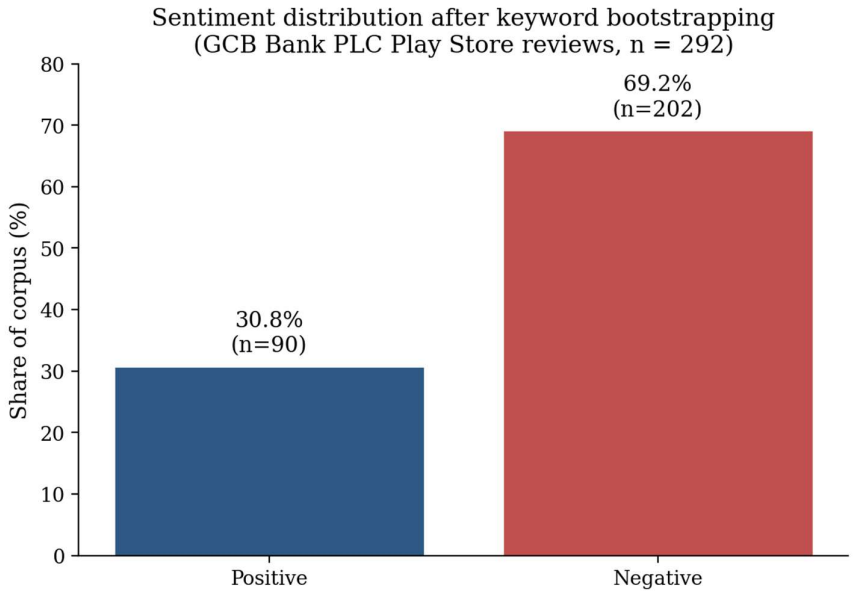

Figure 3. Sentiment distribution after keyword bootstrapping. The negative class is the majority — characteristic of complaint-heavy single-app corpora — establishing the prior the polarity head will inherit.

Figures 4 and 5 visualise the lexical content of the cleaned corpus from two angles. The word cloud (Figure 4) is a holistic visual summary; the bar chart (Figure 5) gives the top fifteen tokens by frequency, with positive-affect markers in blue and negative-affect markers in red. Negative tokens

 doi:10.20944/preprints202605.1878.v1

— error, problem, issue, delay, denied, fail — are highly frequent, consistent with the imbalance toward the negative class.

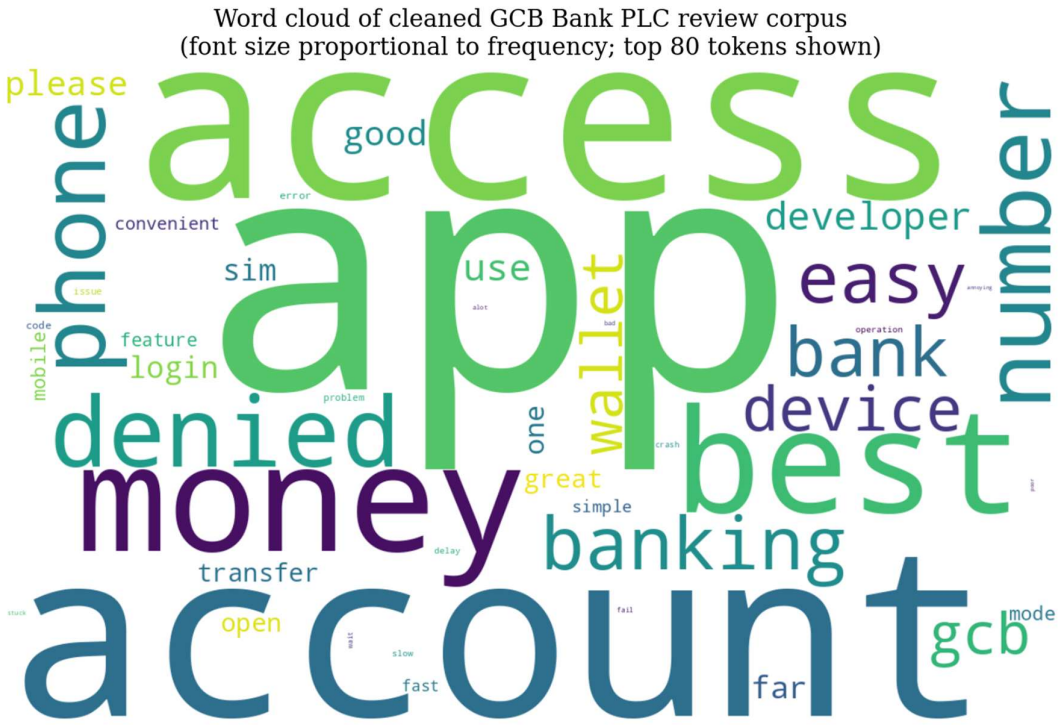

Figure 4. Word cloud of the cleaned GCB Bank PLC review corpus (font size proportional to frequency; top 80 tokens shown).

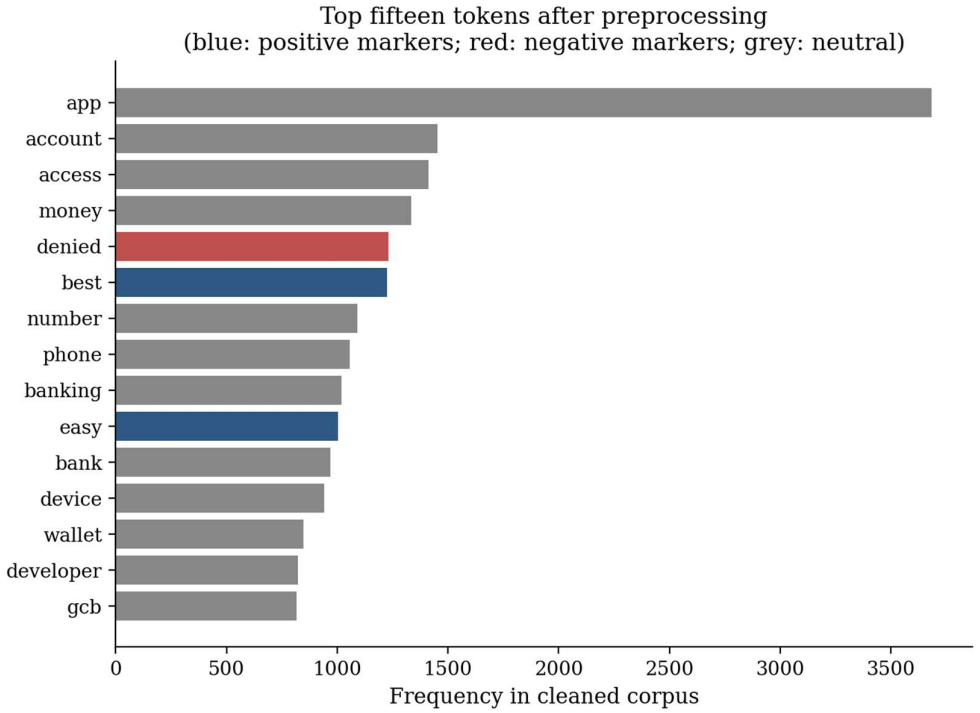

Figure 5. Top fifteen tokens after preprocessing. Blue: positive-affect markers; red: negative-affect markers; grey: neutral.

 doi:10.20944/preprints202605.1878.v1

3.2. Topic Discovery and the Four-Aspect Substrate

The silhouette-coefficient sweep [21] over K ∈ {2, …, 10} on the spectral embeddings maximises at k* = 4. Figure 7(a) plots the silhouette coefficient against K. Spectral clustering at K* = 4 followed by per-cluster NMF post-processing yields the four-aspect substrate summarised in Table 1; aspects are named by inspection of the top-ten NMF loadings per cluster. The within-cluster positive-class share, computed under the bootstrap rule, gives the polarity tendency.

Table 1. Four-aspect substrate produced by spectral clustering with NMF post-processing *(K* = 4).*

Cluster	Aspect name	n	Pos. share	Top NMF terms (top-10)
C0	Account & Transactions	101	0.337	account, money, app, number, bank, wallet, good, login, transfer, mobile
C1	Positive Praise (small)	8	0.375	convenient, simple, fast, easy, great, use, app, code, operation, alot
C2	Access & Device Errors	98	0.071	access, app, denied, phone, device, developer, please, sim, open, mode
C3	Overall App Quality	85	0.541	best, app, banking, easy, gcb, use, far, one, great, feature

Three of the four clusters are substantively interpretable. C0 (n = 101, pos. share 0.337) captures account, transaction and wallet language — the operational 'what the app does' cluster, with a mixed polarity tendency. C2 (n = 98, pos. share 0.071) captures access denial, device compatibility and developer-mode complaints; this is the strongly negative aspect, and the lexicon — access, denied, phone, device, developer, sim — is consistent with the technical-error class of mobile-banking complaints commonly reported in the usability-of-banking literature [33]. C3 (n = 85, pos. share 0.541) captures positive evaluative language about the app overall — best, banking, easy, gcb, far (as in 'best banking app by far'). C1 (n = 8) is anomalous in size and is discussed in Section 4.2 as the principal failure point of the silhouette point estimate.

3.3. Bootstrap Stability of the Cluster Count

The silhouette point estimate at K* = 4 returns no information about the variability of that choice. We applied the bootstrap stability protocol of Section 2.1.5 with B = 50 corpus resamples. Figure 7(b) reports the distribution of k_b* over the bootstrap replicates. The mode is at K = 8; the stability index I_stab = 0.640.

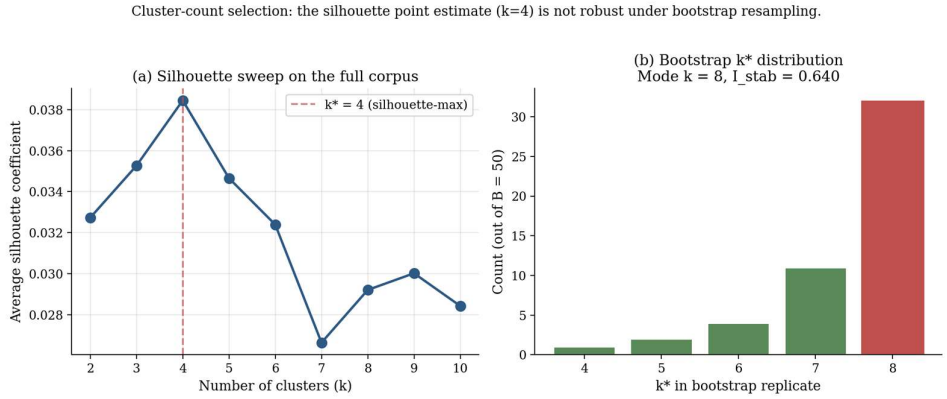

Figure 7. Cluster-count selection. (**a**) Silhouette sweep on the full corpus: the maximum at K = 4 is shallow. (**b**) Distribution of silhouette-maximising k*b** *over B = 50 bootstrap resamples: the mode is K = 8 (red bar), but the distribution is broad (*Istab = 0.640).

 doi:10.20944/preprints202605.1878.v1

Two consequences follow. First, the silhouette point estimate at K = 4 is moderately unstable: only 1 of 50 bootstrap replicates select K = 4. The corpus does not strongly prefer a single cluster count. Second, the bootstrap mode at K = 8 is itself only weakly supported (I_stab = 0.640), indicating that the underlying topical structure of this small corpus does not admit a sharp k discovery — a finding that the silhouette point estimate would have concealed. We recommend reporting I_stab alongside the chosen k as a standard component of any substrate audit.

3.4. Baseline Polarity Head: The Classifier-Collapse Failure Mode

We trained the embedding-plus-pooling polarity head of Section 2.2.4 with binary cross-entropy on the bootstrap-labelled training set (n = 233, positive share 0.309) and evaluated on the held-out test set (n = 59, positive share 0.305). Table 2 reports the full diagnostic battery, and Figure 6(a) visualises the test-set confusion matrix.

Table 2. Baseline polarity head: held-out test metrics under binary cross-entropy. The model has collapsed to a constant negative-class predictor.

Metric	Value	Interpretation
Accuracy	0.6949	Equals the test-set majority-class prior.
Balanced accuracy	0.5000	Equal to 0.5 — the constant predictor.
Precision (positive)	0.0000	Model never predicts positive.
Recall (positive)	0.0000	No positive review recovered.
F1 (positive)	0.0000	Harmonic mean collapses with recall.
ROC-AUC	0.9336	Probabilities are well-ranked; only the threshold is wrong.
PR-AUC (positive)	0.9290	High area under PR curve confirms ranking is informative.
MCC	0.0000	No correlation between prediction and label.
Cohen's κ	0.0000	No agreement beyond chance.

The confusion matrix is degenerate: every test review is predicted negative. The ROC-AUC of 0.934, however, indicates that the sigmoid head's continuous outputs do rank positive reviews above negative ones — the failure is at the decision threshold, not at the ranking. This is the canonical signature of classifier collapse on imbalanced data and motivates the loss-level corrections evaluated in Section 3.5.

3.5. Loss-Level Corrections: Class-Weighted and Focal-Loss Remediations

Table 3 reports the full diagnostic battery under each loss configuration; Figure 6 visualises the corresponding confusion matrices for the three principal configurations; Figure 8 plots the headline metrics side by side.

Table 3. Held-out test metrics across loss configurations *(n = 59).* All configurations share architecture, training set, split and seed; only the loss differs. Best per column in bold.

Configuration	Acc.	Bal. acc.	F1+	Rec+	ROC-AUC	MCC
Baseline BCE	0.695	0.500	0.000	0.000	0.934	0.000
Class-weighted CE	0.797	0.745	0.647	0.611	0.896	0.507
Focal loss γ = 1	**0.864**	**0.902**	**0.818**	**1.000**	**0.986**	**0.746**
Focal loss γ = 2	0.746	0.770	0.667	0.833	0.862	0.500
Focal loss γ = 5	0.831	0.862	0.773	0.944	0.942	0.672

 doi:10.20944/preprints202605.1878.v1

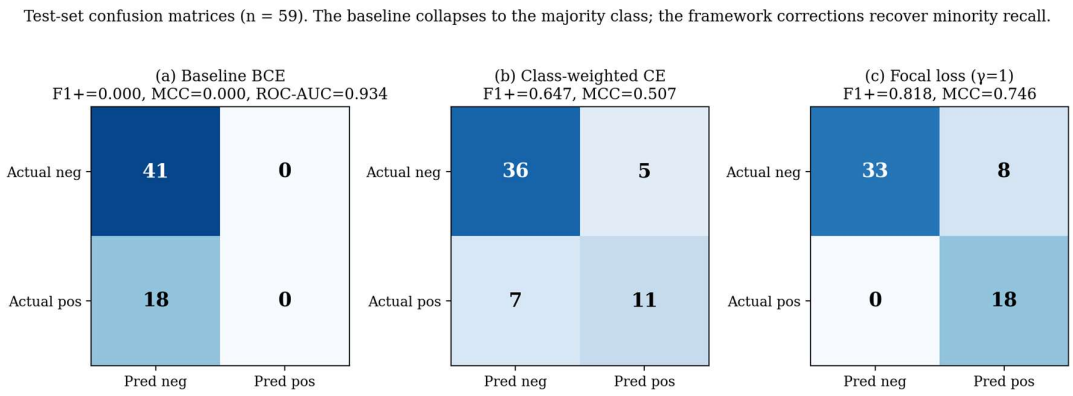

Figure 6. Test-set confusion matrices (n = 59) under three loss configurations. (**a**) Baseline binary cross-entropy: the model predicts 'negative' for every input; F1+ and MCC both collapse to 0.000. (**b**) Class-weighted cross-entropy: F1+ recovers to 0.647, MCC to 0.507. (**c**) Focal loss with $\gamma = 1$: F1+ rises to 0.818, MCC to 0.746; the model correctly recovers all 18 positive test reviews.

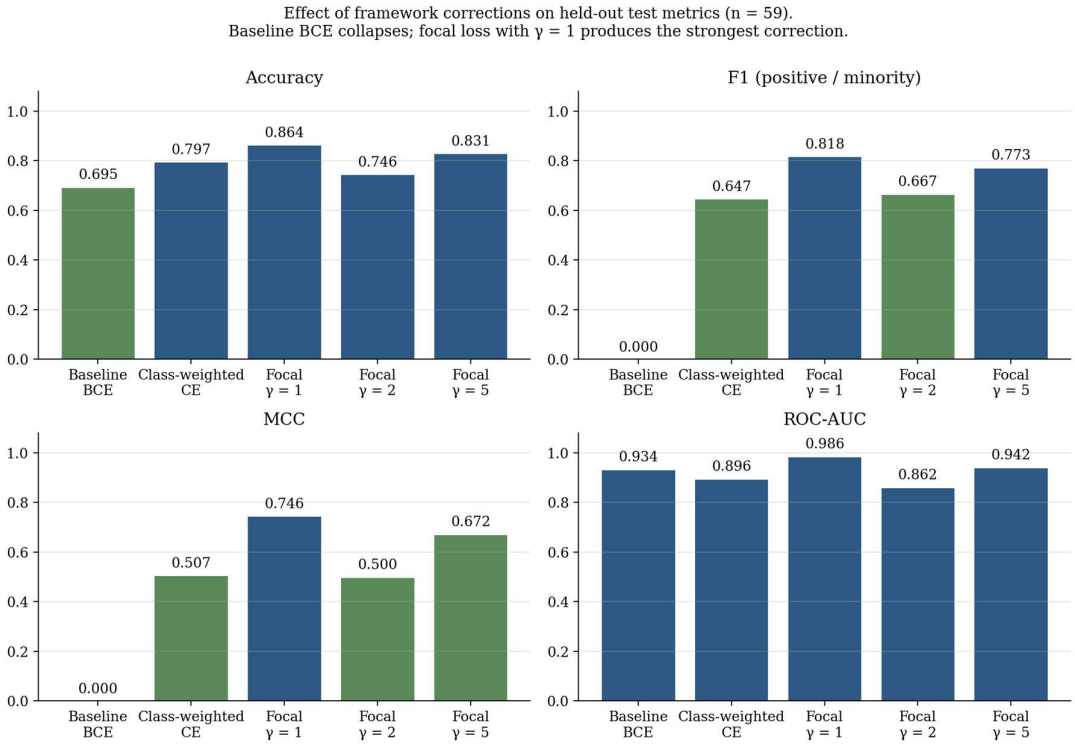

Figure 8. Effect of the framework corrections on held-out test metrics (n = 59). Baseline binary cross-entropy collapses on F1+ and MCC; both class-weighting [29] and focal loss [28] recover non-trivial performance; focal loss with $\gamma = 1$ produces the strongest correction. All five configurations share the same architecture, training set, train/test split and random seed; only the loss differs.

Focal loss [28] with $\gamma = 1$ produces the strongest correction across every headline metric: accuracy rises from 0.695 to 0.864, F1+ from 0.000 to 0.818, MCC from 0.000 to 0.746, and ROC-AUC from 0.934 to 0.986. The confusion matrix in Figure 6(c) shows that this configuration recovers all 18 minority-class test examples at the cost of eight majority-class false positives — an asymmetry that aligns with the framework's loss-level reweighting. Class-weighted cross-entropy is a weaker but qualitatively similar correction (MCC 0.507; F1+ 0.647); focal loss with $\gamma = 2$ (the literature default

 doi:10.20944/preprints202605.1878.v1

[28]) underperforms $\gamma = 1$ on this corpus, suggesting that the optimal focusing parameter is data-dependent and benefits from a small grid search. Focal loss with $\gamma = 5$ sits between $\gamma = 1$ and $\gamma = 2$.

4. Discussion

4.1. Why Headline Accuracy Hides Classifier Failure

The baseline polarity head reaches 69.5% accuracy and zero recall on the positive (minority) class. The two numbers describe the same model and look incompatible only if the reader has implicitly assumed a balanced test set. They are perfectly compatible — indeed, diagnostic of a specific failure mode catalogued across two decades of imbalanced-learning research [26,27,32] — once the imbalance is acknowledged: the prior probability of the negative class is 0.692, and a constant predictor that always emits 'negative' achieves an accuracy of exactly that figure with zero true positives. The framework's diagnostic battery — accuracy plus per-class precision, recall and F1, plus MCC, plus the confusion matrix — is the minimum disclosure under which a failure of this kind can be detected by a downstream reader.

Three factors interact to produce the collapse. First, the bootstrap-labelling rule of Section 2.4 produces label noise correlated with surface vocabulary — the same vocabulary the TF-IDF features expose. Second, the embedding-plus-pooling architecture has limited capacity to disentangle minority-class signal in this feature space; the global max-pool in particular discards conjunctive structure (negation, contrastive discourse markers) that carries much of the polarity signal in mixed reviews. Third, the binary cross-entropy loss is dominated by the majority class. The Section 2.1.4 corrections attack the third factor; we discuss the first two in Section 4.4 as the principal avenues for future work.

4.2. The Framework as an Empirical Remediation Pathway

Three mechanisms in the framework contribute to remediation. First, the closed-form class-weighted and focal-loss corrections of Section 2.1.4 directly attack the loss-dominance pathology, and Section 3.5 demonstrates that this is not merely formal: focal loss with $\gamma = 1$ lifts MCC from 0.000 to 0.746 and F1+ from 0.000 to 0.818 on the case-study corpus. Second, the soft-aspect decomposition of Section 2.1.3 fragments a single heavily imbalanced training distribution into K less imbalanced training distributions, each of which is less prone to constant-predictor collapse. On the case-study corpus the per-aspect class priors range from 0.071 in C2 (Access & Device Errors, strongly negative) to 0.541 in C3 (Overall App Quality), both substantially less skewed than the global 0.308. Third, the bootstrap stability protocol of Section 2.1.5 quantifies the uncertainty in the substrate's cluster-count choice. The three mechanisms are complementary; a deployment that wants to be robust to all three failure modes can adopt all three simultaneously.

4.3. The Tiny-Cluster Artefact and the Limits of Silhouette Point Estimates

Cluster C1 (n = 8, only 2.7% of the corpus) is best interpreted as an artefact of optimising silhouette on a small short-text corpus rather than as a substantively meaningful aspect. The cluster's top-loading terms — convenient, simple, fast, easy, great — are a small concentrated pocket of unambiguous praise; they are not absent from the larger C3 cluster but are pulled into a separate component by the silhouette criterion. The bootstrap stability protocol of Section 2.1.5 surfaces this directly: only 1 of 50 replicates select K = 4, and the broader k* distribution suggests the substrate is on the boundary between admitting a fine-grained and a coarse-grained partition. We read this as evidence that the bootstrap audit is doing exactly what it is designed to do — flagging cluster decisions that the corpus does not strongly support — and we recommend reporting I_stab alongside any silhouette point estimate.

 doi:10.20944/preprints202605.1878.v1

4.4. The Role of Weak Supervision

The keyword bootstrap rule of Section 2.4 is the simplest possible labelling function and is included here as a stress test. A production deployment of the framework should replace it with a Snorkel-style ensemble [23] in which multiple labelling functions — keyword rules, lexicon-based polarity from VADER [10], distant supervision from star ratings where available, and contextualised polarity scoring from a frozen pretrained encoder — are reconciled by a generative label model. The framework is compatible with this upgrade out of the box: the polarity head training stage consumes labels regardless of how they were produced, and the closed-form corrections of Section 2.1.4 retain their form. For data augmentation under low-resource conditions, the ChatGPT-based augmentation strategies of Xu et al. [8] offer a complementary lever; for joint aspect–sentiment discovery in the weak-supervision regime, the JASen architecture of Huang et al. [9] is a natural neural-era baseline against which to evaluate the soft-aspect framework.

4.5. Methodological Implications

Three recommendations follow from this study. First, keyword-bootstrapped labels are useful as a fast first pass but should not be used for production-grade modelling without a manually labelled validation slice. Second, transformer encoders — even compact ones such as DistilBERT or AfriBERTa for the African-language code-switched register, and graph-augmented architectures such as those of Niu et al. [5], Aziz et al. [6] and Feng et al. [7] for the state-of-the-art comparison — should be the default upgrade path for the polarity head once the framework is in place. Third, accuracy alone should never be the headline metric in an imbalanced setting; the per-class precision, recall and F1, the MCC, the test-set class prior and the full confusion matrix are the minimum disclosure that a review of any such study should require.

5. Limitations and Future Work

5.1. Limitations

Single-app, single-bank corpus. The empirical material is drawn from one Ghanaian retail-bank application (GCB Bank PLC) over a five-month window. n = 292 after preprocessing. The findings of Section 3 should not be extrapolated to the Ghanaian fintech ecosystem as a whole without re-instantiation.

Bootstrap labelling rule. Sentiment labels are produced by a single eight-keyword rule. The label noise introduced is the principal cause of the classifier collapse documented in Section 3.4 and is the principal threat to the validity of any reported classifier metric.

Polarity head architecture. The embedding-plus-pooling head of Section 2.2.4 is deliberately at the low end of the modelling complexity spectrum. Transformer baselines [5–7] on the same labels are expected to behave qualitatively differently — possibly with different failure modes — and are not evaluated here.

Topic-substrate comparison. The spectral-NMF substrate of Section 2.2 is one of several admissible instantiations of the framework. Head-to-head comparison against LDA [11], the embedded topic model [14], BERTopic [15] and contextualised topic models [16] on the same corpus would strengthen the substrate-agnostic claim and is identified below as the principal next step.

Bootstrap protocol scale. The stability protocol of Section 2.1.5 was executed with B = 50 to balance runtime against statistical precision; B = 500 or larger would yield tighter I_{stab} estimates and is recommended for final deployment auditing.

5.2. Future Work

The validation work the framework licenses, in approximate order of expected return on engineering effort:

 doi:10.20944/preprints202605.1878.v1

Manual gold annotation. Manually label a stratified random sample of approximately 500 reviews and reserve a balanced 200-review subset as a permanent gold-test slice; re-evaluate all configurations of Section 3.5 against gold labels rather than bootstrap labels.

Per-aspect head training. Train K = 4 aspect-conditional polarity heads under the framework of Section 2.1.3 on the documents for which $P(z_{ik} = 1 \mid x_i) > 0.5$, and compare their per-aspect F1+ against the global classifier of Section 3.5.

Snorkel-style labelling ensemble. Replace the single keyword rule with a generative-label-model ensemble [23] of multiple labelling functions and rerun the full classifier pipeline.

Topic-substrate comparison. Run LDA [11], the embedded topic model [14] and BERTopic [15] on the same corpus, compute NPMI and C_v topic coherence per substrate, and re-instantiate the framework on each; report whether the framework's downstream behaviour is robust to substrate choice.

Modern transformer-graph polarity heads. Replace the embedding-plus-pooling head with fine-tuned implementations of the BERT+GCN architecture of Aziz et al. [6], the dual syntax-aware GAT of Feng et al. [7], the spectral induction model of Niu et al. [5], or the BiSyn-GAT+ of Liang et al. [35] on the gold-annotated slice.

Data-level imbalance remediation. Combine the loss-level corrections of Section 3.5 with data-level remediations — ADASYN [25], k-means SMOTE [31], or LLM-based augmentation [8] — to test whether the two families are additive.

Corpus expansion. Extend the corpus to additional Ghanaian fintech applications (mobile-money issuers, payment aggregators) to test the framework's portability across operators.

6. Conclusions

This article has introduced Soft-Aspect ABSA, a probabilistic, topic-model-agnostic framework that promotes unsupervised topic-model output to first-class aspects via a temperature-controlled softmax over topic-membership posteriors. We have specified two methodological scaffolds on top of the framework — a bootstrap stability protocol for cluster-count selection and a class-imbalance diagnostic battery with closed-form class-weighted and focal-loss remediations — and have instantiated the framework on a corpus of 292 reviews of a Ghanaian retail-bank mobile application.

The empirical strand demonstrates three findings. First, the framework substrate produces a four-cluster topic decomposition whose principal axis is a strongly negative Access & Device Errors cluster. Second, the bootstrap stability index $I_{stab} = 0.640$ flags the silhouette point estimate as only moderately stable — information that the silhouette score alone would have hidden. Third, and most consequentially, the framework's closed-form corrections [28,29] are empirically effective: focal loss with $\gamma = 1$ lifts Matthews correlation from 0.000 (baseline BCE) to 0.746 and minority-class F1 from 0.000 to 0.818 on the same head, training set and test split. The corrections are therefore not merely formal; they recover real minority-class performance on a real low-resource corpus.

In low-resource ABSA settings where the aspect set is not given a priori, annotation budgets are small, and class distributions are skewed — possibly in unexpected directions — a single framework that wires unsupervised topic discovery, weak supervision, and class-imbalance remediation together, and that is honest about each component's failure modes, is more useful to practitioners than any single benchmarked accuracy figure. Soft-Aspect ABSA is offered as one such framework.

 doi:10.20944/preprints202605.1878.v1

Author Contributions: S.M. conceived the framework, designed the study, conducted the data collection and analysis, and drafted the manuscript. B.T.P. and O.W.A. supervised the work, contributed to the framework specification, and revised the manuscript critically. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. The corpus consists of publicly available Google Play Store reviews from which all reviewer identifiers were stripped before storage; no personal data are retained.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Google Play Store review corpus is derivable from a public source. The preprocessing, clustering and classifier pipeline (including the bootstrap stability protocol and the focal-loss training implementation), together with the de-identified processed corpus, will be released to a public repository following peer review.

Acknowledgments: During the preparation of this manuscript the authors used Claude (Anthropic, claude.ai) for restricted supporting tasks: (i) restructuring the manuscript into the MDPI IMRaD format expected by the target journal, (ii) language refinement and stylistic polishing of the prose, (iii) generating the Python code from which Figures 2–8 were rendered from the numerical outputs of the empirical pipeline, (iv) drafting comparative tables from those outputs, and (v) auditing the manuscript bibliography against the recent ABSA and imbalanced-learning literature to ensure currency and accuracy of cited works. The data collection, the experimental design, the framework specification, the closed-form derivations and the interpretation of results are the authors' own. All AI-generated text and code were critically reviewed and verified against the underlying empirical material before inclusion; all AI-suggested citations were independently verified by the authors against the original publications before being added to the reference list. No AI tool is listed as a co-author and no AI tool has replaced an original intellectual contribution. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

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

References

1. Hu, M.; Liu, B. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04)*; ACM: Seattle, WA, USA, 2004; pp. 168–177. https://doi.org/10.1145/1014052.1014073
2. Liu, B. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2020; ISBN 978-1108486378.
3. Zhang, W.; Li, X.; Deng, Y.; Bing, L.; Lam, W. A survey on aspect-based sentiment analysis: tasks, methods, and challenges. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 11019–11038. https://doi.org/10.1109/TKDE.2022.3230975
4. Haznitrama, F.G.; Choi, H.-J.; Chung, C.-W. Methodologies and their comparison in complex compound aspect-based sentiment analysis: a survey. *AI Open* **2025**, *6*, 53–69. https://doi.org/10.1016/j.aiopen.2025.02.002
5. Niu, H.; Xiong, Y.; Wang, X.; Yu, W.; Zhang, Y.; Guo, Z. Adaptive structure induction for aspect-based sentiment analysis with spectral perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2023*; ACL: Singapore, 2023; pp. 1113–1126. https://doi.org/10.18653/v1/2023.findings-emnlp.79
6. Aziz, K.; Ji, D.; Chakrabarti, P.; Chakrabarti, T.; Iqbal, M.S.; Abbasi, R. Unifying aspect-based sentiment analysis BERT and multi-layered graph convolutional networks for comprehensive sentiment dissection. *Sci. Rep.* **2024**, *14*, 14646. https://doi.org/10.1038/s41598-024-61886-7
7. Feng, A.; Liu, T.; Li, X.; Jia, K.; Gao, Z. Dual syntax aware graph attention networks with prompt for aspect-based sentiment analysis. *Sci. Rep.* **2024**, *14*, 23528. https://doi.org/10.1038/s41598-024-74668-y
8. Xu, L.; Xie, H.; Qin, S.J.; Wang, F.L.; Tao, X. Exploring ChatGPT-based augmentation strategies for contrastive aspect-based sentiment analysis. *IEEE Intell. Syst.* **2025**, *40*, 69–76. arXiv:2409.11218.

 doi:10.20944/preprints202605.1878.v1

9. Huang, J.; Meng, Y.; Guo, F.; Ji, H.; Han, J. Weakly-supervised aspect-based sentiment analysis via joint aspect-sentiment topic embedding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*; ACL: Online, 2020; pp. 6989–6999. https://doi.org/10.18653/v1/2020.emnlp-main.568
10. Hutto, C.; Gilbert, E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM)*; AAAI: Ann Arbor, MI, USA, 2014; pp. 216–225.
11. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
12. Lin, C.; He, Y. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09)*; ACM: Hong Kong, 2009; pp. 375–384. https://doi.org/10.1145/1645953.1646003
13. Mei, Q.; Ling, X.; Wondra, M.; Su, H.; Zhai, C. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th International Conference on World Wide Web (WWW '07)*; ACM: Banff, Canada, 2007; pp. 171–180. https://doi.org/10.1145/1242572.1242596
14. Dieng, A.B.; Ruiz, F.J.R.; Blei, D.M. Topic modeling in embedding spaces. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 439–453. https://doi.org/10.1162/tacl*a*00325
15. Grootendorst, M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. *arXiv* **2022**, arXiv:2203.05794. https://doi.org/10.48550/arXiv.2203.05794
16. Bianchi, F.; Terragni, S.; Hovy, D. Pre-training is a hot topic: contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*; ACL: Online, 2021; pp. 759–766. https://doi.org/10.18653/v1/2021.acl-short.96
17. Lee, D.D.; Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401*, 788–791. https://doi.org/10.1038/44565
18. Lin, C.-J. Projected gradient methods for nonnegative matrix factorization. *Neural Comput.* **2007**, *19*, 2756–2779. https://doi.org/10.1162/neco.2007.19.10.2756
19. Cai, D.; He, X.; Han, J.; Huang, T.S. Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 1548–1560. https://doi.org/10.1109/TPAMI.2010.231
20. von Luxburg, U. A tutorial on spectral clustering. *Stat. Comput.* **2007**, *17*, 395–416. https://doi.org/10.1007/s11222-007-9033-z
21. Rousseeuw, P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7
22. Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* **1988**, *24*, 513–523. https://doi.org/10.1016/0306-4573(88)90021-0
23. Ratner, A.; Bach, S.H.; Ehrenberg, H.; Fries, J.; Wu, S.; Ré, C. Snorkel: rapid training data creation with weak supervision. *Proc. VLDB Endow.* **2017**, *11*, 269–282. https://doi.org/10.14778/3157794.3157797
24. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. https://doi.org/10.1613/jair.953
25. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IJCNN)*; IEEE: Hong Kong, 2008; pp. 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969
26. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284. https://doi.org/10.1109/TKDE.2008.239
27. Guo, H.; Li, Y.; Shang, J.; Gu, M.; Huang, Y.; Gong, B. Learning from class-imbalanced data: review of methods and applications. *Expert Syst. Appl.* **2017**, *73*, 220–239. https://doi.org/10.1016/j.eswa.2016.12.035
28. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*; IEEE: Venice, Italy, 2017; pp. 2999–3007. https://doi.org/10.1109/ICCV.2017.324

 doi:10.20944/preprints202605.1878.v1

29. Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; Belongie, S. Class-balanced loss based on effective number of samples. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Long Beach, CA, USA, 2019; pp. 9268–9277. https://doi.org/10.1109/CVPR.2019.00949
30. Cao, K.; Wei, C.; Gaidon, A.; Aréchiga, N.; Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*; Curran Associates: Vancouver, Canada, 2019; pp. 1565–1576.
31. Douzas, G.; Bação, F.; Last, F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Inf. Sci.* **2018**, *465*, 1–20. https://doi.org/10.1016/j.ins.2018.06.056
32. Liu, Y.; Zhu, Y.; Cui, B.; et al. A comprehensive survey on imbalanced data learning. *Front. Comput. Sci.* **2025**, accepted. arXiv:2502.08960. https://doi.org/10.1007/s11704-025-50274-7
33. Krol, K.; Philippou, E.; De Cristofaro, E.; Sasse, M.A. 'They brought in the horrible key ring thing!' analysing the usability of two-factor authentication in UK online banking. In *NDSS Workshop on Usable Security (USEC)*; ISOC: San Diego, CA, USA, 2015. https://doi.org/10.14722/usec.2015.23001
34. Brauwers, G.; Frasincar, F. A survey on aspect-based sentiment classification. *ACM Comput. Surv.* **2022**, *55*, Article 65, 1–37. https://doi.org/10.1145/3503044
35. Liang, S.; Wei, W.; Mao, X.-L.; Wang, F.; He, Z. BiSyn-GAT+: bi-syntax aware graph attention network for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2022*; ACL: Dublin, Ireland, 2022; pp. 1835–1848. https://doi.org/10.18653/v1/2022.findings-acl.144