electronics

MDPI

Article

Aspect-Based Sentiment Analysis of Turkish School Reviews: A Comparative Study of Recurrent and Transformer Architectures

Harun Aksaya [1,*] and Sevinç Gülseçen [2]

[1] Department of Information Technology, Marmara University, 34100 Istanbul, Turkey
[2] Department of Computer Science, Istanbul University, 34100 Istanbul, Turkey; gulsecen@istanbul.edu.tr
* Correspondence: harunaksaya@marmara.edu.tr; Tel.: +90-530-281-37-38

Abstract

Aspect-based sentiment analysis (ABSA) of user-generated Turkish text remains underexplored, and no aspect-level resource exists for the education domain. We introduce a manually annotated dataset of 1533 school reviews from okul.com.tr, yielding 2620 (sentence, aspect) pairs over 19 aspect categories, imbalanced 2.3:1 toward positive polarity; a stratified 200-instance subset gave Cohen's Kappa = 0.702. Given the aspect category, nine systems were compared under a 5-fold cross-validation protocol partitioned at the review level to prevent review-level leakage: four non-pre-trained baselines, two recurrent models (BiLSTM, BiGRU), and three fine-tuned transformers—the Turkish-specific BERTurk together with the multilingual mBERT and XLM-R, the three trained under identical hyperparameters, with class-imbalance handling applied where relevant. TF-IDF with a linear SVM reached a Macro-F1 of 0.836, exceeding both recurrent models, while all transformers exceeded every non-pre-trained system. In this dataset and experimental setting, BERTurk obtained the highest Macro-F1 (0.927) and accuracy (0.938), 0.091 above the strongest classical baseline. The advantage over mBERT concentrated on the minority class: F1-Negative exceeded mBERT's by 0.034 points, roughly twice the accuracy gap. Its margin over XLM-R was small and not statistically established. The benefit of language-specific pre-training therefore appears to narrow as the multilingual model's pre-training corpus grows.

Keywords: aspect-based sentiment analysis; Turkish natural language processing; BERTurk; transformer models; multilingual pre-trained language models; educational text mining; class imbalance

Check for updates

Academic Editor: Arkaitz Zubiaga

Received: 25 August 2026
Revised: 19 October 2026
Accepted: 19 October 2026
Published: 26 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Before social media and online platforms became so pervasive, the opportunity for consumers, students, or patients to share their experiences with a broad audience was quite limited. Today, millions of reviews, ratings, and comments can accumulate on a single platform alone—and manually sifting through them is no longer feasible. This is where sentiment analysis and opinion mining come in: a branch of natural language processing (NLP) research that aims to automatically infer what people feel from what they write [1,2]. Spanning applications from e-commerce and healthcare to political discourse and education, the field has attracted intense academic and industry interest over the past decade.

The field's roots trace back to the early 2000s, when text classification tools were first repurposed for polarity detection. Pang, Lee, and Vaithyanathan [3] applied standard

machine learning classifiers (Naive Bayes, Maximum Entropy, SVM) to movie review polarity and were among the first to demonstrate, systematically, that sentiment classification poses distinct challenges not shared by conventional topic classification. In the same year, Turney [4] proposed an unsupervised alternative, computing a word's semantic orientation from its pointwise mutual information with a small set of positive and negative reference terms—an approach that required no labeled data at all. Both studies established that document-level sentiment classification was feasible, yet both were limited to assigning a single polarity value to an entire document—a review, a critique. Hu and Liu [5] were the first to explicitly identify this limitation, showing that a single product review typically expresses divergent opinions about multiple distinct features, and proposing a feature-based opinion mining framework that extracts sentiment at the level of individual product attributes rather than the document as a whole. That framework is the direct forerunner of what is known today as aspect-based sentiment analysis.

Consider a user writing about a school: "the teachers are caring but the building is in terrible shape." The problem Hu and Liu [5] identified two decades ago applies just as directly here: a conventional sentiment classifier assigns a single polarity to this sentence—positive or negative—and inevitably discards information. Aspect-based sentiment analysis (ABSA) takes a fundamentally different approach: it identifies each aspect target in the text and assigns a distinct sentiment polarity to each [6]. In this sentence, the teacher aspect should be labeled positive; the physical infrastructure aspect, negative. That level of granularity is what makes ABSA genuinely useful for decision support—yet, as Section 2 details, no prior study offers this granularity for Turkish in an educational setting.

This first generation of lexicon- and feature-engineering-based methods gave way, from the mid-2010s onward, to deep neural networks that learn contextual representations directly from data; how this transition unfolded specifically for ABSA, and how the transformer architecture [7] reshaped the field, are discussed in detail in Section 2.2. What matters for the present study is that this architectural progress matured quickly for resource-rich languages such as English, but considerably less so for morphologically rich, comparatively resource-constrained languages such as Turkish—an asymmetry examined in Sections 2.3 and 2.4.

This study addresses the gap that emerges from that asymmetry through three concrete contributions. First, it introduces a new dataset of 2620 labeled (sentence, aspect) pairs across 19 aspect categories, derived from 1533 school reviews collected from okul.com.tr. This resource was built rather than borrowed because no existing Turkish ABSA corpus covers the education domain, and because aspect taxonomies do not transfer across domains: the categories of the available Turkish resources describe restaurants, tourism venues or brand mentions, none of which span the dimensions along which schools are actually evaluated (Section 2.7, Table 1). Second, label reliability is validated through an inter-annotator agreement study involving two independent experts ($\kappa = 0.702$) [8,9]. Third, a dual comparison is presented—recurrent architectures against pre-trained transformers, and monolingual (BERTurk) against multilingual (mBERT, XLM-R) transformers—with all five models evaluated under the same 5-fold cross-validation protocol, benchmarked against four non-pre-trained baselines, and supported by error analysis, significance testing and stability assessment. The related literature is reviewed in Section 2, the dataset and methodology are presented in Section 3, the results follow in Section 4, Section 5 discusses these findings, and Section 6 concludes the study.

Table 1. Turkish ABSA resources—domain, scale, and aspect scheme.

Resource	Domain	Scale	Aspect Scheme	IAA
SemEval-2016 TR [10]	Restaurant	1317 sentences (320 reviews)	12 entity#attribute	Not reported
Salur and Aydin [11]	Tourism	4482 explicit + 21,691 implicit terms	Terms, no taxonomy	7 annotators
Mutlu and Özgür [12]	Twitter (brands)	3952 tweets	6 named targets	$\kappa = 0.855$
Bilgin and Turan [13]	Restaurant	re-uses SemEval-2016 TR	12 entity#attribute	Not applicable
This study	Education (K–12)	2620 pairs (1533 reviews)	19 categories	$\kappa = 0.702$

2. Related Work

This section situates the study within the literature by narrowing from general aspect-based sentiment analysis toward the specific gap addressed here: aspect-based sentiment analysis for Turkish educational text. Each subsection inherits a slightly more specific version of the problem raised in the one before it, converging on the research gap articulated in Section 2.7.

2.1. The SemEval Benchmark Series and Task Standardization

ABSA is not a single task but a family of related subtasks—aspect term extraction, aspect category detection, and aspect sentiment classification among them [14]. Formalizing these subtasks and pairing them with comparable benchmark data is due largely to the SemEval shared task series. The inaugural 2014 edition [6] introduced aspect-term and category-level annotations for restaurant and laptop reviews, giving the field a shared evaluation footing; the following two editions [10,15] extended this framework to multiple languages (including Arabic, Dutch, French, Russian, Spanish, and Turkish) and to out-of-domain evaluation settings. A comprehensive survey by Zhang et al. [14] notes that these three benchmarks remain the field's primary reference point, while also observing that languages other than English are only sparsely represented across them. Efforts to close this benchmark gap continue: Wu et al. [16] introduced M-ABSA in 2025, spanning 21 languages and seven domains, and included Turkish among them. The Turkish portion of M-ABSA, however, was produced through automatic translation followed by human review rather than organic collection from native Turkish text—underscoring that the need for naturally occurring, domain-specific Turkish ABSA data remains unmet.

2.2. The Evolution of Deep Learning Approaches for ABSA

Modeling strategies for ABSA have followed a clear trajectory since the SemEval era. The ground for this trajectory was laid by distributed representation methods that encode words as dense vectors; Mikolov et al.'s [17] word2vec replaced the sparse, hand-engineered feature pipelines that had dominated text classification with learned representations that encode semantic similarity directly in vector space, and these representations formed the input layer for the first generation of recurrent ABSA models. Tang et al. [18] concatenated the target word with the Long Short-Term Memory (LSTM) input, while Wang et al. [19] adapted the attention mechanism originally proposed by Bahdanau et al. [20] for machine translation, aligning attention weights with the aspect term to focus the model on relevant context. The bidirectional recurrent cells from this period—Bidirectional LSTM (BiLSTM) [21] and Bidirectional GRU (BiGRU) [22]—remain in use today as lightweight, interpretable baselines, which is precisely why they are included in this study.

A second wave introduced graph-based methods that expose syntactic dependency structure to the model explicitly. Zhang et al. [23] ran graph convolutional networks (GCNs) over a sentence's dependency tree, capturing long-range syntactic relations between an

aspect and its associated opinion word in a way that word-order-based attention mechanisms could not. The third, and most consequential, wave began with Vaswani et al.'s [7] transformer architecture, which replaced recurrence entirely with self-attention layers, and with Bidirectional Encoder Representations from Transformers (BERT) [24], built on top of it. Sun et al. [25] reframed the aspect-sentence pair as an auxiliary sentence and fed it to BERT in a sentence-pair classification format, achieving the best results reported to that point—this is the direct precedent for the [CLS] aspect [SEP] sentence [SEP] input format adopted in this study.

Two parallel trends define the current period. Zhang et al. [26] reformulated multiple ABSA subtasks—term extraction, category detection, sentiment classification—as a single generative, text-to-text problem, enabling joint learning across tasks at a substantially higher computational cost. In parallel, the zero- and few-shot ABSA capabilities of instruction-tuned large language models (LLMs) have come under scrutiny; Simmering and Huoviala [27] evaluated GPT-4 and GPT-3.5 on the SemEval-2014 benchmark, finding that detailed prompt engineering substantially improves zero-/few-shot performance, yet fine-tuned models at BERT scale remain competitive—and in some configurations superior. This line of work continued into 2024: Ding et al. [28] moved beyond pure prompt engineering, adapting LLMs to ABSA subtasks incrementally through a continual learning framework, and showed that this hybrid approach yields more stable results than either zero-shot prompting or single-pass fine-tuning alone. Taken together, these findings support the continued engineering relevance of fine-tuning a compact transformer on a well-defined, closed-set binary classification task—a computationally predictable choice, which is the approach taken here.

2.3. Multilingual and Language-Specific Pre-Trained Representations

The emergence of multilingual BERT variants (mBERT, XLM-R) raised a natural question: is it preferable to pre-train across dozens of languages on a shared sub-word vocabulary, or to specialize a model on a single language? Part of XLM-R's improvement over mBERT is not attributable to language coverage at all: Liu et al.'s [29] RoBERTa demonstrated that a more carefully optimized pre-training recipe—larger batches, longer training, dynamic masking—substantially improves BERT-style models independent of the pre-training corpus's language composition, which is precisely why holding the training recipe constant matters when comparing multilingual and monolingual models. The language-coverage question proper is not specific to Turkish, however—it has been tested repeatedly across typologically diverse languages. Martin et al. [30] showed that CamemBERT, trained for French, consistently outperformed an equivalently sized mBERT on part-of-speech tagging, dependency parsing, and named entity recognition, and that a modest 4 GB web corpus was sufficient to obtain this advantage. Antoun et al. [31] confirmed a similar pattern for Arabic with AraBERT, reporting clear gains over multilingual mBERT across nearly every downstream task evaluated.

The mechanism behind this recurring pattern has been examined directly by Rust et al. [32]: a shared multilingual sub-word vocabulary frequently over-segments low-resource or morphologically complex languages during tokenization, constraining the model's capacity to represent that language's semantic units efficiently. This indicates that the gains observed for CamemBERT and AraBERT are not incidental but stem from a systematic relationship between the language coverage of the pre-training corpus and tokenizer quality. Taken together, this body of work frames the comparison between BERTurk, mBERT, and XLM-R adopted in this study not as a novel idea in itself, but as the application of an established methodological tradition—testing the benefit of language-

specific pre-training under controlled conditions—to Turkish and, specifically, to the education domain.

2.4. Turkish Natural Language Processing and Sentiment Analysis

Turkish's agglutinative structure and rich inflectional morphology complicate the direct transfer of English-centric NLP tooling. Early work addressed this largely through lexicon-based methods, machine translation, or cross-lingual transfer. Dehkharghani et al. [33] introduced SentiTurkNet, assigning positive, negative, and neutral polarity scores to every synset in the Turkish WordNet, and showed quantitatively that this resource outperformed polarity scores obtained by directly translating from the English SentiWordNet—a concrete demonstration of why a language-specific resource matters. Alongside this lexicon-based line of work, Demirtas and Pechenizkiy [34] proposed a polarity detection approach built on machine-translated English-labeled data, while Kaya et al. [35] compared classical machine learning classifiers (Naive Bayes, Maximum Entropy, SVM) on Turkish political news text. A shared limitation across this line of work is the reliance on indirect solutions or static lexicons in the absence of domain-specific Turkish training data. The introduction of BERTurk [36] changed this picture; pre-trained directly on large-scale Turkish web corpora, it delivers a consistent quality gain over lexicon-based, translation-based, or multilingual alternatives on fine-tuning-based downstream tasks.

The field has remained active over the past two years. Çakıcı et al. [37] systematically surveyed 78 Turkish sentiment analysis studies published between 2012 and 2025 and cross-validated the more than 80 datasets that this literature produced, finding that the field's resources remain fragmented and difficult to compare against one another. Zümberoğlu et al. [38] proposed both dataset improvements and model innovations for Turkish sentiment analysis, showing that BERTurk-based approaches still leave room for further gains. Both recent studies confirm that general (document/sentence-level) Turkish sentiment classification is maturing, yet neither addresses the aspect level—the gap taken up in Section 2.5.

2.5. Turkish Aspect-Based Sentiment Analysis

Compared to general (document-level) Turkish sentiment analysis, Turkish ABSA literature remains quite narrow. Salur and Aydin [11] compiled a Turkish ABSA corpus for tourism reviews using seven annotators, concretely illustrating how labor-intensive producing high-quality labeled data in this space can be. Mutlu and Özgür [12] introduced both a new dataset and BERTurk-based models for targeted sentiment analysis, providing the first quantitative documentation of BERTurk's measurable advantage over multilingual counterparts. More recently, Bilgin and Turan [13] treated Turkish ABSA as a two-stage pipeline on the SemEval-2016 Turkish restaurant dataset, fine-tuning five transformer variants—BERT, ConvBERT, ELECTRA, DeBERTa and DistilBERT—first for aspect term and category extraction and then for polarity classification, reporting 92.20% accuracy on aspect identification and 86.91% on sentiment classification. Their findings reinforce the advantage of fine-tuned transformers over lexicon-based alternatives for Turkish and, at the same time, illustrate the domain concentration noted above: as with the tourism corpus of Salur and Aydin, the evaluation rests on a consumer-service domain. A more recent contribution approaches the problem from the multi-task side: Yaman and Aslan [39] fine-tune BERTurk jointly on three rated aspects—delivery speed, service and taste—across 676,214 Turkish food-delivery reviews, and report that a shared-encoder multi-task model with a class-weighted loss outperforms both single-task fine-tuning and TF-IDF baselines. Their aspect ratings are supplied by the platform rather than manually annotated, so the resource is not an annotated ABSA corpus in the sense used here, but the result reinforces

two observations that matter for this study: class weighting is an effective response to the skew these corpora exhibit, and Turkish work of this kind continues to be situated in consumer-service domains. Together, these studies indicate that Turkish ABSA has yet to diversify beyond a small number of application domains, and that studies offering a domain-specific aspect taxonomy beyond general-purpose targeted sentiment analysis are still missing.

2.6. Opinion Mining in Educational Contexts

Sentiment analysis applied to educational text has largely stayed at the document level. Altrabsheh et al. [40] proposed a general sentiment classification framework for classroom feedback text, and Wen et al. [41] examined how sentiment shifts in Massive Open Online Course (MOOC) discussion forums relate to student engagement. Both studies, however, assign a single sentiment label to an entire text, without distinguishing which specific service or instructional dimension a comment addresses.

More recent work partially closes this gap by operating at the aspect level. Sindhu et al. [42] proposed a two-layer LSTM architecture for evaluating faculty teaching performance from university student feedback, where the first layer detects aspects and the second assigns a sentiment polarity to each. Kastrati et al. [43] performed aspect category detection on MOOC reviews using a weakly supervised framework, substantially reducing the manual annotation cost typically required. This line strengthened further in 2023: Ren et al. [44] proposed a framework that automatically scores student feedback for teaching evaluation at the aspect level, while Kathuria et al. [45] combined lexicon-based and machine learning methods in a hybrid approach (AOH-Senti) that separately evaluates instructional dimensions such as punctuality, pacing, and subject mastery. The line has continued into the most recent literature, and beyond English. Butt et al. [46] segment 6025 Spanish-language student evaluations into opinion units and classify each across nine predefined teaching aspects, comparing fine-tuned BERT and RoBERTa against GPT-4o and LLaMA-3.1 under few-shot prompting. Their study is the closest structural analogue to the present work—a fixed aspect inventory applied to institutional feedback in a language other than English—and it indicates that the aspect-level framing transfers across languages. These studies together demonstrate that aspect-aware analysis in education is not only feasible but an actively pursued line of research. All of them, however, address higher-education or MOOC settings; no study addressing K-12 (primary and secondary school) parent or student reviews has been identified—precisely the gap the okul.com.tr dataset used in this study occupies.

2.7. Research Gap

The literature reviewed from Sections 2.1–2.6, narrowing progressively, surfaces three intersecting gaps. First, although the advantage of language-specific pre-training over multilingual alternatives (Section 2.3) has been replicated across typologically diverse language families, no study has quantitatively tested this advantage for Turkish at the aspect level under a controlled, 5-fold cross-validation protocol—existing Turkish ABSA work (Section 2.5) evaluates BERTurk against only a single multilingual baseline. Second, the aspect-level educational-text literature (Section 2.6) is confined to English university and MOOC settings; no study has addressed Turkish primary or secondary school reviews. Third, the intersection of these two gaps—the absence of a publicly available, aspect-level labeled dataset for Turkish educational text—is confirmed by both the general Turkish NLP resources reviewed in Section 2.4 and the Turkish ABSA studies reviewed in Section 2.5.

Table 1 places the present corpus alongside the Turkish ABSA resources discussed above, which makes the nature of the gap concrete.

Three observations follow, which together indicate why a new resource was required rather than an adaptation of an existing one.

First, no Turkish ABSA resource covers the education domain, and this reflects a difference in applicability rather than in scale. Aspect taxonomies are domain-bound by construction: the twelve SemEval categories are organized around entities such as FOOD, DRINKS and AMBIENCE, which do not correspond to the dimensions along which schools are evaluated, such as teaching staff, counseling services, parent communication and transport. A model trained on that taxonomy could therefore not be applied to school reviews without redefining the label space.

Second, the two remaining Turkish resources operate at annotation granularities that do not support aspect-category classification. Salur and Aydin annotate aspect terms without consolidating them into a category taxonomy, so the task they enable is term extraction rather than category-level polarity assignment. Mutlu and Özgür annotate sentiment toward six named brands; this is targeted sentiment analysis, in which the target is a surface entity string rather than a conceptual category, and it therefore addresses a different formulation of the problem.

Third, on the dimension of scale, the present corpus compares favorably with the established reference point for Turkish ABSA. At 2620 annotated (sentence, aspect) pairs it is roughly twice the size of the SemEval-2016 Turkish subset, which at 1317 sentences has served as the evaluation benchmark for Turkish ABSA work including the most recent study in the area [13], and it carries the largest aspect taxonomy of any of these resources at 19 categories. The corpus is nonetheless small in absolute terms, a limitation we return to in Section 5.4.

This study targets exactly these three gaps through the contributions summarized in Section 1: a new dataset, a validated annotation protocol, and a dual comparison spanning both recurrent-versus-transformer and monolingual-versus-multilingual transformer architectures. The dataset and experimental protocol are described in detail in Section 3.

3. Materials and Methods

3.1. Data Collection

School choice is among the more consequential decisions families make, and in Turkey it is increasingly informed by public review platforms rather than by word of mouth alone. The feedback accumulating on these platforms is also of direct operational interest to school administrations, for which it constitutes an unsolicited and continuous signal about service quality across distinct dimensions—teaching, facilities, counseling, transport. Extracting that signal requires resolving sentiment at the level of the individual dimension rather than the review as a whole, which is what motivates an aspect-level resource for this domain. The raw data used in this study was collected from okul.com.tr, a Turkish platform that aggregates parent and student reviews for primary and secondary schools across Turkey. It provides real user feedback organized by individual schools, making it a suitable source for this research. A total of 1533 reviews were manually read and processed; from each review, one or more aspect–sentiment pairs were extracted. Since a single review can address multiple topics simultaneously—for instance, evaluating both teacher quality and school facilities in the same text—a multi-aspect annotation scheme was adopted. Filtering was applied at the review level and before aspect extraction: reviews containing no identifiable aspect expression, and reviews in which the target of an evaluation could not be determined, were not taken into the corpus. Aspect expressions were therefore extracted only from reviews that had passed this screen, and no aspect expression was annotated and subsequently discarded. A count of the reviews removed at this stage was not kept.

Data collection and preparation. The reviews were collected from okul.com.tr over the 2024–2025 period. Collection was restricted to Turkish-language reviews of primary and secondary schools, and reviews containing insults or profanity were excluded; a check of the final vocabulary confirms that no such terms remain. Eleven pairs of reviews in the corpus carry identical text, giving 1522 distinct texts across 1533 reviews. These were posted by different users and were therefore retained as separate records rather than de-duplicated. No preprocessing was applied to the review text prior to annotation beyond the trimming of whitespace: the text was neither lower-cased nor stemmed, no stop words were removed, and no spelling correction was performed. No author identifiers were retained at collection, and the stored text contains no e-mail addresses, URLs, user handles, telephone numbers, identification numbers or dates; no personal names and no named institutions appear in the corpus, so no further anonymization step was required.

3.2. Aspect Category Construction

The raw annotation process produced 464 unique aspect expressions, many of which overlap semantically—for example, "öğretmen" (teacher), "öğretmenler" (teachers), and "hocalar" (a colloquial term for teachers) are three surface forms referring to the same concept. To reduce sparsity and make model training tractable, all expressions were mapped to 19 canonical aspect categories.

The 19 categories were not defined in advance. Annotation proceeded with free-form aspect labels: the first pass yielded 248 distinct labels, which rose to 495 as the corpus was extended and were consolidated to 464 in the final annotated set. The canonical inventory was then induced from these surface labels by grouping, and only 7 of the 248 labels from the first pass appear unchanged among the 19 category names. The inventory therefore reflects the distinctions the reviews themselves drew rather than a taxonomy fixed before annotation.

The mapping followed a two-stage process. A manually compiled dictionary of 488 entries assigned raw expressions directly to one of the 19 canonical categories; the number of dictionary variants per category ranges from 91 for Education Quality to a handful for the sparsest categories. Expressions absent from the dictionary were routed to a keyword-based fallback rule set built on 152 stem keywords. Of the 464 unique raw expressions, 446 (96.1%) were resolved by dictionary matching and 18 (3.9%) by the fallback rules; counted over expression occurrences rather than unique surface forms, the dictionary resolved 2762 of 2814 occurrences (98.2%). The expressions routed through the fallback were predominantly broad or compound terms—Ders ("course"), Eğitim Anlayışı ("educational approach"), Tanıtım ("promotion") and Mentörlük ("mentoring") account for 38 of the 52 fallback occurrences.

Because several raw expressions within one review can map to the same canonical category, consolidation also merges duplicate labels inside a review: Kütüphane ("library") and Laboratuvar ("laboratory"), for instance, both map to Physical Facilities and yield a single annotation. Merging applies only where the duplicated labels share a polarity. Where two raw expressions in the same review map to the same canonical category but carry opposite polarities, both assignments are retained. A review stating that Fiziksel İmkan ("physical facilities", positive) is good while Atölye ("workshop", negative) is absent yields both a positive and a negative instance of Physical Facilities, because the two expressions were distinct before consolidation. Forty (review, category) pairs in the final corpus carry both polarities in this way; the consequence for measurement is reported in Section 4.4. This merging reduced the 2814 raw aspect labels to 2620 and affected 175 of the 1533 reviews (11.4%)—160 reviews lost one label, 12 lost two, two lost three and one lost

four. No review lost its only label, so the merging changes annotation granularity rather than corpus coverage. The final 19 categories are listed in Table 2.

Table 2. Canonical aspect categories, frequency, and example expressions.

#	Category (TR)	Category (EN)	Frequency	Example Expressions
1	Eğitim Kalitesi	Education Quality	532	eğitim kalitesi, kaliteli eğitim, akademik başarı
2	Genel Değerlendirme	General Evaluation	458	genel, okul, her şey
3	Fiziksel İmkanlar	Physical Facilities	364	bina, tesis, sınıf, bahçe
4	Öğretmen Kadrosu	Teaching Staff	360	öğretmen, hoca, öğretmenler
5	Sosyal ve Kültürel Etkinlikler	Social & Cultural Activities	164	etkinlik, gezi, kermes
6	Yönetim ve İdare	Management & Administration	100	müdür, yönetim, idare
7	Yabancı Dil Eğitimi	Foreign Language Education	91	ingilizce, yabancı dil, dil eğitimi
8	Personel ve Hizmet	Staff & Service	81	personel, hizmet, güler yüzlü
9	Güvenlik	Safety & Security	67	güvenlik, güvenli, kamera
10	Ücret ve Finansal	Fees & Financial	58	ücret, fiyat, para
11	Yemek ve Beslenme	Food & Nutrition	55	yemek, kantin, beslenme
12	Rehberlik ve Psikolojik Destek	Counseling & Psychological Support	52	rehber, psikolojik, danışman
13	Temizlik ve Hijyen	Cleanliness & Hygiene	46	temizlik, hijyen, tuvalet
14	Veli İletişimi	Parent Communication	46	veli toplantısı, iletişim, bilgi
15	Öğrenci Gelişimi	Student Development	44	gelişim, kişisel, sosyal beceri
16	Ders Çeşitliliği	Course Variety	34	seçmeli ders, kurs çeşitliliği
17	Ulaşım ve Konum	Transportation & Location	30	servis, ulaşım, konum
18	Disiplin	Discipline	26	disiplin, kural, ceza
19	Teknoloji	Technology	12	teknoloji, tablet, bilgisayar

Frequencies are computed from the final dataset of (sentence, aspect) pairs and sum to 2620; example expressions are illustrative and are reproduced in the original Turkish; the English category name in the third column serves as their gloss.

3.3. Annotation and Inter-Annotator Agreement

The corpus was annotated by the two authors of this study, who come from different educational backgrounds and hold qualifications at bachelor's and doctoral level. Annotation worked from a written guideline prepared in advance. The guideline set out the procedure to be followed—how aspect expressions were to be identified and how polarity was to be assigned—but deliberately gave no indication of the label expected for any particular expression, so that the annotators' judgments would not be steered. Reviews in which the target of an evaluation could not be determined had been screened out before this stage (Section 3.1), so annotation did not require ambiguous aspect expressions to be resolved. The two authors annotated the corpus jointly rather than independently: they worked through the reviews together and agreed each aspect expression and its polarity as it was assigned. The corpus therefore carries a single agreed annotation, and there is no internal disagreement rate to report for it. Polarity was assigned to each aspect expression

separately rather than to the review as a whole, so a review may carry expressions of opposite polarity; 242 of the 1533 reviews (15.8%) do. Because the annotation was carried out by the authors and provides no independent check of its own, label reliability was additionally assessed by two experts who took no part in it.

To independently validate label reliability, 200 samples were selected from the dataset using stratified sampling (140 positive, 60 negative) and sent to two independent domain experts. Each expert saw only the aspect category and sentence text—no access to the original labels or each other's decisions. Raw agreement and Cohen's Kappa are computed between the two annotators and therefore do not depend on any reference label. The per-annotator figures reported below are a separate quantity: they measure each expert's agreement with the original reference labels for the same 200 instances. Those labels were produced by the authors; since no externally established gold standard exists for this corpus, they are referred to here as the original reference labels rather than as ground truth, and the figures as agreement rather than accuracy. They are released with the dataset, and they are the labels used in all experiments: the adjudication of the 26 disagreements characterized those cases and did not replace the corpus labels.

Agreement was measured using Cohen's Kappa (κ) [8], which adjusts for chance agreement and is therefore more informative than a raw agreement percentage. For two annotators assigning one of two labels, κ is defined as

$$\kappa = (p_o - p_e)/(1 - p_e)$$

where p_o is the observed proportion of agreement and p_e the proportion expected if the two annotators labeled independently, p_e = Σ_c P_1(c)P_2(c) with P_k(c) the share of instances annotator k assigned to class c. A value of 0 indicates agreement at chance level and 1 indicates perfect agreement. The resulting κ = 0.702 falls in the Substantial Agreement range (0.61–0.80) as defined by Landis and Koch [9]. Estimated from 200 instances, the coefficient carries appreciable sampling uncertainty: a bootstrap over 10,000 resamples of the subset gives a 95% confidence interval of [0.590, 0.804], and the analytical normal approximation gives [0.595, 0.809]. The interval therefore extends from the upper part of the moderate-agreement band to the lower edge of the almost-perfect band, and the point estimate should be read with that width in mind. Disagreements occurred at a higher rate on negative-truth samples (9 of 60, 15.0%) than on positive-truth samples (17 of 140, 12.1%), consistent with the implicit nature of negative sentiment in Turkish educational discourse. Results are visualized in Figures 1–3. Table 3 summarizes the agreement statistics for the annotated subset.

Figure 1 shows the inter-annotator agreement outcomes. The gauge places Cohen's Kappa (κ = 0.7018) within the Substantial Agreement band of the Landis and Koch scale. Across the 200-sample stratified subset, the two annotators reached consensus on 174 cases (87.0%), with individual agreement with the original reference labels of 93.0% and 94.0%, respectively. The 26 conflicting instances were adjudicated by the first and second authors, who reviewed each case jointly—with the sentence, its aspect category and both annotators' labels in view—and agreed on a final polarity. Adjudication produced 17 positive and 9 negative labels. The adjudicated label coincided with Annotator 1's judgment in 12 of the 26 cases and with Annotator 2's in the remaining 14, an almost even split indicating that adjudication did not systematically favor either annotator. The conflicts were also dispersed rather than concentrated, spanning 13 of the 19 aspect categories, which suggests that they arose from the difficulty of individual instances rather than from a structural weakness in any one category.

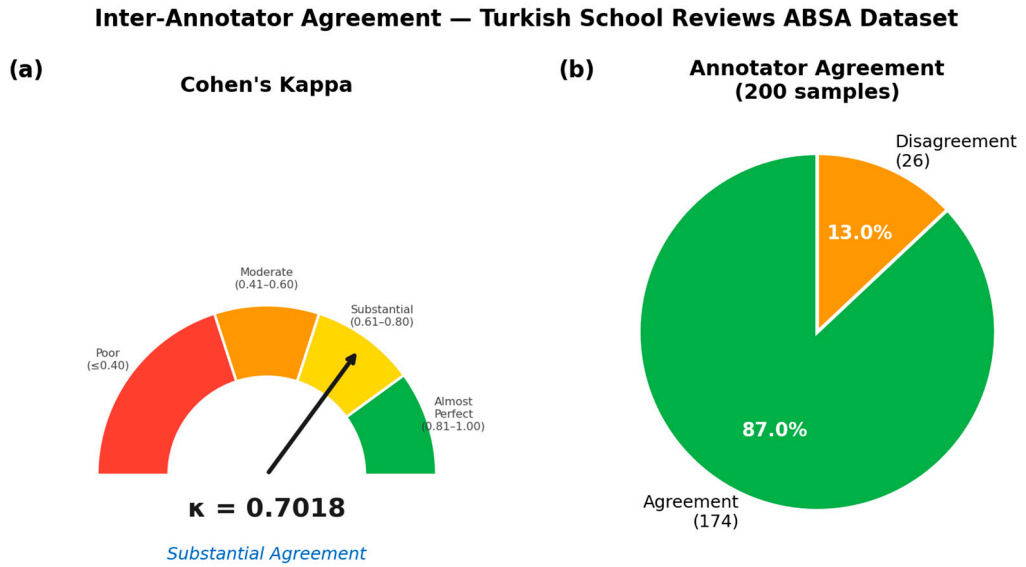


Figure 1. Inter–annotator agreement summary. (**a**) Cohen's Kappa positioned on the Landis and Koch scale; (**b**) proportion of agreeing and disagreeing judgments over the 200–instance subset. The corresponding numerical values are reported in Table 3 and are not repeated here.

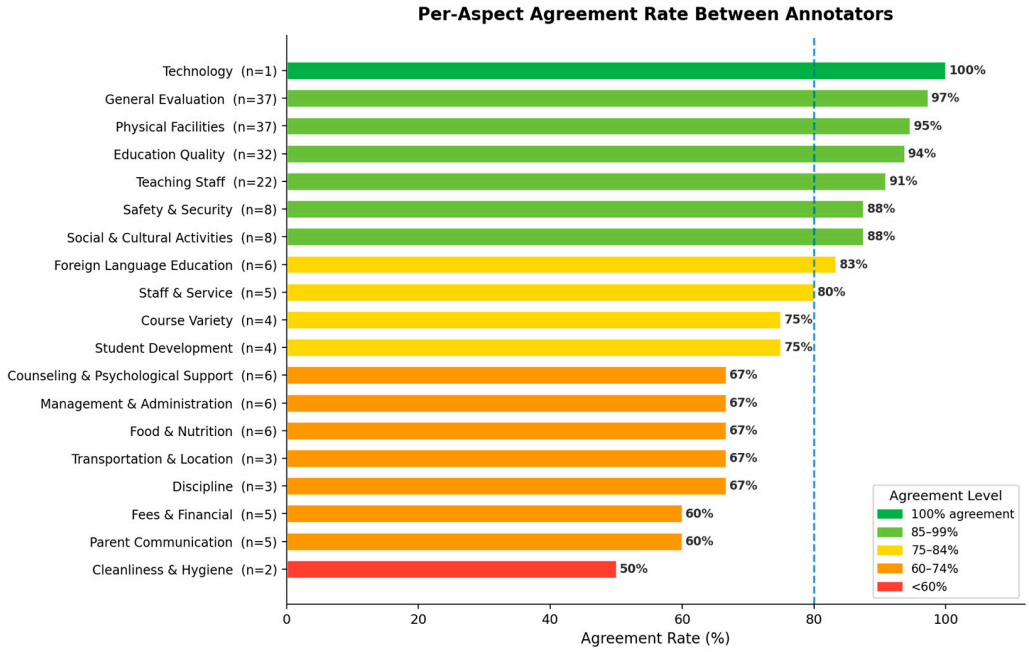


Figure 2. Per–aspect agreement rates.

Figure 2 shows the per-aspect agreement rates across all 19 annotation categories. Four categories—General Evaluation (97%), Physical Facilities (95%), Education Quality (94%), and Teaching Staff (91%)—exceeded the 90% threshold, together accounting for 128 of the 200 sampled instances. Safety & Security and Social & Cultural Activities both reached 88% agreement, and Foreign Language Education reached 83%. Staff & Service (80%), Course Variety (75%), and Student Development (75%) showed moderate agreement. Five categories—Counseling & Psychological Support, Management & Administration, Food & Nutrition, Transportation & Location, and Discipline—converged on 67% agreement, while Fees & Financial and Parent Communication both reached 60%. The single-instance Technology category (n = 1) was excluded from trend interpretation due to insufficient sample size, leaving Cleanliness & Hygiene as the lowest-agreement category at 50% (n = 2). The dashed reference line at 80% indicates that agreement was concentrated in the

high-frequency, general-purpose categories, while several lower-frequency, more subjective categories fell below this threshold.

Figure 3 shows the label distribution between the two annotators across the 200–sample subset. The diagonal cells represent concordant labels: 123 positive–positive and 51 negative–negative agreements, yielding 174 matched pairs in total (87.0%). Of the 26 disagreements, 17 cases involved Annotator 1 assigning a negative label while Annotator 2 assigned positive, compared to only 9 cases in the reverse direction. This asymmetry indicates that Annotator 1 applied a somewhat stricter labeling criterion, showing a greater tendency to interpret ambiguous expressions as negative—consistent with the marginally lower agreement with the original reference labels of 93.0% relative to Annotator 2 (94.0%).

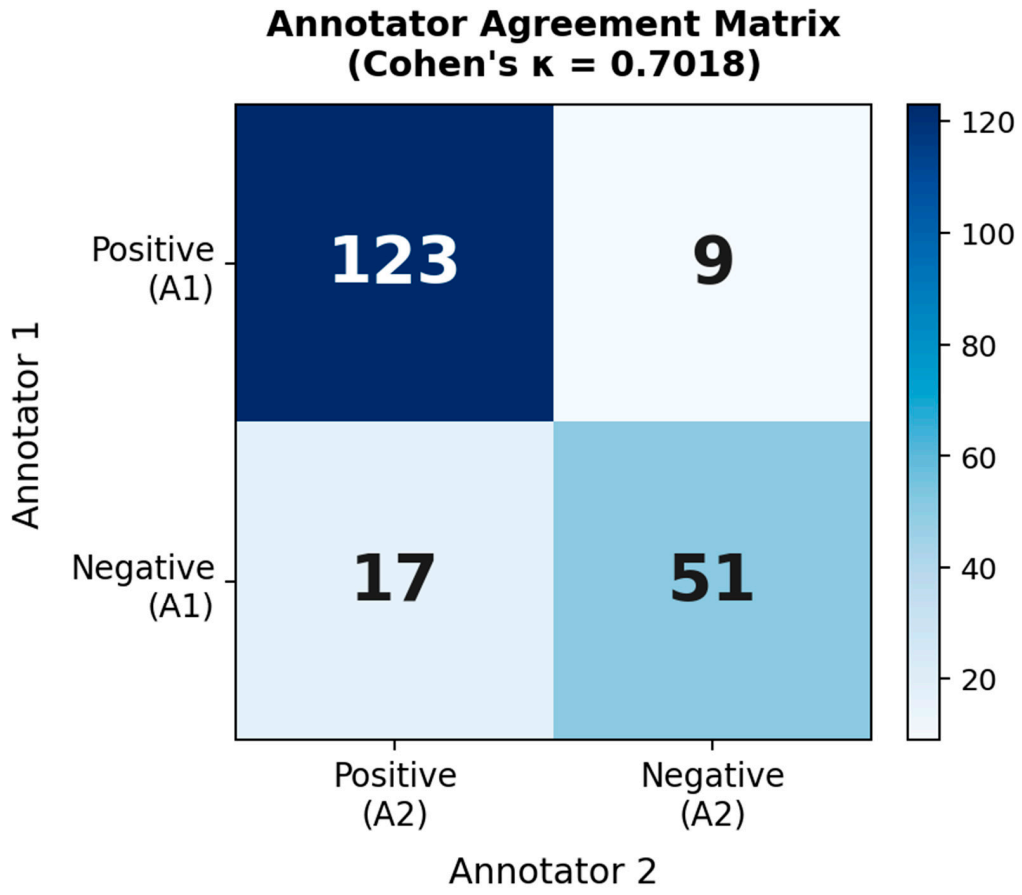


Figure 3. Annotator confusion matrix.

Table 3. Inter-annotator agreement results.

Metric	Value
Total samples	200
Agreements	174
Disagreements	26
Agreement rate	87.0%
Cohen's Kappa (κ)	0.702
Interpretation	Substantial Agreement

3.4. Dataset Description and Analysis

After aspect grouping and final labeling, the dataset consists of 2620 labeled (sentence, aspect) pairs derived from 1533 unique reviews. The class distribution is 69.6% positive (1824 samples) and 30.4% negative (796 samples), yielding a roughly 2.3:1 class imbalance

that is addressed during training (see Section 3.6). Table 4 summarizes the principal descriptive statistics of the final dataset.

Table 4. Dataset statistics.

Property	Value
Total instances	2620
Unique reviews	1533
Aspect categories	19
Positive instances	1824 (69.6%)
Negative instances	796 (30.4%)
Mean sentence length (per aspect pair)	~13 words (median = 12)
Review length, mean/median	12.0/11 words
Review length, p25/p75/p95	9/14/22 words
Review length, min/max	2/39 words
Mean aspects per review	1.71
Reviews spanning more than one aspect category	794 (51.8%)
Reviews containing both polarities	242 (15.8%)
Distinct aspect combinations	227
Aspect distribution entropy	3.50 of 4.25 bits (normalized 0.82)

Sentence length is computed across all 2620 (sentence, aspect) pairs, so reviews with multiple aspects are counted once per aspect; see Figure 4 for the length distribution computed per unique review (1533 reviews).

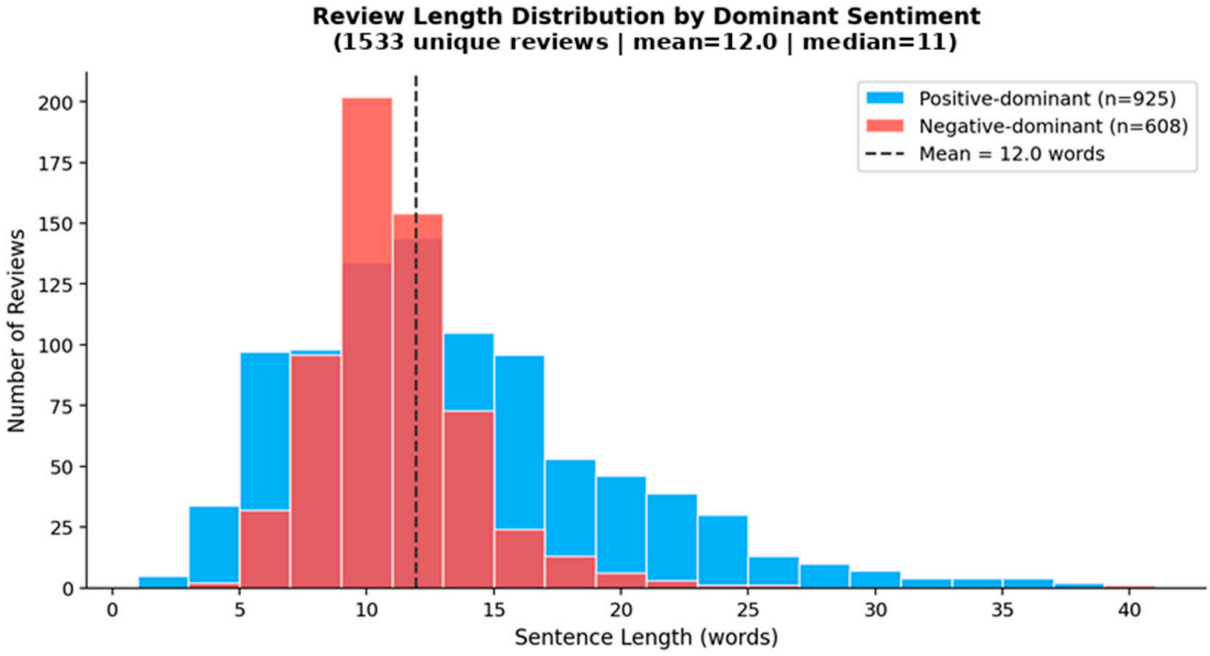


Figure 4. Sentence length distribution.

Figures 4–9 present a statistical overview of the dataset. The aspect frequency distribution (Figure 6) shows that Education Quality (n = 532) and General Evaluation (n = 458) are the most frequently annotated categories. The sentence length distribution (Figure 4) indicates that reviews have a mean length of 12.0 words (median = 11), with the majority concentrated in the 5–15-word range. The co-occurrence heatmap (Figure 9) reveals that the strongest co-occurrences are observed between Education Quality and General Evaluation (n = 104) and Education Quality and Teaching Staff (n = 100), suggesting that reviewers tend to evaluate these dimensions as part of a holistic educational experience.

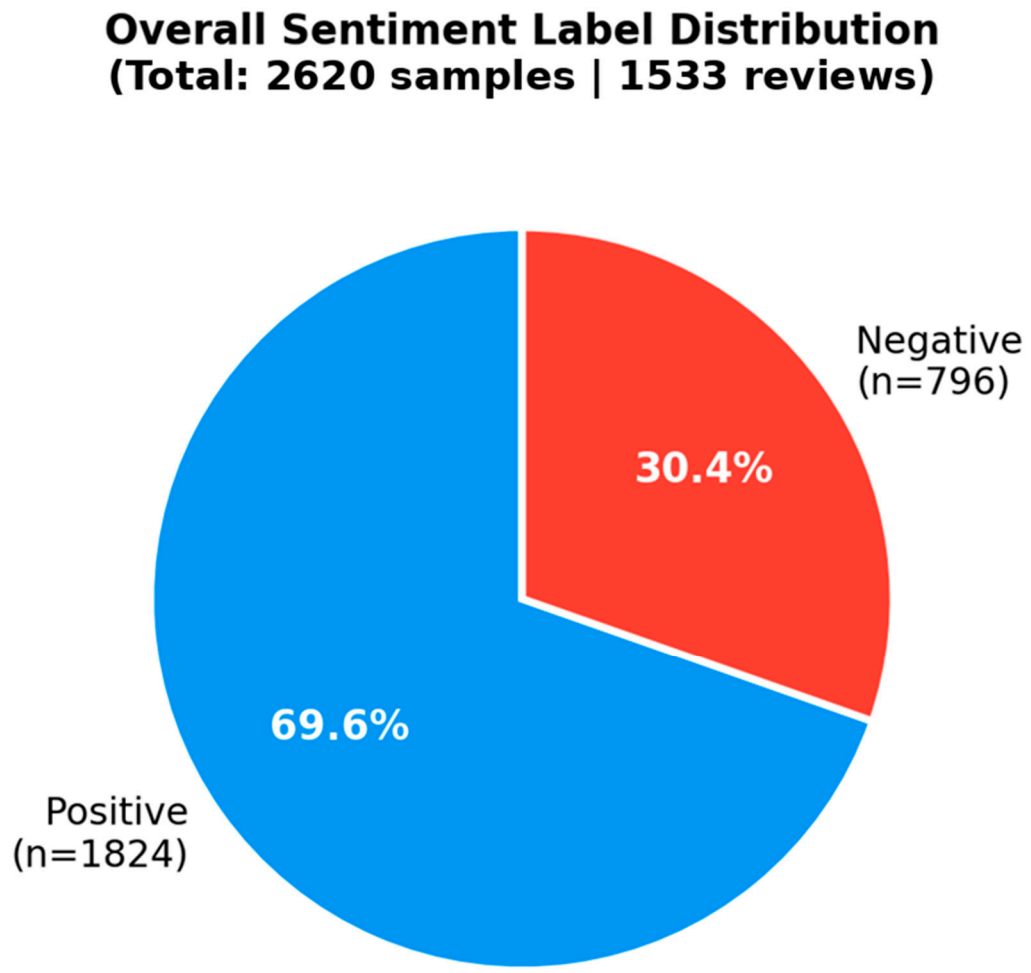


Figure 5. Label distribution.

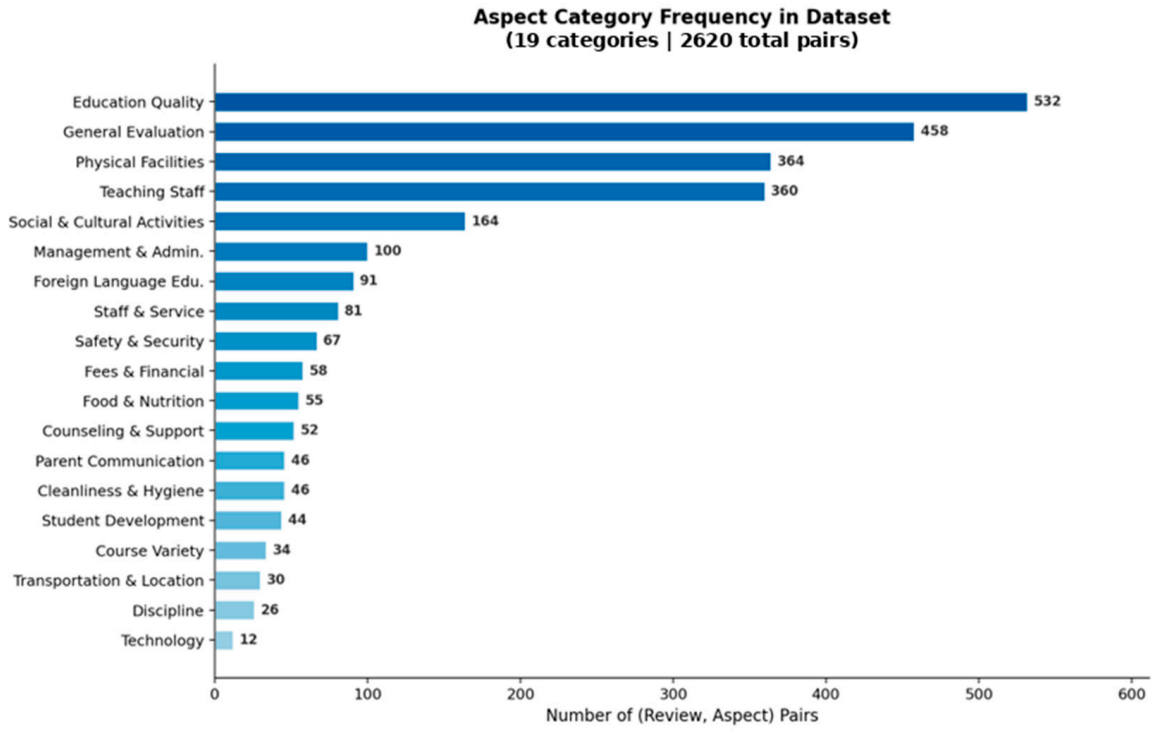


Figure 6. Aspect category frequencies.

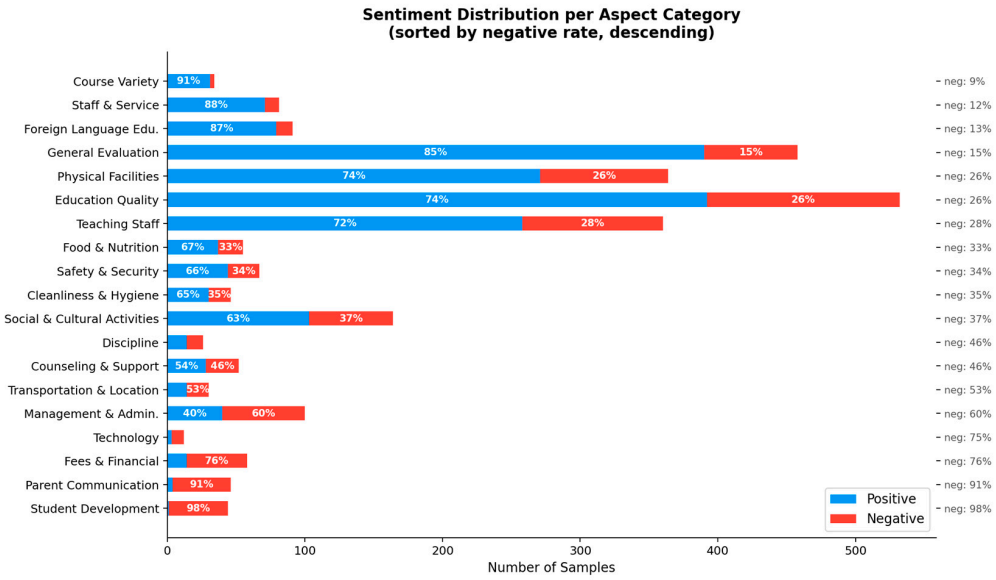


Figure 7. Per-aspect sentiment distribution.

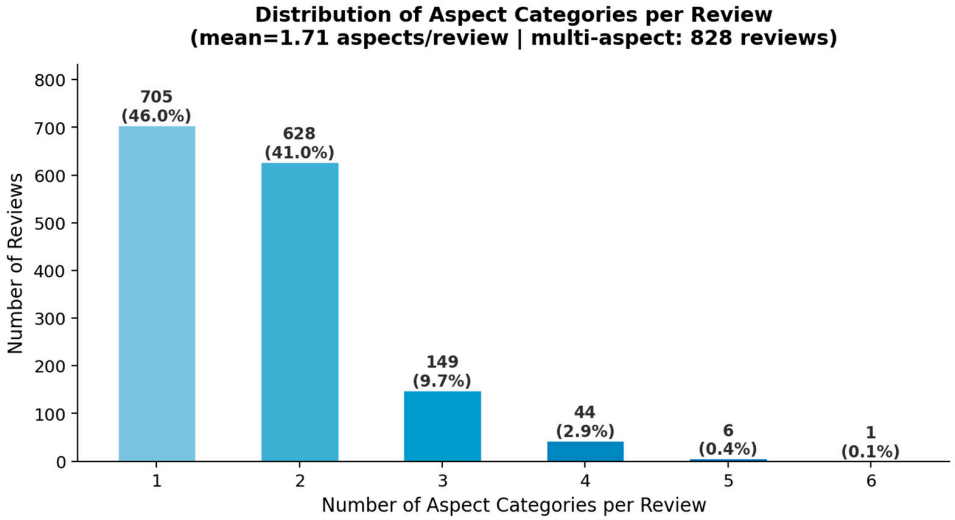


Figure 8. Number of aspect annotations per review.

Figure 5 shows the overall sentiment label distribution across all 2620 (sentence, aspect) pairs. Positive labels account for 69.6% of the dataset (n = 1824), while negative labels make up the remaining 30.4% (n = 796), yielding a class imbalance ratio of approximately 2.3:1. This skew toward positive sentiment reflects the nature of voluntary school reviews, where users tend to post when satisfied, and directly motivates the use of weighted loss during model training.

Figure 6 shows the frequency of each aspect category across the 2620 annotated pairs. Education Quality (532) and General Evaluation (458) are the two most represented categories, together accounting for nearly 38% of the dataset. Physical Facilities (364) and Teaching Staff (360) follow closely. At the other end, Technology (12), Discipline (26), and Transportation & Location (30) are the least frequent categories, which may limit the reliability of per-category model evaluation for these groups.

Figure 7 shows the sentiment distribution within each aspect category, sorted by descending negative rate. Student Development (neg: 98%), Parent Communication (neg: 91%), and Fees & Financial (neg: 76%) exhibit the highest proportions of negative sentiment, reflecting domains where parental dissatisfaction tends to be more explicitly expressed. Conversely, Course Variety (neg: 9%), Staff & Service (neg: 12%), and Foreign Language

Education (neg: 13%) are predominantly positive. High-frequency categories such as Education Quality and Teaching Staff show moderate negative rates of 26% and 28%, respectively, indicating that sentiment polarity varies substantially across aspect types despite the overall positive skew of the dataset.

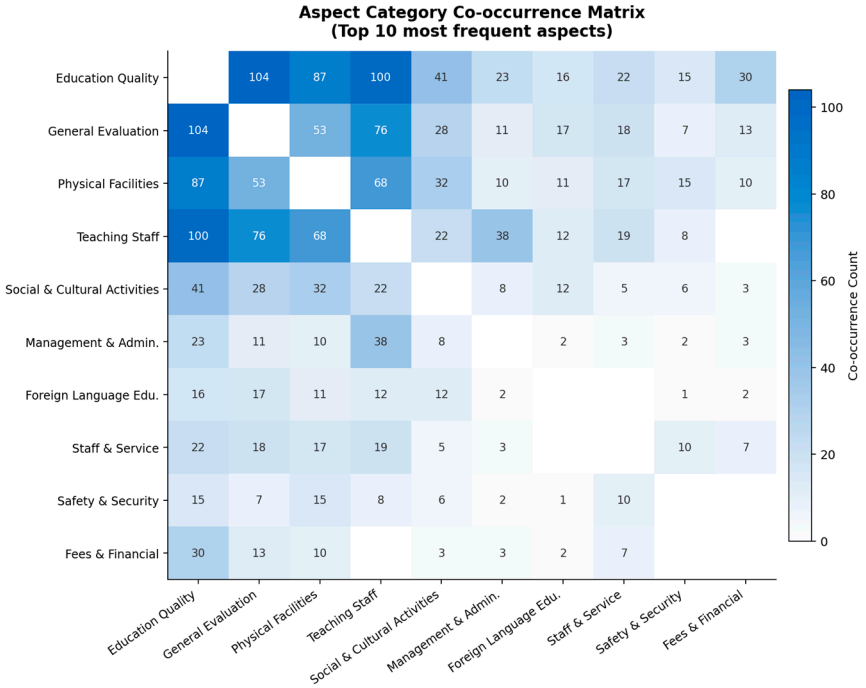


Figure 9. Aspect co-occurrence heatmap.

Figure 4 shows the review length distribution across 1533 unique reviews, grouped by dominant sentiment. The distribution is right-skewed, with a mean of 12.0 words and a median of 11 words; most reviews fall between 5 and 20 words. Positive-dominant (n = 925) and negative-dominant (n = 608) reviews follow a similar length profile, with both groups peaking in the 9–12-word range. The short, focused nature of reviews is consistent with the informal character of the platform and suggests that aspect-level sentiment is generally conveyed within brief sentences.

Figure 8 shows the distribution of aspect categories per review. Nearly half of all reviews (46.0%, n = 705) address a single aspect, while 41.0% (n = 628) cover exactly two. Reviews with three or more aspects account for the remaining 13.0%, with a small number reaching up to six categories. The mean of 1.71 aspects per review and the 828 reviews carrying more than one aspect annotation confirm that a single-label approach would have discarded a substantial share of the available semantic content, validating the multi-aspect annotation design.

Figure 9 shows the co-occurrence counts among the ten most frequent aspect categories. Education Quality and General Evaluation share the highest co-occurrence value (104), followed by Education Quality & Teaching Staff (100) and Education Quality & Physical Facilities (87). Teaching Staff & Management & Administration (38) also stands out as a notable pair, pointing to a perceived connection between instructional and administrative dimensions of school quality. Categories such as Foreign Language Education and Safety & Security show sparse co-occurrence with most others, consistent with their lower overall frequency in the dataset.

Beyond the marginal distributions, three properties characterize the corpus and bear on how the results should be read.

The first concerns how unevenly annotation mass is spread over the taxonomy. The five most frequent categories account for 71.7% of all pairs, while the ten least frequent together account for 15.4%; Education Quality (n = 532) is 44 times larger than Technology (n = 12). Measured on the category distribution, the entropy is 3.50 bits against the 4.25 bits a uniform distribution over 19 categories would give (normalized entropy 0.82, Gini coefficient 0.54). The distribution is therefore skewed but not degenerate: only two categories, Technology (n = 12) and Discipline (n = 26), fall below 30 instances, and together they represent 1.5% of the corpus.

The second concerns polarity, which is not distributed uniformly across categories. The corpus-level ratio of 2.3:1 conceals considerable variation: the proportion of negative instances ranges from 9% (Course Variety) to 98% (Student Development), and only 8 of the 19 categories fall within the 30–70% band. Categories describing concrete, provisioned services tend toward positive evaluations, whereas those describing developmental or relational outcomes attract criticism far more often. Aggregate accuracy figures therefore average over categories whose difficulty differs substantially.

The third concerns the structural diversity of the annotations themselves. The 1533 reviews yield 227 distinct combinations of aspect categories; 51.8% span more than one aspect category, and 15.8% carry both polarities within a single review. This last figure is the most consequential for the task formulation, since these are precisely the reviews on which a document-level classifier cannot produce a correct answer, and Section 4.4 shows that they are also where the models' residual errors concentrate.

Two limitations follow directly from these properties. Sparse categories cannot support reliable per-category estimates: 4 of the 38 (category, polarity) cells contain fewer than five instances, which is why performance is reported at corpus level in Section 4. And because the corpus is drawn from a single voluntary-participation platform, the positive skew and the register of the reviews reflect that source; the distributions reported here should not be read as properties of Turkish school feedback in general. These points are taken up again in Section 5.4.

3.5. Model Architectures

ABSA is not a single task but a family of subtasks (Section 2.1), of which this study addresses aspect sentiment classification: given a sentence and an aspect category drawn from a fixed inventory, the model assigns a polarity to that category. Aspect term extraction and aspect category detection are treated as upstream problems and are not attempted here.

This scoping follows from the question the study asks. The comparison of interest is between pre-training regimes—recurrent against transformer, and monolingual against multilingual—and such a comparison is interpretable only if the models are evaluated on identical inputs. In an end-to-end pipeline the polarity decision is conditioned on whatever aspects the extraction stage recovers, so a difference in final scores would confound the quality of the sentiment model with the quality of the extraction stage, and the effect of language-specific pre-training could not be isolated. Supplying the aspect category as input removes this confound and keeps the observed differences attributable to the models themselves.

The formulation also corresponds to the intended application. In a school-administration dashboard the categories are fixed in advance—the 19 canonical categories of Section 3.2—and the question put to the system is how a given dimension is evaluated in a given review, not which dimensions are present. The input encoding used throughout, [CLS] aspect [SEP] sentence [SEP], follows the auxiliary-sentence construction of Sun et al. [25] and is applied identically to all three transformer models.

What this scoping leaves out is stated in Section 5.4: a deployed system would additionally require aspect term extraction, evaluated jointly with the sentiment classifier, which falls outside the scope of this study.

3.5.1. Bidirectional LSTM (BiLSTM)

The Long Short-Term Memory (LSTM) network [21] addresses the vanishing gradient problem that limits standard RNNs from learning long-range dependencies. Its gating mechanism—forget, input, and output gates—allows the model to selectively retain or discard information across time steps. In the bidirectional variant, the model processes the sequence in both left-to-right and right-to-left directions, giving each time step access to both past and future context.

The BiLSTM model uses a mean pooling layer that averages all hidden states along the sequence to produce a fixed-size sentence representation, which feeds into a two-class linear classifier. Each input concatenates the aspect category and sentence text (sentence + " " + aspect), tokenized by whitespace up to a maximum of 50 tokens. Embedding size is 128 and dropout is 0.3. Grid search identified hidden size = 64, layers = 2 as the best configuration (~250K parameters).

3.5.2. Bidirectional GRU (BiGRU)

The Gated Recurrent Unit (GRU) [22] offers a streamlined alternative to LSTM. Using only reset and update gates—compared to LSTM's three—it contains fewer parameters per hidden unit and tends to be less prone to overfitting on small datasets. The architecture retains the same mean pooling layer and linear classifier as BiLSTM, replacing only the recurrent cell.

Grid search found hidden size = 256, layers = 2 to be optimal (~2.1M parameters). The difference in best hidden size between BiLSTM and BiGRU is not accidental: at the same hidden dimension, LSTM carries far more parameters than GRU due to its four weight matrices (three gates plus the candidate cell-state update, versus GRU's three). On a dataset of 2620 samples, a larger BiLSTM risks overfitting, while BiGRU's more compact gating structure generalizes better at higher capacity.

3.5.3. BERTurk

BERTurk (dbmdz/bert-base-turkish-cased) [36] is a cased BERT model [24] pre-trained on large Turkish web corpora. With approximately 110M parameters, 12 encoder layers, and 768-dimensional hidden representations, it produces context-sensitive token embeddings that capture rich morphological and semantic information. Each (aspect, sentence) pair is encoded in the sentence-pair classification format:

[CLS] aspect_category [SEP] sentence_text [SEP]

This formulation lets the model jointly attend to the aspect query and the sentence context during self-attention. The [CLS] token representation from the final encoder layer is passed to a randomly initialized two-class linear head, and the entire model is fine-tuned end-to-end.

3.5.4. mBERT

mBERT (bert-base-multilingual-cased) is a multilingual BERT model developed by Google and pre-trained on Wikipedia in 104 languages. It uses the same encoder architecture as BERTurk—12 encoder layers and 768-dimensional hidden representations—but its vocabulary of approximately 119K sub-word units is considerably larger, which raises the total parameter count to approximately 178M. This breadth enables cross-lingual transfer but requires the model to partition its representational capacity across over a hundred languages rather than specializing in one.

mBERT is evaluated under an identical fine-tuning protocol to BERTurk: the same [CLS] aspect [SEP] sentence [SEP] input format, the same learning rate (2×10^{-5}), and the same 5-fold cross-validation setup. This holds the fine-tuning procedure constant between the two models, so the comparison varies the pre-training language and data while keeping the training protocol fixed. It does not isolate pre-training as the sole source of any difference: the two models also differ in vocabulary size and in the composition of their pre-training corpora, and the learning rate common to both was selected on BERTurk rather than independently (Section 3.6).

3.5.5. XLM-R (XLM-RoBERTa)

XLM-R (xlm-roberta-base) is a multilingual model developed by Meta that substantially improves on mBERT [47]. It extends pre-training to 100 languages using 2.5 TB of CommonCrawl data—far beyond Wikipedia—while adopting the improved training recipe introduced by RoBERTa [29]: larger batch sizes, longer training, and dynamic masking. It has 12 encoder layers and 768-dimensional hidden representations, and its vocabulary of approximately 250K sub-word units brings the total parameter count to approximately 278M.

The key architectural distinction from mBERT is that XLM-R does not use `token_type_ids`. This is handled in the training script by a model-type flag that conditionally omits the segment-ID argument; all other hyperparameters and the input format remain identical to mBERT.

3.5.6. Rationale for the Selected Architectures

The five architectures were selected according to a single criterion: to vary the pre-training regime while holding the remaining design choices as constant as the comparison allows. That control is not complete, and Section 3.6 records where it is not: the learning rate was tuned on BERTurk and applied to mBERT and XLM-R without a separate search, so the comparison within the transformer group holds the training procedure constant but not the provenance of the value it uses. All five receive the same two inputs—the aspect category and the sentence—are trained with the same class-weighted loss, and are evaluated under the same review-level 5-fold protocol. The encoding of those inputs differs between the two families: the transformers use the [CLS] aspect [SEP] sentence [SEP] format of Sun et al. [25], while the recurrent models receive the sentence and the aspect concatenated as a single token sequence, having no segment embeddings or separator tokens to exploit. The comparison therefore holds constant the information available to each model rather than its encoding, and to that extent the encoding difference is confounded with the difference in representation origin. The models were therefore chosen for the position each occupies in the comparison rather than for expected performance, and none of them is redundant in that design.

The two recurrent models provide the reference point without pre-training. BiLSTM and BiGRU learn their representations from the 2620 instances of this corpus alone, which is what makes the comparison informative: the two families differ in pre-training as well as in architecture, capacity and representation strategy, so the contrast bounds how much of the difference can be ascribed to pre-training rather than isolating it: without a non-pre-trained baseline, the observation that transformers perform better would not distinguish the contribution of the architecture from that of the pre-trained representations. Both are included because they represent the two standard gating structures of the recurrent family (Section 2.2), and because their difference in parameter count allows the effect of capacity to be observed within that family.

BERTurk is the model under test. It is the component that carries the hypothesis of the study, namely that pre-training on a single morphologically rich language yields representations better suited to that language than multilingual pre-training of comparable architecture.

mBERT and XLM-R serve as the controls for that hypothesis, and two of them are required rather than one. The two differ substantially in the scale of their pre-training: mBERT is trained on Wikipedia across 104 languages, whereas XLM-R is trained on 2.5 TB of CommonCrawl data across 100 languages under an improved training recipe. Including both therefore turns a binary question into a measurable one, since it allows the monolingual advantage to be examined as a function of how much multilingual pre-training data the comparison model has seen. The results in Section 4.6 show that the advantage is not constant under this variation.

Two families of models were deliberately left out. Architectures developed specifically for ABSA—dependency-based graph convolutional networks and aspect-oriented attention mechanisms—were not included because, as discussed in Section 2.2, a thoroughly pre-trained transformer appears to learn much of the aspect–context relationship implicitly at this dataset scale, and because the question addressed here concerns the pre-training regime rather than architectural design. Large language models under zero- or few-shot prompting were also not included; this is recorded as a limitation in Section 5.4 rather than as a design choice, since it represents a comparison that would be informative but falls outside the scope of this study.

3.6. Training Setup and Cross-Validation Protocol

All models were trained and evaluated under a 5-fold cross-validation scheme. Folds are partitioned at the review (sentence_id) level—all aspect–sentiment pairs belonging to the same review always appear in the same fold, preventing data leakage between the training and test partitions at the review level. The eleven duplicated review texts described in Section 3.1 fall outside this mechanism, since they carry distinct review identifiers: in nine of the eleven pairs the same text appears in a training fold and in a test fold. The instances involved account for 26 of the 2620 pairs (1.0%). Within each training fold, 15% of training reviews are held out as a validation set; early-stopping monitors macro-F1 on this set throughout training. For the RNN models, the vocabulary is rebuilt from scratch using only that fold's training data, ensuring no token-level information from the test set leaks into the model. Table 5 lists the training hyperparameters shared by the two recurrent models.

Table 5. RNN training hyperparameters.

Hyperparameter	BiLSTM	BiGRU
Optimizer	Adam	Adam
Learning rate	1×10^{-3}	1×10^{-3}
Batch size	32	32
Max epochs	30	30
Early-stopping patience	5	5
Gradient clipping	1.0	1.0
Dropout	0.3	0.3
Hidden size	64	256
Layers	2	2

Hyperparameter selection. The configurations in Tables 5 and 6 were obtained from grid searches carried out before the cross-validation runs, on a single review-level 70/15/15 split of the corpus: for the recurrent models, three hidden sizes (64, 128 and 256) crossed

with one and two layers, giving six configurations; for the transformers, three learning rates (2×10^{-5}, 3×10^{-5} and 5×10^{-5}). In each search the configuration scoring highest on the held-out partition of that split was adopted and then applied unchanged to all five folds. The learning-rate search was run on BERTurk only: 2×10^{-5} was selected there and then applied to mBERT and XLM-R without a separate search for either. The three transformers were therefore trained under identical hyperparameters, but the value was tuned on one of them, and this asymmetry favors BERTurk to an extent the present design does not measure. That split and the cross-validation corpus share 1532 of 1533 reviews, so the selected configuration was chosen with reference to instances that subsequently appeared in test folds. This is a form of selection leakage, distinct from the review-level leakage that the fold partitioning prevents, and the cross-validation scores may be optimistic because of it. Section 4.9 quantifies the effect for the two recurrent models by repeating the selection inside each fold, and states why the same procedure was not computationally feasible for the transformers.

Table 6. BERTurk training hyperparameters.

Hyperparameter	Value
Optimizer	AdamW
Learning rate	2×10^{-5}
Weight decay	0.01
Batch size	16
Max epochs	10
Early-stopping patience	3
Gradient clipping	1.0
Warmup ratio	10%
Max token length	128

Class-imbalance handling: A weighted cross-entropy loss is applied in all models, with weights recomputed from each fold's training subset:

$$w_negative = (N_positive)/(N_negative),\ w_positive = 1.0$$

Because the weights are derived from each training fold rather than from the corpus as a whole, their values vary slightly across folds. The weight applied to the negative class took the values **2.279, 2.279, 2.381, 2.300 and 2.223** in folds 1–5 respectively (mean 2.292), tracking the fold-to-fold variation in the training-set class ratio; the positive class is always weighted 1.0. No other imbalance-handling strategy—oversampling, undersampling or focal loss—was applied, so the reported results reflect this single mechanism.

BERTurk fine-tuning uses AdamW [48] with a linear warmup–decay learning rate schedule: LR increases linearly over the first 10% of training steps, then decreases linearly to zero. This warmup prevents the pre-trained weights from being disrupted by large gradient updates early in fine-tuning. The complete fine-tuning configuration is given in Table 6, and the same settings are applied to mBERT and XLM-R. All experiments were run on a single Apple M4 Mac mini with 16 GB of unified memory. No GPU or Neural Engine acceleration was used: the training scripts fall back to the CPU backend when CUDA is unavailable, so every model reported here was trained on CPU. The resulting training costs are reported in Section 4.7.

All experiments were implemented in Python 3.14 using PyTorch 2.14 and the Hugging Face Transformers library. A base random seed of 42 was fixed across all runs; each fold was additionally initialized with a fold-specific seed (42 + fold index) to ensure independent randomness across folds.

3.7. Evaluation Metrics

The primary evaluation metric is Macro-F1, which averages the per-class F1 scores with equal weight regardless of class size. This is the appropriate choice under class imbalance: a classifier that simply predicts the majority (positive) class can achieve high accuracy on a 2.3:1 skewed dataset while completely failing on the minority class. Macro-F1 penalizes this behavior directly.

The following secondary metrics are also reported:

Accuracy: Proportion of correctly classified instances

F1-Positive: Binary F1 for the positive class

F1-Negative: Binary F1 for the negative (minority) class—the most informative indicator of how well a model handles the harder, underrepresented class

For a class c, let TP_c, FP_c and FN_c denote true positives, false positives and false negatives. The reported metrics are then

$$\text{Precision_c} = (\text{TP_c})/(\text{TP_c} + \text{FP_c}),\ \text{Recall_c} = (\text{TP_c})/(\text{TP_c} + \text{FN_c})$$

$$\text{F1_c} = 2\cdot(\text{Precision_c}\cdot\text{Recall_c})/(\text{Precision_c} + \text{Recall_c}),\ \text{Macro-F1} = (1)/(|\,C\,|)\ \Sigma_c \in C\ \text{F1_c}$$

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

with C = negative, positive. Because Macro-F1 averages F1_c over classes without weighting by class size, the minority class contributes as much as the majority class.

All metrics are reported as mean ± standard deviation across five folds, enabling joint assessment of both performance level and stability.

Use of generative AI. Generative AI tools were used in the preparation of this manuscript for translation between Turkish and English, for spelling and punctuation checking, and for formatting and consistency checks. The authors reviewed all such output and take full responsibility for the content of this paper.

Additional details regarding the experimental methodology are provided in the Supplementary Materials.

4. Results

4.1. Overall Model Performance

Table 7 presents the mean performance and standard deviations across the five folds. Four non-pre-trained baselines are included alongside the neural models: a majority-class predictor; a predictor that uses the aspect category alone, ignoring the sentence entirely; and TF-IDF representations classified by logistic regression and by a linear SVM. All are evaluated under the same review-level partitioning. Class weighting applies to the systems that are trained with a loss function; the majority-class and aspect-only predictors are rule-based and involve no training loss. Significance tests for the differences discussed below are reported in Section 4.8.

Table 7. Five-fold cross-validation results—mean ± standard deviation.

Model	Accuracy	Macro-F1	F1-Negative	F1-Positive
Majority class	0.6961 ± 0.019	0.4103 ± 0.007	0.0000 ± 0.000	0.8207 ± 0.013
Aspect category only	0.7463 ± 0.021	0.6147 ± 0.025	0.3897 ± 0.038	0.8398 ± 0.015
TF-IDF + Logistic Regression	0.8500 ± 0.014	0.8264 ± 0.013	0.7627 ± 0.018	0.8900 ± 0.012
TF-IDF + Linear SVM	0.8604 ± 0.015	0.8358 ± 0.017	0.7728 ± 0.025	0.8989 ± 0.012
BiLSTM (h = 64, L = 2)	0.8121 ± 0.022	0.7804 ± 0.020	0.6995 ± 0.033	0.8613 ± 0.025
BiGRU (h = 256, L = 2)	0.8292 ± 0.018	0.7999 ± 0.010	0.7258 ± 0.020	0.8741 ± 0.022
BERTurk	0.9382 ± 0.010	0.9266 ± 0.010	0.8975 ± 0.013	0.9557 ± 0.008

The majority-class predictor reaches 0.6961 accuracy but a Macro-F1 of only 0.4103, since it never predicts the minority class; this is the floor against which the other models should be read, and it illustrates why accuracy alone is not informative under a 2.3:1 imbalance.

The aspect-only predictor is more informative than the majority baseline, and is included for a specific reason. It assigns to each instance the majority polarity of its aspect category in the training folds, without seeing the sentence at all, and still reaches 0.7463 accuracy and a Macro-F1 of 0.6147. This reflects a genuine property of the corpus: polarity is unevenly distributed across categories (Section 3.4), and several categories are close to single-polarity—Student Development is 98% negative and Course Variety 91% positive. Any model with access to the aspect label therefore starts from a non-trivial floor, and the distance above that floor is the part attributable to reading the sentence. On that measure BERTurk gains 0.3119 Macro-F1 points over the aspect-only predictor and TF-IDF + SVM gains 0.2211, which indicates that neither is relying on the category label alone.

More consequential is the position of the two TF-IDF baselines, which exceed both recurrent models: TF-IDF with a linear SVM obtains a Macro-F1 of 0.8358 against 0.7999 for BiGRU, a difference of 0.0359 points, which the review-level analysis in Section 4.8 supports ($p = 0.0002$). The margin is nevertheless sensitive to how the recurrent configuration is chosen: it narrows to 0.0192 points when that choice is made within each fold (Section 4.9). Representations learned from scratch on 2620 instances are therefore not sufficient to outperform well-tuned surface features on this corpus. This does not diminish the role of the recurrent models in the comparison, which is to provide a neural reference point without pre-training (Section 3.5.6); rather, it sharpens the interpretation of the transformer results, since the gain they deliver is measured against a baseline family that is itself competitive. BERTurk exceeds the strongest classical baseline by 0.0907 Macro-F1 points.

BiGRU obtains a higher mean than BiLSTM by 0.0195 Macro-F1 points and 0.0171 accuracy points, although this difference falls within fold-level variance and is not statistically reliable (Section 4.8). The gap is larger on F1-Negative—0.7258 versus 0.6995—suggesting that the higher-capacity GRU architecture learns minority-class features more effectively on this dataset.

BERTurk's advantage over BiGRU is considerably larger: +0.1267 Macro-F1 (+15.8%), +0.1090 accuracy (+13.1%). The most striking gain is on F1-Negative, which jumps from 0.7258 to 0.8975—an increase of 0.1717 points (+23.7%). BERTurk therefore substantially outperformed BiGRU under the present experimental setting, and the margin is widest on the instances where negative sentiment is implicit. The two systems differ in pre-training, architecture, capacity and representation, so the observation is reported without attributing it to a single mechanism.

4.2. Per-Fold Results

Notably, BERTurk's weakest fold (Fold 4: 0.9123) exceeds BiGRU's best individual fold result (Fold 4: 0.8171). This relationship indicates that BERTurk's advantage is unlikely to be an artifact of favorable data splits: it holds across all five partitions.

Figure 10 presents the mean ± standard deviation across all four evaluation metrics for each model.

The error bars are also informative. They show the standard deviation of each metric across the five folds, not a confidence interval. BERTurk obtains both the highest scores and the smallest standard deviations: 0.008 on F1-Positive and 0.013 on F1-Negative. This combination of high performance and low variance is characteristic of a model whose

representations are already well-adapted to the target language prior to fine-tuning. Table 8 reports BERTurk's per-fold scores in full.

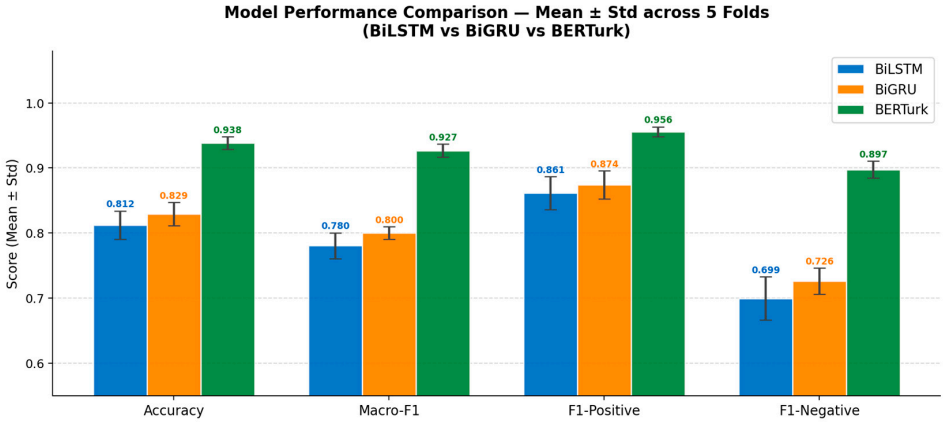


Figure 10. Mean ± standard deviation across all metrics—BiLSTM, BiGRU, and BERTurk. Error bars indicate standard deviation across five folds.

Table 8. BERTurk—per-fold detailed results.

Fold	Accuracy	Macro-F1	F1-Negative	F1-Positive	Epochs
1	0.9508	0.9399	0.9145	0.9654	10
2	0.9340	0.9213	0.8896	0.9529	9
3	0.9323	0.9234	0.8974	0.9495	10
4	0.9257	0.9123	0.8780	0.9465	5
5	0.9483	0.9359	0.9078	0.9640	9
Mean	0.9382	0.9266	0.8975	0.9557	~8.6

4.3. Training Dynamics

Figure 11 shows the training curves for BiLSTM and BiGRU on each model's best-performing fold.

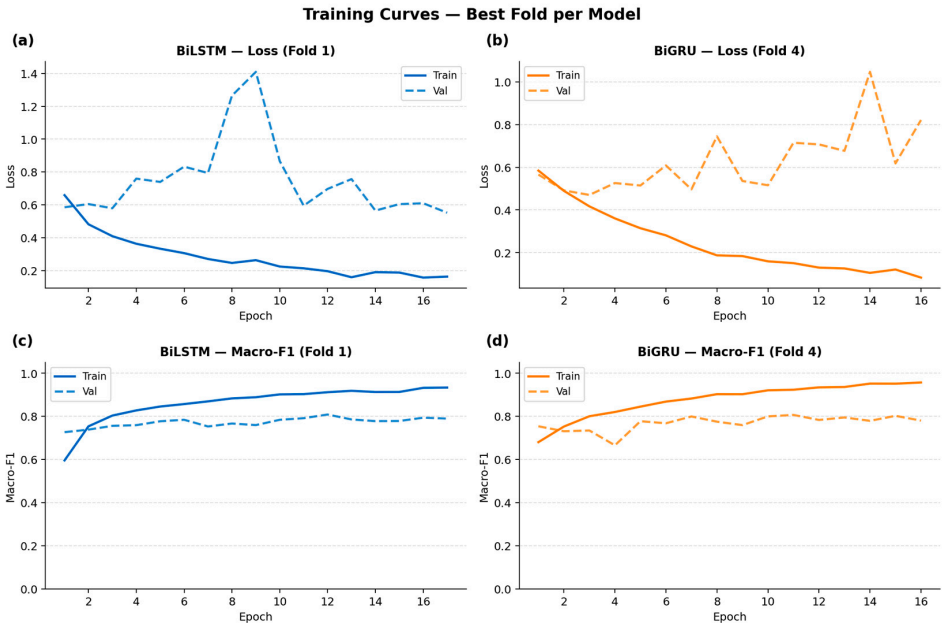


Figure 11. Training curves on each model's best fold. (**a**) BiLSTM cross-entropy loss; (**b**) BiGRU cross-entropy loss; (**c**) BiLSTM Macro-F1; (**d**) BiGRU Macro-F1. In all panels the solid line denotes the training set and the dashed line the validation set.

Both recurrent models converged smoothly, with BiLSTM triggering early stopping at an average of 12.4 epochs and BiGRU at 14.8. The gap between training and validation loss remained within acceptable bounds throughout, with no sign of severe overfitting—a reasonable outcome given the relatively small dataset size and the dropout regularization applied.

Figure 12 shows BERTurk's training curves for its best fold.

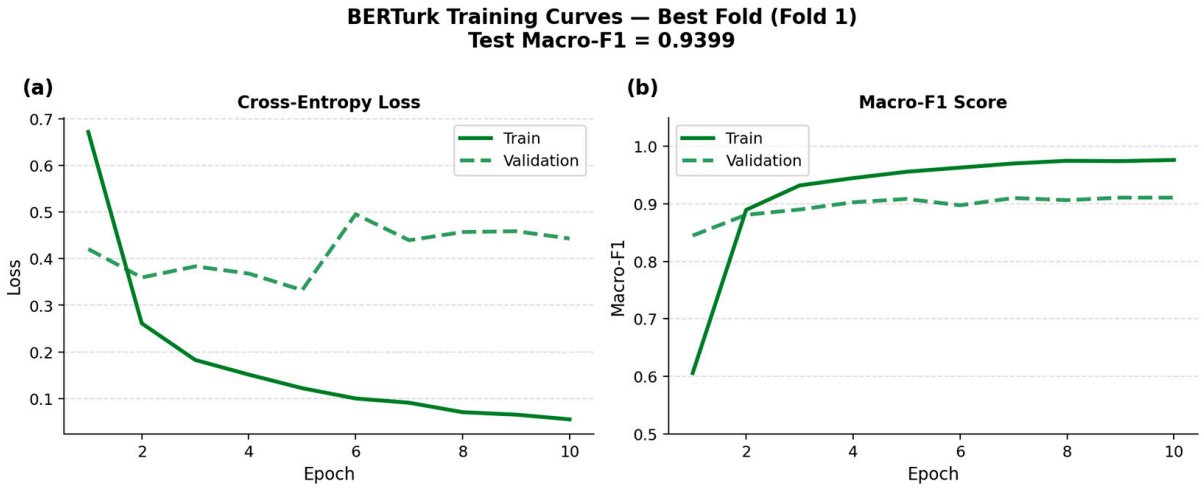


Figure 12. BERTurk training curves—best fold (Fold 1, Macro-F1 = 0.9399). (**a**) Cross-entropy loss; (**b**) Macro-F1. In both panels the solid line denotes the training set and the dashed line the validation set.

The learning dynamics differ noticeably from the recurrent models. It is the training Macro-F1 that climbs steeply over the first two to three epochs, rising from 0.61 to 0.93 on the best fold. Validation Macro-F1 behaves quite differently: it already stands at 0.84 after a single epoch—a direct consequence of the pre-trained representations transferring to the downstream task almost immediately—and improves only marginally thereafter, reaching 0.91 before plateauing. Early stopping was triggered at an average of 8.6 epochs. Fold 4 converged earliest at epoch 5, while folds 1 and 3 ran the full 10 epochs. From the second epoch onward, training loss falls below validation loss, and validation loss drifts upward in the later epochs; because early-stopping monitors validation Macro-F1 rather than validation loss, this did not translate into degraded test performance.

4.4. Error Analysis—Confusion Matrices

Figure 13 presents aggregated confusion matrices for BiLSTM and BiGRU, combining predictions from all five folds.

In absolute terms the larger share of errors falls on the positive class—for BiLSTM, 278 positive instances are classified as negative against 215 negative instances classified as positive—but this reflects the 2.3:1 class imbalance rather than the relative difficulty of the two classes. Normalized by class size the picture reverses: BiLSTM misclassifies 27.0% of negative instances against 15.2% of positive ones, and BiGRU 24.7% against 13.7%. The minority class is therefore close to twice as error-prone in both models. BiGRU lowers the error rate on the negative class relative to BiLSTM, as reflected in its higher F1-Negative (0.7258 vs. 0.6995), but the asymmetry persists. This is consistent with the known difficulty of detecting negative sentiment in Turkish educational reviews, where criticism is often expressed indirectly or embedded within otherwise neutral phrasing.

Figure 14 shows BERTurk's aggregated confusion matrix across all five folds.

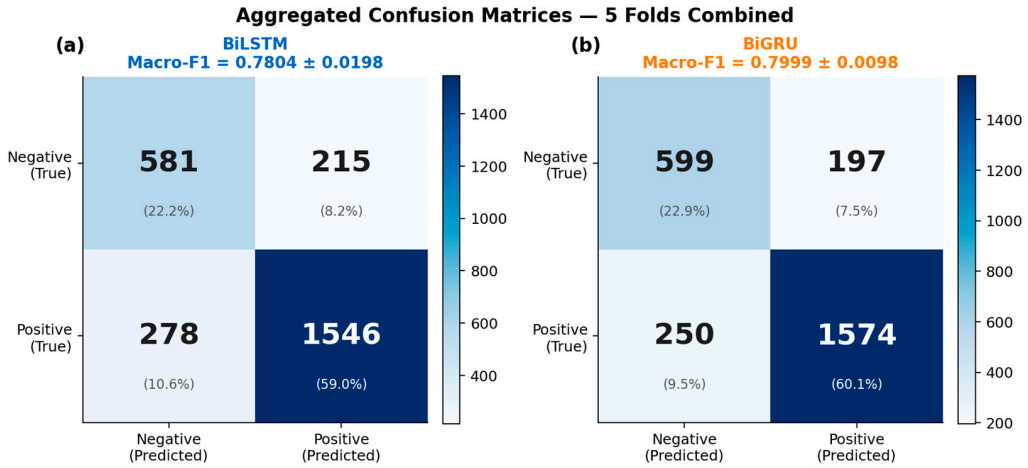


Figure 13. Aggregated confusion matrices over all five folds. (**a**) BiLSTM; (**b**) BiGRU. Values are raw instance counts and percentages are relative to the total sample. Rows correspond to the true class and columns to the predicted class, ordered negative then positive.

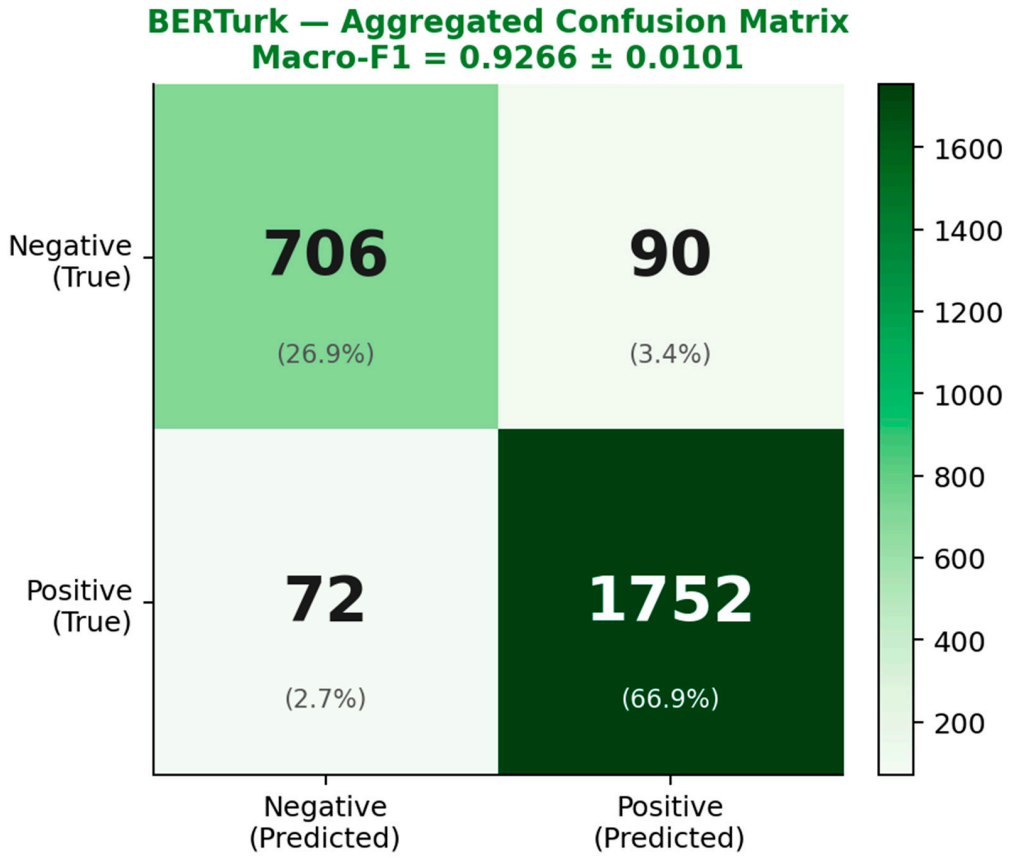


Figure 14. BERTurk aggregated confusion matrix (all five folds combined). Macro-F1 = 0.9266 ± 0.010. Rows correspond to the true class and columns to the predicted class, ordered negative then positive.

BERTurk reduces both error types substantially relative to the recurrent baselines: 90 negative instances are classified as positive and 72 positive instances as negative, against 215 and 278 respectively for BiLSTM. The error rate on the negative class falls from 27.0% to 11.3%, and on the positive class from 15.2% to 3.9%, indicating more balanced learning across both classes. The asymmetry between the classes nevertheless remains—the minority class is still misclassified at nearly three times the rate of the majority class—confirming that class imbalance and context-dependent negative sentiment continue to pose the primary challenge, even for a transformer-based model.

Forty of these errors are not resolvable by any classifier. Where consolidation placed two opposite-polarity expressions in the same canonical category (Section 3.2), the model receives an identical input—the same category and the same sentence—for two instances whose gold labels are opposite, so exactly one of the two must be counted as an error whatever the model predicts. The predictions bear this out exactly: across the five folds BERTurk is correct on one instance of each of the forty pairs, never on both and never on neither. These forty instances account for 24.7% of the 162 errors and place a ceiling of 0.9847 on the accuracy attainable on this corpus, against the 0.9382 BERTurk achieves. The remaining 122 errors are model errors, and their characteristics are described below.

To characterize these residual errors, we recovered all 162 instances that BERTurk misclassified across the five folds and examined them against the corpus as a whole. One property is markedly over-represented among the errors. Of the misclassified instances, 72.2% occur in sentences that carry aspects of both polarities, against 20.3% of the corpus overall—531 of the 2620 pairs; both percentages are counted over (sentence, aspect) pairs, the unit of classification; the same property counted over unique reviews gives the 242 of 1533 (15.8%) reported in Table 4—an over-representation of 3.6×. A closely related signal points the same way: 48.8% of the errors contain a contrastive connective (ama, ancak, fakat, rağmen, olsa da—"but", "however", "although", "despite", "even though"), against 16.6% of the corpus, an over-representation of 2.9×. Errors are therefore not spread evenly over the data; they concentrate in sentences whose sentiment is internally divided.

The following examples, drawn from the misclassified set, illustrate the mechanism (the target aspect is given in brackets, followed by the gold label and the model's prediction):

1. "Fiziksel imkanlar çok iyi, öğretmen kadrosu deneyimli ama okul müdürü bürokratik ve soğuk."—"The physical facilities are very good and the teaching staff experienced, but the principal is bureaucratic and cold." [Teaching Staff: positive → predicted negative]. The negative clause that follows the connective governs the sentence's closing tone, and the model transfers that tone to the target aspect even though the aspect itself is praised earlier in the sentence.

2. "Fiziksel imkanlar iyiydi ama kütüphane yalnızca belirli saatlerde açık, erişim kısıtlı."—"The physical facilities were good, but the library is open only at certain hours and access is limited." [Physical Facilities: negative → predicted positive]. Here a single aspect category carries both polarities within one sentence; the model settles on the polarity expressed first and misses the qualification that follows.

3. "Okulun odağı maalesef para. Yoksa eğitim kadroları da imkanları da kötü değil."—"The school's focus is unfortunately money. Otherwise neither the teaching staff nor the facilities are bad." [Physical Facilities: positive → predicted negative]. The aspect is endorsed only through litotes—the negation of a negative predicate (kötü değil, "not bad")—embedded in an otherwise critical passage, a construction that requires the model to resolve both the negation and the discourse shift signaled by yoksa.

The pattern is consistent across all three: the model attributes the dominant tone of the sentence to the target aspect rather than isolating the span that actually governs it. This pattern suggests that the difficulty lies in associating an opinion with the correct aspect rather than in recognizing sentiment as such, which may account for the persistence of these errors despite the gains that pre-training otherwise provides. It is also consistent with the premise of this work, since a document-level classifier would assign a single polarity to each of these sentences and thereby discard the distinction that the task requires.

4.5. Stability Analysis

Figures 15 and 16 visualize the fold-level distribution of Macro-F1 and F1-Negative for all models using box plots with individual fold points overlaid.

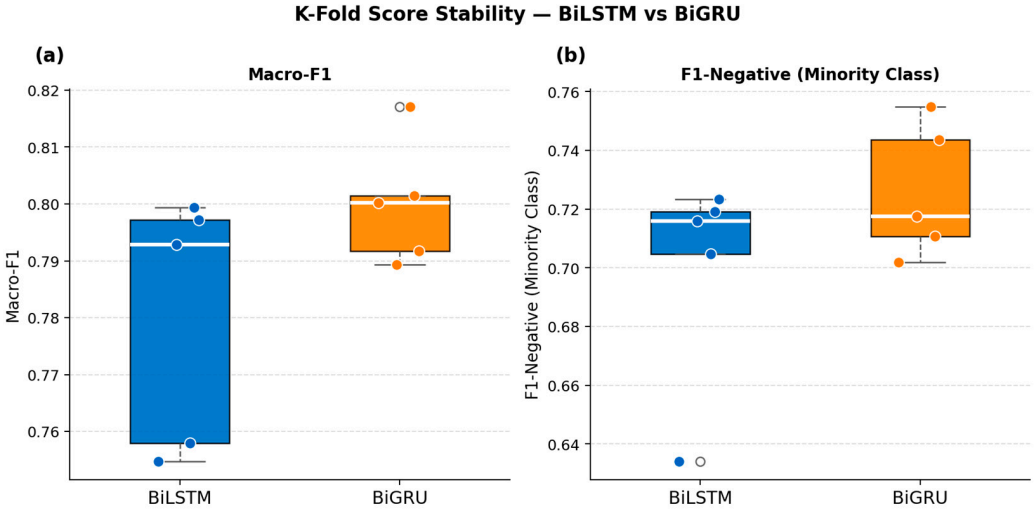


Figure 15. K-fold stability box plots—BiLSTM and BiGRU. Individual fold scores are shown as overlaid dots. (**a**) Macro-F1; (**b**) F1-Negative.

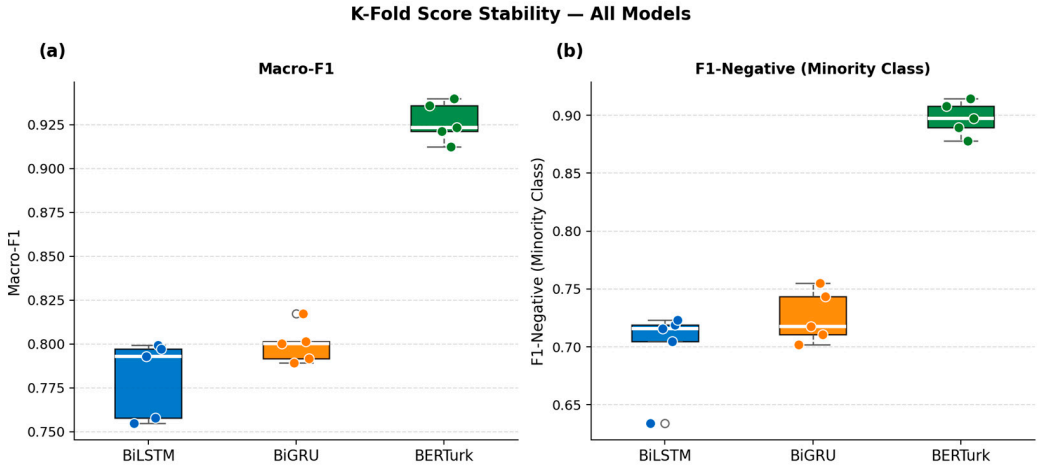


Figure 16. K-fold stability box plots for BiLSTM, BiGRU and BERTurk. (**a**) Macro-F1; (**b**) F1-Negative. All models are plotted on the same scale for direct comparison.

BiLSTM exhibits the widest spread across folds (Macro-F1 std = 0.020; F1-Negative std = 0.033), indicating that its performance is notably sensitive to how the data is partitioned. BiGRU tightens this variance considerably (std = 0.010 for Macro-F1), and BERTurk matches BiGRU's stability while operating at a substantially higher performance level—all five BERTurk folds fall between 0.91 and 0.94 Macro-F1. On F1-Negative, BiLSTM's high variance (std = 0.033) is particularly notable, as it suggests the model's ability to identify negative instances fluctuates meaningfully depending on which reviews appear in training versus test.

Two properties of the evaluation protocol bear on how much confidence these numbers carry, given the size of the corpus. First, under 5-fold cross-validation every one of the 2620 instances is held out exactly once, so the reported scores are computed over the entire corpus rather than over a single hold-out partition of a few hundred instances; the aggregate accuracy of 0.9382 for BERTurk carries a binomial 95% confidence interval of $\pm$0.0092. Second, the fold-level spread is narrow: BERTurk's Macro-F1 ranges from 0.9123 to 0.9399 across the five partitions (std = 0.0101). The principal risk when training on a small corpus is that performance becomes contingent on which examples happen to fall in the test set, and that contingency would surface as wide fold-level variance. It

does not appear here, which suggests that the ordering of the models is not an artifact of a particular partition.

Stability across folds does not by itself establish stability across the kinds of input the corpus contains, and the latter is examined here by partitioning BERTurk's pooled predictions along three axes. Performance is not uniform. It declines as sentences lengthen, from a Macro-F1 of 0.9511 on reviews of nine words or fewer to 0.9176 in the 10–15-word band and 0.8423 on reviews of sixteen words or more; and it is lower on reviews annotated with more than one aspect (0.9106) than on single-aspect reviews (0.9640). Both gradients point in the same direction as the error analysis of Section 4.4, where the residual errors were concentrated in long, internally divided reviews, and they delimit where the reported aggregate figures are optimistic.

The third axis requires care in interpretation. Grouping categories by frequency, Macro-F1 is 0.8918 on the four categories with at least 300 instances, 0.9588 in the 50–299 band and 0.9783 on the categories below 50. The apparent improvement on sparse categories is not evidence of better generalization: those categories are close to single-polarity, so the aspect-only baseline of Section 4.1 already performs well on them, and the high scores largely reflect that. The frequent categories—Education Quality, General Evaluation—are the semantically broad ones on which polarity is genuinely mixed, and they are correspondingly harder.

Because the task is aspect-based by definition, aggregate figures may conceal variation across the 19 categories. Table 9 therefore reports BERTurk's pooled performance for each category separately.

Table 9. BERTurk performance by aspect category (pooled over the five folds).

Aspect Category	*n*	Negative (%)	Accuracy	Macro-F1
Education Quality	532	26	0.908	0.877
General Evaluation	458	15	0.954	0.910
Physical Facilities	364	26	0.887	0.843
Teaching Staff	360	28	0.944	0.933
Social & Cultural Activities	164	37	0.933	0.928
Management & Administration	100	60	1.000	1.000
Foreign Language Education	91	13	0.978	0.952
Staff & Service	81	12	0.963	0.924
Safety & Security	67	34	1.000	1.000
Fees & Financial	58	76	0.897	0.851
Food & Nutrition	55	33	0.945	0.939
Counseling & Psychological Support	52	46	0.981	0.981
Cleanliness & Hygiene	46	35	1.000	1.000
Parent Communication	46	91	1.000	1.000
Student Development	44	98	1.000	1.000
Course Variety	34	9	0.941	0.817
Transportation & Location	30	53	0.933	0.933
Discipline	26	46	0.962	0.961
Technology	12	75	1.000	1.000

Categories are ordered by frequency. The negative share is given because it bounds how much the category label alone reveals about the polarity. Estimates for the smallest categories—Technology (n = 12) and Discipline (n = 26)—rest on too few instances to be treated as reliable, and the same caution applies to the minority class of strongly

skewed categories such as Student Development (one positive instance) and Parent Communication (four).

Two patterns are visible. The lowest scores belong to the largest and semantically broadest categories: Physical Facilities (0.843) and Education Quality (0.877), which together account for 34% of the corpus and are also the categories whose polarity is most genuinely mixed. Conversely, every category reaching a Macro-F1 of 1.000 is either small or close to single-polarity, so these values reflect the predictability of the category label as much as the model's reading of the sentence—the same effect quantified by the aspect-only baseline in Section 4.1. Course Variety is the clearest illustration: accuracy is high (0.941) but Macro-F1 falls to 0.817 because only three of its 34 instances are negative, and errors on that handful dominate the macro average.

4.6. Multilingual Model Comparison

This section presents a head-to-head comparison of BERTurk against two multilingual transformer alternatives—mBERT and XLM-R—fine-tuned on the identical task under the same hyperparameter configuration (LR = 2×10^{-5}, batch = 16, max epochs = 10, early-stopping patience = 3) and the same 5-fold cross-validation protocol. Table 10 reports the resulting head-to-head comparison.

Table 10. Multilingual comparison—BERTurk, mBERT, and XLM-R (5-fold mean ± std).

Model	Pre-Training	Accuracy	Macro-F1	F1-Neg.	F1-Pos.
mBERT	104 languages	0.9195 ± 0.009	0.9033 ± 0.010	0.8640 ± 0.015	0.9427 ± 0.007
XLM-R	100 languages	0.9344 ± 0.011	0.9207 ± 0.014	0.8877 ± 0.022	0.9536 ± 0.008
BERTurk	Turkish only	0.9382 ± 0.010	0.9266 ± 0.010	0.8975 ± 0.013	0.9557 ± 0.008

Figure 17 shows the mean ± standard deviation bar chart across all four metrics.

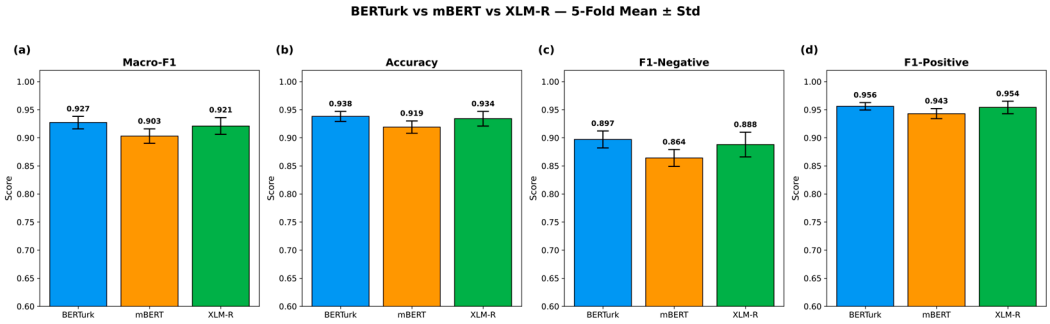


Figure 17. Five-fold mean ± standard deviation for BERTurk, mBERT and XLM-R. (**a**) Macro-F1; (**b**) Accuracy; (**c**) F1-Negative; (**d**) F1-Positive. Error bars show the standard deviation across the five folds. BERTurk achieves the highest scores on every metric.

Figure 18 presents stability box plots for Macro-F1 across the five cross-validation folds.

The numerical comparison yields three observations, which the fold-level data support to differing degrees; the corresponding tests are reported in Section 4.8. First, BERTurk obtains the highest mean Macro-F1 (0.9266 ± 0.010), ahead of XLM-R (0.9207 ± 0.014) by +0.006 points and mBERT (0.9033 ± 0.010) by +0.023 points; of these two margins, only the one over mBERT is substantial relative to fold-level variance, whereas the margin over XLM-R falls within it. Second, XLM-R obtains higher means than mBERT on every metric—+0.017 Macro-F1 and +0.024 F1-Negative—which is consistent with the benefit of XLM-R's 2.5 TB CommonCrawl pre-training corpus over mBERT's Wikipedia-only data, and this difference is established by the review-level analysis of Section 4.8

(p = 0.0050). Third, XLM-R exhibits notably higher fold-level variance (Macro-F1 std = 0.014; F1-Negative std = 0.022) than BERTurk (std = 0.010 and 0.013, respectively), which suggests that monolingual pre-training yields not only a higher mean but also more consistent minority-class detection across data splits.

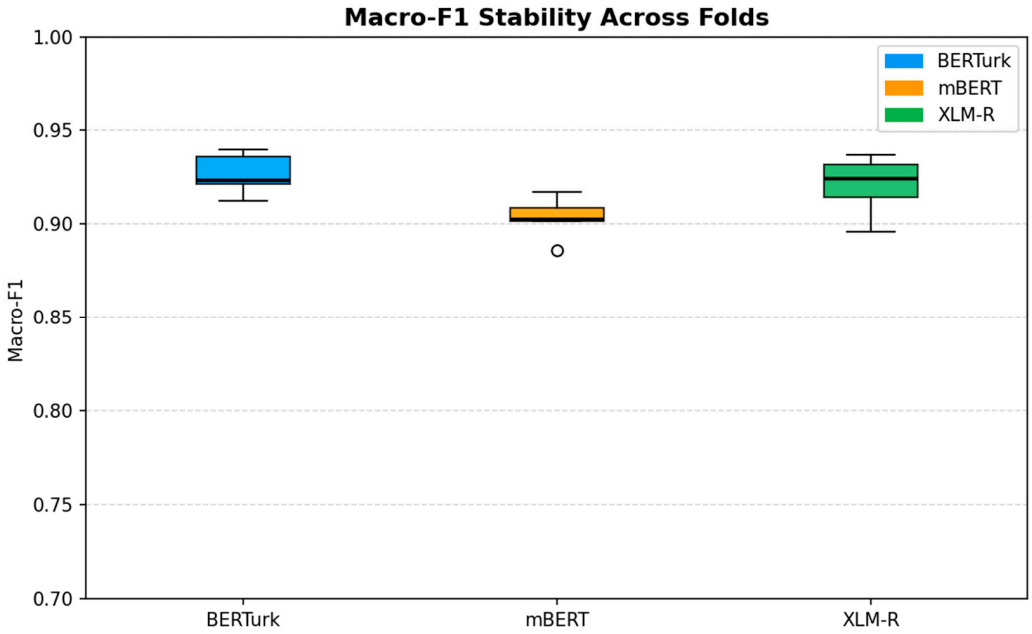


Figure 18. Macro-F1 stability box plots for all three models. BERTurk demonstrates both a higher median and a narrower interquartile range than either multilingual alternative.

These results align with the broader NLP literature showing that language-specific pre-training yields superior representations for morphologically rich languages [12,32]. Turkish is an agglutinative language in which a single surface form can encode negation, modality, and case simultaneously through stacked suffixes. mBERT and XLM-R represent such forms through a shared multilingual sub-word vocabulary (BPE/SentencePiece), which may fail to segment Turkish morphemes in linguistically meaningful ways. BERTurk, pre-trained on approximately 44 GB of Turkish text, uses a dedicated 32,000-token vocabulary derived from Turkish alone. This account is offered as a plausible mechanism consistent with the observed pattern rather than as one demonstrated here: the present study does not measure morpheme segmentation quality, tokenizer fertility or suffix representation, and establishing that the vocabulary is the operative cause would require experiments that isolate it from the other ways in which these models differ.

The F1-Negative metric most clearly illustrates this dynamic: the gap between BERTurk and mBERT is +0.034 points on F1-Negative versus only +0.019 on accuracy—approximately twice as large. Given that negative sentiment in Turkish educational reviews is frequently expressed indirectly or through subtle morphological markers, accurately detecting the minority class requires precisely the kind of fine-grained morphological and contextual understanding that monolingual pre-training provides. These findings support a preference for language-specific pre-training in Turkish ABSA, subject to the qualification recorded in Section 4.8: the advantage over mBERT is established by the review-level analysis, whereas the margin over XLM-R is not supported by it. Read with those qualifications, the results remain informative for researchers working on low-to-medium resource Turkic languages.

4.7. Computational Cost

Because the practical value of a model depends on what it costs to train as well as on how well it performs, Table 11 reports the wall-clock training cost of all five architectures

under the protocol described in Section 3.6. All times were measured on the same machine (Apple M4 Mac mini, 16 GB unified memory, CPU backend) and cover the complete 5-fold run, including early stopping.

Table 11. Training cost—5-fold cross-validation on a single Apple M4 CPU.

Model	Parameters	Time per Fold	Total (5 Folds)	Macro-F1
BiLSTM (h = 64, L = 2)	~250K	0.8 min	4.0 min	0.7804
BiGRU (h = 256, L = 2)	~2.1M	5.1 min	25.5 min	0.7999
BERTurk	~110M	2.7 h	13.7 h	0.9266
mBERT	~178M	4.2 h	20.8 h	0.9033
XLM-R	~278M	6.8 h	33.9 h	0.9207

Two observations follow. First, the accuracy obtained by moving from a recurrent model to a transformer carries a substantial computational cost: BERTurk requires approximately 200 times the training time of BiLSTM for a gain of 0.146 Macro-F1 points. Where computational resources are limited, however, the recurrent models are not the appropriate economy choice: TF-IDF with a linear SVM obtains a higher Macro-F1 than either of them (0.8358 against 0.7999 for BiGRU, Table 7) without gradient-based training. Table 11 accordingly reports the five neural architectures, for which training cost is the quantity of interest.

Second, and more relevant to the monolingual-versus-multilingual comparison, BERTurk obtained both the highest Macro-F1 and the lowest training cost among the three transformers. It trains in 13.7 h against 33.9 h for XLM-R—a 2.5-fold difference—while scoring 0.006 Macro-F1 points higher. Because all inputs are padded to a fixed length of 128 tokens (Section 3.6), the three transformers process sequences of identical length, so the difference does not originate in tokenization. It follows vocabulary size instead: BERTurk's dedicated Turkish vocabulary contains approximately 32K sub-word units against roughly 119K for mBERT and 250K for XLM-R, and the corresponding embedding matrices account for approximately 25M, 92M and 192M parameters. The comparison between BERTurk and mBERT is the most informative of the three, since the two models share an encoder architecture and differ primarily in vocabulary size, and mBERT requires roughly half as much time again to train. Language-specific pre-training therefore improves the accuracy–cost trade-off along both axes at once, which strengthens the practical case made in Section 5.3.

It should be noted that these absolute times reflect CPU-only training and would fall by roughly an order of magnitude on a discrete GPU; the relative ordering of the models, however, is determined by parameter count and vocabulary size and would be preserved.

4.8. Statistical Significance of Model Differences

The comparisons in the preceding sections rest on differences in mean scores, and the question of which are supported by the data is addressed here. Two features of the design constrain how that question can be answered. Comparing the five fold-level scores provides a sample of five, which is small; and those five scores are not independent, because any two folds share three-fifths of their training data. The primary analysis reported here therefore does not use the fold scores at all.

Cross-validation assigns every instance in the corpus to exactly one test fold, so the five folds together yield one out-of-fold prediction per instance: 2620 predictions per system, each produced by a model that did not see that instance in training. These pooled predictions are compared directly. The resampling unit is the review rather than the (sentence, aspect) pair, because the 2620 pairs come from 1533 reviews and pairs

from the same review are not independent—they share the sentence, the author and the context. Resampling whole reviews preserves that dependence; resampling pairs would treat it as absent and overstate precision. Two tests are reported: a bootstrap over 10,000 resamples of the 1533 reviews, giving a 95% confidence interval for the difference in pooled Macro-F1, and a permutation test over 10,000 draws in which the two systems' predictions are exchanged within randomly selected reviews, giving a two-sided p-value under the hypothesis that the systems are equivalent. The same seven comparisons are evaluated as before, and a Holm–Bonferroni correction is applied across them. Table 12 reports the results.

Table 12. Pairwise comparisons on pooled out-of-fold predictions (resampling unit: review, n = 1533).

Comparison	Δ Macro-F1	95% CI (Bootstrap)	p (Permutation)	p (Holm)	Sig.
SVM vs. Majority	+0.4257	(+0.4087, +0.4421)	0.0001	0.0007	yes
BERTurk vs. SVM	+0.0903	(+0.0756, +0.1052)	0.0001	0.0007	yes
BERTurk vs. mBERT	+0.0230	(+0.0106, +0.0347)	0.0002	0.0010	yes
SVM vs. BiGRU	+0.0342	(+0.0180, +0.0505)	0.0002	0.0010	yes
XLM-R vs. mBERT	+0.0174	(+0.0053, +0.0297)	0.0050	0.0150	yes
BiGRU vs. BiLSTM	+0.0197	(+0.0027, +0.0363)	0.0227	0.0454	yes
BERTurk vs. XLM-R	+0.0056	(−0.0040, +0.0153)	0.2688	0.2688	no

SVM denotes the TF-IDF + linear SVM baseline. Differences are computed on Macro-F1 over the pooled out-of-fold predictions, which differs by at most 0.0003 from the fold-averaged values in Table 7 (BERTurk 0.9264 pooled against 0.9266 averaged). With 10,000 permutations the smallest attainable p-value is 0.0001.

Six of the seven comparisons are supported and one is not. The separation between the pre-trained transformers and everything below them is large and unambiguous: BERTurk exceeds the strongest classical baseline by 0.0903 Macro-F1 points, with a confidence interval nowhere near zero. This group of comparisons supports a direct empirical statement—the pre-trained transformer models evaluated here outperformed the non-pre-trained baselines—and the manuscript makes no stronger claim from it. The systems differ simultaneously in architecture, capacity, tokenization and pre-training history, and the design does not isolate any one of these, so the difference is not attributed to a single mechanism.

The comparisons within the pre-trained group behave differently from one another. BERTurk's margin over mBERT (+0.0230) is established, as is XLM-R's over mBERT (+0.0174). BERTurk's margin over XLM-R (+0.0056) is not: its confidence interval spans zero and the permutation test returns p = 0.2688. The pattern is consistent with the reading given in Section 4.6—the advantage of Turkish-specific pre-training is evident against the smaller multilingual model and is not resolvable against the larger one.

The fold-level paired t-tests reported in the previous version of this analysis agree with Table 12 in direction for all seven comparisons, but reach the corrected threshold for only two of them. The difference is one of power rather than of finding: with five folds a paired comparison attains 80% power only for effects of dz $\geq$ 1.68, whereas the review-level analysis resamples 1533 units. The four comparisons that change status—BERTurk against mBERT, TF-IDF + SVM against BiGRU, XLM-R against mBERT, and BiGRU against BiLSTM—were previously described as directionally consistent but not established; the stronger test establishes them. We note that the overlap between training folds, which is sometimes expected to make fold-level tests anti-conservative, is not the dominant effect at this sample size; the loss of power from n = 5 is.

One comparison does not change. BERTurk against XLM-R is unsupported under both analyses, and the ranking within the transformer group is accordingly presented as provisional throughout.

4.9. Nested Cross-Validation of the Recurrent Configurations

The configurations used above were selected on a single held-out split rather than within each cross-validation fold (Section 3.6), and that split shares 1532 of its 1533 reviews with the corpus used for cross-validation. The selected configuration was therefore chosen with reference to instances that later appeared in test folds, and the reported scores may be optimistic as a result. This section quantifies the effect for the two recurrent models by repeating the selection inside each fold.

For each of the five outer folds, all six configurations in the grid were trained on that fold's training partition, and the configuration scoring highest on the fold's internal validation set was then evaluated once on the outer test fold. Selection used the validation set alone, so the outer test fold took no part in the choice. The same run also re-evaluated the published configuration on every outer fold, which allows the two protocols to be compared on identical data. Table 13 reports both.

Table 13. Nested versus fixed-configuration selection (Macro-F1).

Model	Fold	Configuration Selected in Fold	Nested	Fixed Configuration
BiLSTM	1	h = 64, L = 2	0.7994	0.7994
BiLSTM	2	h = 128, L = 2	0.7930	0.7929
BiLSTM	3	h = 64, L = 1	0.7771	0.7972
BiLSTM	4	h = 128, L = 2	0.7834	0.7579
BiLSTM	5	h = 64, L = 1	0.8541	0.7547
BiLSTM	mean	—	0.8014 ± 0.0274	0.7804 ± 0.0198
BiGRU	1	h = 128, L = 1	0.7960	0.8014
BiGRU	2	h = 64, L = 2	0.8191	0.7893
BiGRU	3	h = 64, L = 1	0.8229	0.7917
BiGRU	4	h = 128, L = 2	0.8189	0.8171
BiGRU	5	h = 256, L = 1	0.8262	0.8002
BiGRU	mean	—	0.8166 ± 0.0107	0.7999 ± 0.0098

The fixed configuration is h = 64, L = 2 for BiLSTM and h = 256, L = 2 for BiGRU. The fixed configuration column reproduces the per-fold values underlying Table 7 to four decimal places, so the two columns differ only in how the configuration was chosen.

Three observations follow. First, per-fold selection raises rather than lowers the recurrent scores: BiLSTM gains 0.0210 Macro-F1 points and BiGRU 0.0167. The published configuration was re-selected in one of five folds for BiLSTM and in none for BiGRU, which indicates that no single configuration in this grid is preferred across folds; selecting within each fold accommodates that variation, whereas a fixed choice does not. Neither increase is itself distinguishable from zero at this sample size (BiLSTM $p = 0.3722$, BiGRU $p = 0.0955$), so the difference between the protocols is better described as bounded than as established.

Second, correcting the bias does not weaken the central comparison of this study: BERTurk's margin over BiGRU falls from 0.127 to 0.110 Macro-F1 points and remains unambiguous (dz = 5.54, $p = 0.0002$). Whether the transformer scores carry a comparable bias was not tested, and Section 5.4 states this rather than assuming it away.

Third, one secondary comparison is affected. The margin of TF-IDF + SVM over BiGRU, given in Section 4.1 as 0.0359 points, falls to 0.0192 under nested selection, and its confidence interval moves from (+0.0048, +0.0670) to (−0.0066, +0.0451). Under the fixed-configuration protocol this margin is supported by the review-level analysis of Section 4.8. The nested comparison reported here rests on fold-level scores alone, because out-of-fold predictions were not retained for the configurations selected within each fold, and at that level the interval spans zero. The margin is therefore better described as sensitive to how the recurrent configuration is chosen than as settled in either direction. The corresponding comparison against BiLSTM is unaffected (+0.0344, CI +0.0014 to +0.0675), so the broader

observation that a well-tuned feature-based baseline is competitive with the recurrent models stands.

The same procedure was not applied to the three transformers. A nested search over the three learning rates examined in the original grid would require fifteen fine-tuning runs per model, which at the per-fold costs in Table 11 amounts to approximately 41 h for BERTurk, 62 h for mBERT and 102 h for XLM-R—some 205 h in total on the CPU-only hardware used here, against 1.6 h for the complete recurrent grid. The transformer results therefore retain the fixed-configuration protocol, and Table 7 reports all nine systems under that one protocol so that models are not compared across differing selection procedures. The possibility that the transformer scores carry a comparable selection bias is not excluded by the analysis above and is noted among the limitations in Section 5.4.

5. Discussion

5.1. Interpretation of Findings

Taken at face value, the numbers in Section 4 trace a clear ranking: BERTurk > XLM-R > mBERT > BiGRU > BiLSTM. What matters more than the ranking itself, however, is what drives it. The gap between the models trained from scratch and the pre-trained transformers—up to 0.15 Macro-F1 points—is too large to attribute to architectural depth alone; it points instead to the linguistic prior encoded during pre-training. The classical baselines added in Section 4.1 place this on a second footing: TF-IDF with a linear SVM exceeds both recurrent models, so the transformer advantage is not measured against neural models alone but against a competitive feature-based approach as well, and BERTurk still exceeds the strongest such baseline by 0.0907 Macro-F1 points (Section 4.8). BiLSTM and BiGRU must learn word-level representations from scratch using only the 2620 instances available in this corpus, whereas BERTurk enters fine-tuning already equipped with a broad model of Turkish agglutinative morphology. On a small, domain-specific corpus, this prior substantially offsets the constraint that dataset size would otherwise place on achievable performance.

The same pattern re-emerges, in a subtler form, between the monolingual and multilingual transformers, where architecture is held constant and the pre-training corpus differs in its language coverage (Section 4.6). The comparison is not fully controlled: as Section 3.6 records, the learning rate was tuned on BERTurk and transferred to the other two, and that asymmetry favors BERTurk by an amount this design does not measure. BERTurk's margin over mBERT (+0.023 Macro-F1) is established by the review-level analysis, whereas its margin over XLM-R (+0.006 Macro-F1) is not supported by it (Section 4.8); together these suggest that the benefit of language-specific pre-training is not a fixed quantity—it appears to scale inversely with how much pre-training data the multilingual competitor was exposed to. Rather than supporting an unqualified claim that monolingual pre-training is always preferable, the data point toward a more calibrated conclusion: the advantage shrinks as multilingual pre-training corpora grow.

Perhaps the most engineering-relevant observation is that gains on F1-Negative are consistently larger than gains on accuracy across every model comparison in Section 4.6. Model choice does not affect the two classes uniformly. This is a concrete demonstration that comparing classifiers on a single aggregate metric can be actively misleading under class imbalance, and it argues for reporting per-class metrics as a default practice rather than an afterthought in imbalanced classification tasks.

5.2. Relation to Prior Work

Table 14 places the results reported here alongside the Turkish ABSA studies reviewed in Section 2.5.

Table 14. Reported results in Turkish ABSA studies.

Study	Data/Domain	Task	Classes	Reported Result
Salur and Aydin [11]	Tourism (TripAdvisor)	Aspect term extraction	—	Different task
Mutlu and Özgür [12]	Twitter, 6 brand targets	Targeted sentiment	3	Macro-F1 0.669
Bilgin and Turan [13]	Restaurant (SemEval-2016 TR)	Aspect extraction + polarity	3	Accuracy 0.8691 (polarity)
This study	Education (school reviews)	Aspect sentiment classification	2	Macro-F1 0.9266/Accuracy 0.9382

These figures are not directly comparable, and we set them side by side only to situate the present work rather than to claim an improvement over it. Each study evaluates on a different corpus, and two differences in particular preclude a direct reading. First, both Mutlu and Özgür and Bilgin and Turan operate under a three-class scheme that includes a neutral category, which is structurally harder than the binary scheme used here; a binary score cannot be compared with a ternary one without qualification. Second, the task formulations differ: Bilgin and Turan evaluate a two-stage pipeline in which polarity is assigned to automatically extracted aspects, whereas the aspect category is given as input here (Section 3.5), and Mutlu and Özgür assign sentiment toward named entities rather than conceptual categories.

A more informative calibration of the present result comes from within the corpus itself. The difficulty of this task is bounded below by the baselines reported in Section 4.1: a predictor using the aspect category alone reaches a Macro-F1 of 0.6147, and TF-IDF with a linear SVM reaches 0.8358. BERTurk exceeds these by 0.3119 and 0.0907 points respectively, both measured on the same data under the same protocol. This is the sense in which the result is competitive—not that it exceeds scores obtained on other corpora, but that it exceeds strong non-pre-trained references on the corpus it was evaluated on.

It should also be stated plainly that this study does not propose a new modeling method. Its contributions are a resource and a controlled comparison, and the transformer architectures evaluated here are used as published. Claims of state-of-the-art performance would accordingly not be meaningful; what the results establish is which pre-training regime is preferable for this task and domain, measured under conditions that hold everything else constant.

The findings largely reinforce earlier work on Turkish ABSA. Mutlu and Özgür [12] reported that BERTurk-based models carry a clear edge over multilingual counterparts on targeted sentiment analysis; this study reproduces that advantage quantitatively at the aspect level and in a different domain (educational reviews), and extends it by demonstrating the advantage systematically against both mBERT and XLM-R rather than a single multilingual baseline. Salur and Aydin [11] emphasized, in the tourism domain, how difficult it is to obtain competitive Turkish ABSA results without high-quality labeled data; the substantial inter-annotator agreement obtained here ($\kappa = 0.702$) indicates that the annotation protocol used in this study meets a comparable reliability threshold.

In the broader ABSA literature, Wang et al. [19] and Sun et al. [25] introduced dedicated attention mechanisms and auxiliary sentence formulations specifically to expose the aspect-sentence relationship to the model. That the plain `[CLS] aspect [SEP] sentence [SEP]` format used here achieves strong results without any such architectural intervention suggests that the self-attention layers of a thoroughly pre-trained transformer already learn

much of that relationship implicitly—at least at this dataset scale and for a binary polarity task. This is consistent with a wider trend in NLP in which architectural complexity is increasingly substituted by representation quality inherited from pre-training.

The class asymmetry identified in the error analysis (Section 4.4)—every model misclassifies negative instances at roughly two to three times the rate of positive ones—is likewise consistent with prior observations. Work on educational text sentiment [40,41] has repeatedly documented that criticism in this domain tends to be indirect or softened. That BERTurk narrows but does not eliminate this gap suggests that implicit, context-dependent negativity is bounded not only by pre-training scale but also by the size of the fine-tuning corpus itself.

5.3. Practical and Methodological Implications

One direct engineering implication concerns the trade-off between model choice and deployment cost. BERTurk, at roughly 110M parameters, can be fine-tuned in reasonable time on a single mid-range GPU, and its Macro-F1 advantage over XLM-R, which has roughly 2.5 times as many parameters, is only +0.006. That marginal gap makes a strong case for BERTurk in any deployment where the target content is exclusively Turkish and multilingual coverage is not a requirement. Where multilingual support is genuinely needed—for instance, a single model serving an international school network—XLM-R obtained higher means than mBERT on every metric and is on that basis the better-supported fallback of the two, although the difference between them was not established at fold level (Section 4.8).

The practical value of aspect-level labeling is visible directly in the dataset statistics: 54% of reviews carry more than one aspect annotation (Section 3.4, Figure 8). A document-level sentiment system would inevitably average out mixed signals in a majority of these cases, misclassifying a review such as "the teachers are caring but the building is in terrible shape" as neutral or one-sided. The aspect-aware approach demonstrated here offers the granularity needed for school-administration dashboards or parent-feedback systems to pinpoint which specific service dimension requires attention, rather than surfacing an undifferentiated sentiment score.

Methodologically, the two-stage dictionary-plus-keyword aspect mapping pipeline described in Section 3.2 resolved 98.2% of 464 raw expressions automatically, substantially reducing the manual annotation burden. That pipeline is directly reusable by researchers collecting data from structurally similar review platforms—university evaluation sites or course-feedback portals, for example—without requiring a redesign of the aspect taxonomy from scratch.

5.4. Limitations

Several limitations bound the interpretation of these results. First, the corpus is small in absolute terms. At 2620 annotated pairs it exceeds the SemEval-2016 Turkish subset that has served as the reference benchmark for Turkish ABSA (Section 2.7, Table 1), and the evaluation protocol mitigates the consequences in the ways set out in Section 4.5—every instance is held out exactly once and fold-level variance is narrow. Neither observation removes the constraint. Its clearest effect is on the long tail of the aspect taxonomy: categories such as Technology (n = 12) and Discipline (n = 26) are too sparse for per-category performance to be estimated reliably, and conclusions in this paper are therefore drawn at the level of the corpus as a whole rather than per category. A larger corpus would also be required before the absolute scores reported here could be treated as stable estimates of what these architectures achieve on Turkish educational text in general.

Second, the dataset originates from a single platform (okul.com.tr); its voluntary-participation model, where satisfied users are more likely to post, produces the positive skew observed in the corpus (69.6% positive). Data collected from a different platform or a mandatory feedback channel—an official school evaluation survey, for instance—could exhibit a markedly different class distribution and register.

Third, the inter-annotator agreement study covered only 200 of the 2620 instances (7.6% of the corpus), using two annotators. While $\kappa = 0.702$ provides a solid reliability signal for that subset, extending this conclusion to the remainder of the corpus rests on an extrapolation; full-corpus multi-annotator labeling was not performed.

Fourth, the study is constrained to a binary (positive/negative) sentiment scheme. Genuinely neutral or mixed statements—expressions along the lines of "not bad, but could be better"—are necessarily forced into one polarity or the other under this scheme, which flattens some of the nuance present in naturally occurring reviews.

Fifth, the comparison is restricted to systems trained or fine-tuned on this corpus: four non-pre-trained baselines, two recurrent models trained from scratch, and three fine-tuned transformers. Few-shot or zero-shot prompting with large language models constitutes an increasingly relevant comparison point that falls outside the scope of this study.

Sixth, the models were not evaluated on an independent dataset. No aspect-level Turkish corpus exists for the education domain, which is the gap this study addresses, and transfer to the available Turkish ABSA resources is not informative: the SemEval-2016 Turkish subset is annotated with twelve entity#attribute categories describing restaurants, so a model trained on the present taxonomy would be evaluated on a label space it has never seen, and the resulting scores would measure taxonomy mismatch rather than generalization. What can be established within the present corpus is reported in Section 4.5: performance is stable across folds but not uniform across input types, declining on long reviews (0.8423 Macro-F1 on sixteen words or more, against 0.9511 on short ones) and on reviews carrying several aspects. Evaluation on an independently collected corpus, ideally from a different platform and educational tier, remains necessary before the present results can be taken to characterize Turkish educational text in general.

Seventh, polarity is distributed very unevenly across aspect categories. The corpus-level ratio of 2.3:1 conceals a range from 9% negative for Course Variety to 98% negative for Student Development, and only eight of the 19 categories fall within the 30–70% band (Section 3.4). Because the aspect category is supplied to the classifier as part of the input, a model can exploit this association directly: the aspect-only baseline of Section 4.1, which never sees the sentence, already reaches a Macro-F1 of 0.6147. The models evaluated here exceed that floor by a wide margin—BERTurk by 0.3119 points—so they are clearly not relying on the category label alone, but the aggregate figures nonetheless rest in part on category-specific priors rather than on reading the review text alone. This also explains why performance is higher on sparse categories (0.9783) than on frequent ones (0.8918), as noted in Section 4.5.

Eighth, the hyperparameter configurations were selected on a split that overlaps the cross-validation corpus (Section 3.6), and Section 4.9 shows that repeating the selection inside each fold changes the recurrent scores by 0.017 to 0.021 Macro-F1 points. The equivalent check could not be run for the three transformers within the available compute budget, so a selection bias of comparable magnitude cannot be ruled out for them. Its consequences would differ by comparison: an effect of that size is well below the 0.0675 points separating the weakest transformer from the strongest classical baseline, so the grouping reported in Table 7 is robust to it, but it exceeds the 0.0059 points separating BERTurk from XLM-R, so the ordering within the transformer group should be read as provisional—the reading Section 4.8 already gives it on independent grounds.

Ninth, reviews containing insults or profanity were excluded at collection (Section 3.1). Because strongly critical reviews are the ones most likely to contain such language, this screen may have removed negative reviews at a higher rate than positive ones, which would add to the positive skew already produced by the platform's voluntary-participation model. The number of reviews removed at this stage was not recorded, so the size of this effect cannot be estimated here; it is noted as a possible source of selection bias rather than quantified.

Finally, this work addresses a classification task in which the aspect category is provided as input. A fully deployed system would additionally require automatic aspect term extraction from raw text, evaluated jointly with the sentiment classifiers presented here as part of an end-to-end pipeline—an evaluation not undertaken in the current study.

5.5. Future Work

The limitations above point toward several natural extensions. Broadening data collection across multiple platforms and educational tiers—universities, tutoring centers, and similar settings—could reduce both class imbalance and platform-specific bias. Incorporating a neutral or mixed sentiment class would require re-evaluation under a ternary classification setup but would yield a system more faithful to real-world usage. Combining aspect term extraction with sentiment classification in a joint architecture would move the system toward genuine end-to-end applicability.

From an engineering standpoint, evaluating distilled or quantized variants of BERTurk under latency-sensitive deployment scenarios—a real-time school feedback dashboard, for example—is a natural next step that would translate the accuracy-focused findings reported here into resource-constrained settings. Finally, benchmarking zero-shot and few-shot large language model performance against fine-tuned BERTurk under the same 5-fold protocol would provide a timely comparison point given how rapidly this area of the field is moving.

Two further comparisons bear directly on the interpretation offered here and are left open. The first would separate capacity from pre-training: the recurrent models hold approximately 250K and 2.1M parameters against BERTurk's 110M, so the present design does not isolate the two. The ordering within the pre-trained group argues against capacity as the explanation—mBERT (178M) and XLM-R (278M) both score below BERTurk (110M), as Table 11 shows—but a recurrent model scaled to comparable size would test this directly, with the caveat that such a model would be heavily overparameterized relative to a corpus of 2620 instances. The second concerns imbalance handling: class-weighted cross-entropy is applied throughout and no alternative was evaluated (Section 3.6), so whether the ranking of the architectures is stable under focal loss or under resampling is not established here.

6. Conclusions

This study addressed three concrete gaps in aspect-based sentiment analysis (ABSA) for Turkish educational text: the absence of a domain-specific labeled dataset, the complete lack of ABSA work targeting educational reviews, and the absence of a controlled, quantitative test of language-specific versus multilingual pre-training for Turkish aspect-based sentiment analysis. A new dataset of 2620 labeled (sentence, aspect) pairs across 19 canonical aspect categories, derived from 1533 reviews collected from okul.com.tr, was introduced, with label reliability confirmed through an inter-annotator agreement of $\kappa = 0.702$ (substantial agreement).

Comparing nine systems under a review-level 5-fold cross-validation protocol produced a clear ordering. Among the systems without pre-training, TF-IDF with a linear SVM (Macro-F1 = 0.8358) exceeded both recurrent models (BiGRU 0.7999, BiLSTM 0.7804); all of

them trailed well behind the pre-trained transformers. BERTurk, pre-trained on Turkish text, delivered the strongest results at 0.9266 Macro-F1 and 0.9382 accuracy, exceeding the strongest non-pre-trained baseline by +0.091 points and the multilingual alternatives mBERT (0.9033) by +0.023 and XLM-R (0.9207) by +0.006, the latter margin falling within fold-level variance. The gap widened on the minority class: BERTurk's F1-Negative (0.8975) exceeded mBERT's by +0.034 points, indicating that the benefit of language-specific pre-training is not evenly distributed across classes but concentrates where the task is hardest.

BERTurk achieved the highest mean performance in this dataset and experimental setting. Its advantage over mBERT is established by the review-level analysis, while its advantage over XLM-R was small and was not statistically established. These results therefore indicate that language-specific pre-training remains worth considering for a morphologically rich, agglutinative language such as Turkish, without establishing that it is superior to multilingual pre-training at the scale XLM-R represents. They also show that a comparatively modest model (~110M parameters), fine-tuned on a domain-specific corpus, can achieve both high accuracy and a computational footprint compatible with practical deployment—a combination particularly relevant for resource-constrained research settings and real-time feedback systems.

The practical contribution of this work lies in laying the groundwork for an aspect-aware decision-support component for school-administration and parent-communication workflows; with more than half of the reviews touching multiple aspects, the results confirm that document-level, single-label approaches would be inadequate for this domain. The limitations discussed in Section 5.4—single-platform sourcing, the binary sentiment scheme, and the assumption of a known aspect category as input—together with the future directions outlined in Section 5.5 outline a concrete roadmap for subsequent work. Taken together, this study contributes both a new resource and a reproducible evaluation protocol that can inform future ABSA research on Turkish and typologically similar languages.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/electronics15214949/s1. Figure S1: Macro-F1 score per fold for BiLSTM, BiGRU and BERTurk; Figure S2: Macro-F1 per fold for BERTurk, mBERT and XLM-R; Figure S3: Per-fold F1-Negative comparison for the three transformers. All three report fold-level values that are also given numerically in Tables 8 and 10.

Author Contributions: Conceptualization, H.A. and S.G.; methodology, H.A.; software, H.A.; validation, H.A. and S.G.; formal analysis, H.A.; writing—original draft preparation, H.A.; writing—review and editing, S.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was not funded by any external organization.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: A representative sample of the annotated dataset (stratified by aspect category and polarity) and the complete source code for every system evaluated here are publicly available at https://github.com/haksaya/ABSA-Paper-Repo (accessed on 21 July 2026). The full dataset of 2620 annotated pairs is available from the corresponding author upon reasonable request, subject to the terms of use of the source platform.

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

Abbreviations

The following abbreviations are used in this manuscript:

ABSA	Aspect-Based Sentiment Analysis
NLP	Natural Language Processing
BERT	Bidirectional Encoder Representations from Transformers
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
BiLSTM	Bidirectional Long Short-Term Memory
BiGRU	Bidirectional Gated Recurrent Unit
mBERT	Multilingual BERT
XLM-R	XLM-RoBERTa
LLM	Large Language Model
GCN	Graph Convolutional Network

References

1. Liu, B. *Sentiment Analysis and Opinion Mining*; Morgan & Claypool: San Rafael, CA, USA, 2012. [CrossRef]
2. Pang, B.; Lee, L. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.* **2008**, *2*, 1–135. [CrossRef]
3. Pang, B.; Lee, L.; Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), Philadelphia, PA, USA, 6–7 July 2002; pp. 79–86. [CrossRef]
4. Turney, P.D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting of the ACL, Philadelphia, PA, USA, 6–12 July 2002; pp. 417–424. [CrossRef]
5. 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), Philadelphia, PA, USA, 24–25 August 2004; pp. 168–177. [CrossRef]
6. Pontiki, M.; Galanis, D.; Pavlopoulos, J.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S. SemEval-2014 Task 4: Aspect based sentiment analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval), Dublin, Ireland, 23–24 August 2014; pp. 27–35. [CrossRef]
7. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
8. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
9. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [CrossRef]
10. 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), San Diego, CA, USA, 16–17 June 2016; pp. 19–30. [CrossRef]
11. Salur, M.U.; Aydin, I. An annotated Turkish aspect based sentiment analysis corpus for smart tourism. In Proceedings of the IEEE Innovations in Intelligent Systems and Applications Conference (ASYU), Elazığ, Turkey, 6–8 October 2021. [CrossRef]
12. Mutlu, M.M.; Özgür, A. A dataset and BERT-based models for targeted sentiment analysis on Turkish texts. In Proceedings of the 60th Annual Meeting of the ACL: Student Research Workshop, Dublin, Ireland, 22–27 May 2022; pp. 338–346. [CrossRef]
13. Bilgin, M.; Turan, M. Aspect-based sentiment analysis for Turkish reviews using token and sequential classification methods. *Int. J. Comput. Inf. Technol.* **2025**, *14*, 90–99. [CrossRef]
14. 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. [CrossRef]
15. Pontiki, M.; Galanis, D.; Papageorgiou, H.; Manandhar, S.; Androutsopoulos, I. SemEval-2015 Task 12: Aspect based sentiment analysis. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval), Denver, CO, USA, 4–5 June 2015; pp. 486–495. [CrossRef]
16. Wu, C.; Ma, B.; Liu, Y.; Zhang, Z.; Deng, N.; Li, Y.; Chen, B.; Zhang, Y.; Xue, Y.; Plank, B. M-ABSA: A multilingual dataset for aspect-based sentiment analysis. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Suzhou, China, 4–9 November 2025; pp. 2530–2557. [CrossRef]
17. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems (NeurIPS), Lake Tahoe, NV, USA, 5–10 December 2013; pp. 3111–3119.

18. Tang, D.; Qin, B.; Feng, X.; Liu, T. Effective LSTMs for target-dependent sentiment classification. In Proceedings of the International Conference on Computational Linguistics, Osaka, Japan, 11–16 December 2016; pp. 3298–3307. Available online: https://aclanthology.org/C16-1311 (accessed on 10 May 2026).
19. Wang, Y.; Huang, M.; Zhu, X.; Zhao, L. Attention-based LSTM for aspect-level sentiment classification. In Proceedings the Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 606–615. [CrossRef]
20. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
21. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
22. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 1724–1734. [CrossRef]
23. Zhang, C.; Li, Q.; Song, D. Aspect-based sentiment classification with aspect-specific graph convolutional networks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing, Hong Kong, China, 3–7 November 2019; pp. 4568–4578. Available online: https://aclanthology.org/D19-1464/ (accessed on 15 May 2026).
24. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, 2–7 June 2019; pp. 4171–4186. [CrossRef]
25. Sun, C.; Huang, L.; Qiu, X. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, 2–7 June 2019; pp. 380–385. [CrossRef]
26. Zhang, W.; Li, X.; Deng, Y.; Bing, L.; Lam, W. Towards generative aspect-based sentiment analysis. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Virtual, 1–6 August 2021; Volume 2, pp. 504–510. Available online: https://aclanthology.org/2021.acl-short.64/ (accessed on 7 June 2026).
27. Simmering, P.F.; Huoviala, P. Large language models for aspect-based sentiment analysis. *arXiv* **2023**, arXiv:2310.18025. [CrossRef]
28. Ding, X.; Zhou, J.; Dou, L.; Chen, Q.; Wu, Y.; Chen, A.; He, L. Boosting large language models with continual learning for aspect-based sentiment analysis. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP, Miami, FL, USA, 12–16 November 2024; pp. 4367–4377. [CrossRef]
29. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv* **2019**, arXiv:1907.11692. [CrossRef]
30. Martin, L.; Muller, B.; Ortiz Suárez, P.J.; Dupont, Y.; Romary, L.; de la Clergerie, É.; Seddah, D.; Sagot, B. CamemBERT: A tasty French language model. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Virtual, 5–10 July 2020; pp. 7203–7219. [CrossRef]
31. Antoun, W.; Baly, F.; Hajj, H. AraBERT: Transformer-based model for Arabic language understanding. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT), Marseille, France, 12 May 2020; pp. 9–15. Available online: https://aclanthology.org/2020.osact-1.2/ (accessed on 15 June 2026).
32. Rust, P.; Pfeiffer, J.; Vulić, I.; Ruder, S.; Gurevych, I. How good is your tokenizer? On the monolingual performance of multilingual language models. In Proceedings of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing, Virtual, 1–6 August 2021; pp. 3118–3135. [CrossRef]
33. Dehkharghani, R.; Saygin, Y.; Yanikoglu, B.; Oflazer, K. SentiTurkNet: A Turkish polarity lexicon for sentiment analysis. *Lang. Resour. Eval.* **2016**, *50*, 667–685. [CrossRef]
34. Demirtas, E.; Pechenizkiy, M. Cross-lingual polarity detection with machine translation. In Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM), Chicago, IL, USA, 11 August 2013; pp. 1–8. [CrossRef]
35. Kaya, M.; Fidan, G.; Toroslu, I.H. Sentiment analysis of Turkish political news. In Proceedings of the 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Macau, China, 4–7 December 2012; pp. 174–180. [CrossRef]
36. Schweter, S. BERTurk—BERT Models for Turkish, Version 1.0.0. Zenodo 2020. Available online: https://zenodo.org/records/3770924 (accessed on 18 October 2026).
37. Çakıcı, Ş.; Karaduman, D.; Çırlan, M.A.; Hürriyetoğlu, A. A cross-validation study of Turkish sentiment analysis datasets and tools. *Lang. Resour. Eval.* **2025**, *59*, 4003–4041. [CrossRef]
38. Zümberoğlu, K.B.; Dik, S.Z.; Karadeniz, B.S.; Sahmoud, S. Towards better sentiment analysis in the Turkish language: Dataset improvements and model innovations. *Appl. Sci.* **2025**, *15*, 2062. [CrossRef]

39. Yaman, S.; Aslan, Ö. Multi-aspect sentiment analysis on Turkish food delivery reviews: A multi-task BERTurk approach. *Int. J. Innov. Eng. Appl.* **2026**, *10*, 121–130. [CrossRef]
40. Altrabsheh, N.; Gaber, M.M.; Cocea, M. SA-E: Sentiment analysis for education. *Front. Artif. Intell. Appl.* **2013**, *255*, 353–362. [CrossRef]
41. Wen, M.; Yang, D.; Rosé, C.P. Sentiment analysis in MOOC discussion forums: What does it tell us? In Proceedings of Educational Data Mining (EDM), London, UK, 4–7 July 2014; pp. 130–137. Available online: https://educationaldatamining.org/EDM2014/uploads/procs2014/long%20papers/130_EDM-2014-Full.pdf (accessed on 20 April 2026).
42. Sindhu, I.; Daudpota, S.M.; Badar, K.; Bakhtyar, M.; Baber, J.; Nurunnabi, M. Aspect-based opinion mining on student's feedback for faculty teaching performance evaluation. *IEEE Access* **2019**, *7*, 108729–108741. [CrossRef]
43. Kastrati, Z.; Imran, A.S.; Kurti, A. Weakly supervised framework for aspect-based sentiment analysis on students' reviews of MOOCs. *IEEE Access* **2020**, *8*, 106799–106810. [CrossRef]
44. Ren, P.; Yang, L.; Luo, F. Automatic scoring of student feedback for teaching evaluation based on aspect-level sentiment analysis. *Educ. Inf. Technol.* **2023**, *28*, 797–814. [CrossRef]
45. Kathuria, A.; Gupta, A.; Singla, R.K. AOH-Senti: Aspect-oriented hybrid approach to sentiment analysis of students' feedback. *SN Comput. Sci.* **2023**, *4*, 152. [CrossRef]
46. Butt, S.; Núñez Daruich, S.D.; Alvarado-Uribe, J.; Ceballos, H.G. Cutting-edge technologies for analyzing student feedback to inform institutional decision-making in higher education. *Foresight STI Gov.* **2025**, *19*, 68–80. [CrossRef]
47. 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, Virtual, 9 June 2020; pp. 8440–8451. [CrossRef]
48. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. In Proceedings of the 7th International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.