Aspect-Based Sentiment Analysis for Turkish Reviews Using Token and Sequential Classification Methods

Metin Bilgin
[1]Department of Computer Engineering
Bursa Uludağ University of Turkey
Bursa,Turkey
Email: metinbilgin [AT] uludag.edu.tr

Melek Turan
[1]Department of Computer Engineering
Bursa Uludağ University
[2]Özveri Ar-ge Merkezi
Bursa, Turkey
Email: Melekturan454 [AT] gmail.com

Abstract— **Aspect-Based Sentiment Analysis (ABSA) aims to identify sentiments expressed toward specific aspects or attributes of entities in text. This study addresses the under-explored area of ABSA in the Turkish language by extracting aspect terms (targets) and their categories from customer reviews and determining the sentiment polarity for each aspect. Turkish, being a morphologically rich and structurally complex language, poses unique challenges that often hinder the direct application of methods developed for other languages. Hence, developing sentiment analysis approaches tailored to Turkish is of significant importance. We propose a two-stage pipeline: a token-level classification to recognize aspect terms and assign them to one of 12 predefined aspect categories, followed by a sequence-level (sentence-level) classification to predict sentiment (positive, negative, or neutral) for each identified aspect. We fine-tuned five transformer-based language models (BERT, ConvBERT, ELECTRA, DeBERTa, and DistilBERT) for aspect term and category extraction, and four models (BERT, ConvBERT, ELECTRA, DistilBERT) for sentiment classification. Experimental results on the SemEval-2016 Turkish ABSA Restaurant dataset show that the BERT model achieved the highest accuracy (92.20%) for aspect term and category identification, closely followed by ConvBERT (91.68%). For sentiment analysis, ConvBERT performed best with an accuracy of 86.91%, outperforming ELECTRA (85.34%), BERT (82.75%), and DistilBERT (77.48%). These findings demonstrate that pre-trained transformer models can effectively handle fine-grained sentiment analysis in Turkish, substantially improving on previous approaches. The proposed pipeline and comparative results provide a novel benchmark for Turkish ABSA, with potential applications in analyzing Turkish customer feedback to glean actionable insights.**

Keywords- token classification; sequential model classification; aspect term; aspect-based sentiment analysis; deep learning, turkish

I. INTRODUCTION

Aspect-based sentiment analysis (ABSA) is a research domain that examines and retrieves emotional expressions and related elements in written texts. Its primary objective is to analyze content in order to recognize and assess the emotional aspects associated with the subjects or entities mentioned. ABSA employs a predefined set of classifications to detect emotional expressions related to any subject or entity. Typically, these classifications are derived from textual sources such as customer opinions, social media blog posts, or product reviews. By pinpointing the emotional aspects linked to different elements of a product or service, ABSA provides valuable insights that companies can use to enhance their products, refine their marketing tactics, or improve service quality.

While ABSA has been extensively studied in English and other major languages, research on this task in Turkish remains limited. Turkish, with its agglutinative nature and complex morphology, presents unique challenges that require models to capture subtle linguistic nuances. Early studies on Turkish sentiment analysis primarily relied on manual feature engineering or frequency-based methods [1], [2] and achieved moderate success. These methods, however, often struggled to generalize across diverse linguistic structures and failed to capture the fine-grained distinctions necessary for robust ABSA.

To bridge this gap, our study employs state-of-the-art pre-trained transformer models to perform ABSA on Turkish texts. The novelty of our work lies in its two-stage pipeline: first, we use token-level classification for target term extraction and category assignment; second, we apply sequential classification to determine the sentiment (positive, negative, or neutral) associated with each target. For the sequential classification task, we compared four models BERT [3], DistilBERT [4], ConvBERT [5], and ELECTRA [6] and for target term extraction, we evaluated five models BERT, ELECTRA, DistilBERT, ConvBERT, and DeBERTa [7]. Our approach not only addresses the challenges posed by Turkish morphology but also establishes a new benchmark for Turkish ABSA by clearly demonstrating the superior performance of transformer-based methods in this context.

The remainder of this paper is organized as follows. Section 2 summarizes previous studies, Section 3 details the

dataset, Section 4 describes the proposed methodology, Section 5 presents performance results, Section 6 discusses key findings and limitations, and finally, Section 7 concludes the work and highlights its significance.

II. RELATED WORK

Research on aspect-based sentiment analysis has evolved considerably over the past two decades. Early foundational work by Wilson et al. [8] is widely regarded as a pioneering study in ABSA. Wilson and colleagues developed linguistic rule-based techniques to identify opinion-bearing phrases and determine their polarity in context [8]. Following this, the rapid development of machine learning approaches led to a variety of models and algorithms for ABSA. The introduction of neural networks brought significant improvements: for example, Recurrent Neural Networks (RNN) and in particular Long Short-Term Memory (LSTM) networks allowed modeling of sequence data for sentiment tasks. Cho et al. [9] proposed the Gated Recurrent Unit (GRU) to improve upon standard RNNs' memory mechanisms, enabling more effective handling of long-term dependencies in text. With these advancements, researchers began achieving better accuracy in detecting aspect-specific sentiments. To address limitations of earlier neural models, innovative architectures were explored. In 2018, Wang et al. [10] introduced a capsule network approach for sentiment analysis, aiming to capture hierarchical feature relationships. Their model represented aspects and sentiments as vectors ("capsules") and achieved notable performance gains by reconstructing input representations and estimating sentiment presence concurrently [10]. In 2019, Hou et al. [11] presented the SA-GCN (Selective Attention Based Graph Convolutional Networks) model based on a capsule network and graph convolutional networks, which achieved an accuracy score of 85.8% on the ABSA SemEval restaurant dataset [12], [11]. Since 2019, there has been a notable utilization of the BERT model and the attention mechanism for the purpose of categorizing textual attributes and discerning emotional connotations. Zhang et al. [13] introduced the MIN (Multiple Interactive Attention Network) model as a BERT-based model. The method obtains a parallel hidden state using a partial transformer after completing the pre-training process. In the ABSA task with the SemEval restaurant dataset, the MIN model achieved an accuracy score of 82.69%. In 2021, the ABSA-DeBERTa model was proposed by Silva et al. [14], who integrated the BERT model with the disentangled attention mechanism in order to perform target phrase classification and identify the corresponding emotional meanings for the ABSA task. Following the completion of the investigation, an evaluation was conducted on the restaurant dataset, wherein the ABSA-DeBERTa model exhibited a remarkable accuracy score of 89.46%. Yang et al. [15] introduced the LSA (local sentiment aggregation paradigm), in conjunction with the LSA+DeBERTa-V3-Large model, to enable comprehensive modeling of sentiment coherence. They used the LSA method to better comprehend the relationships between sentences, paragraphs, and words expressing emotion in text and learn emotional meaning more consistently. The method performed exceptionally well on the restaurant dataset, achieving an accuracy score of 90.33%. Most recently, Scaria et al. [16] proposed InstructABSA, which leverages natural language instructions to guide the model in identifying aspect sentiments. This innovative approach attained 89.06% accuracy on the SemEval restaurant dataset and 88.37% on a laptop reviews dataset, representing the state-of-the-art in general ABSA tasks.

Research specifically focusing on Turkish texts began to gain momentum after 2016. Kama et al. [1] proposed a three-step approach for extracting features from informal text documents. In this approach, they employed frequency-based feature extraction (FBFE), frequency-based feature extraction with sentiment word support (FBFESWS) and web search-based feature extraction (WSBFE). The study was conducted on a dataset of user reviews for a mobile phone model, and an F-Score of 69.79% along with a precision of 59.24% were obtained for WSBFE [1]. Türkmen et al. [17] proposed an ABSA-based method by analysing hotel reviews. They developed a scoring algorithm using reviews collected from the web and determined the sentiment scores of the directions. Kama et al. [2] then carried out a new study to match features and sentiment words of informal Turkish texts, aiming to improve the performance of ABSA systems. As a result, F-Score and precision values of 85% and 91%, respectively, were achieved. Ekinci et al. [18] presented a system for automatically discovering binary features using Turkish reviews of products from various domains. The study involved organising the dataset, determining n-gram frequencies, filtering these frequencies, and creating multiple word aspects. Consequently, precision and accuracy scores of 82% and 83% were obtained. Çetin and Eryiğit [19] worked on the extraction of the target category and target term. They devised a labelling algorithm based on word vectors and lexical analysis data for this purpose. The study, which focused on the SemEval 2016 ABSA restaurant dataset, yielded an F-Score of 46.7% when simultaneously identifying the target category and term. Kama et al. [20] further developed a tool for ABSA. By using online product reviews as a dataset, they extracted implicit and explicit aspects with a frequency-based method. This study produced an F-Score of 86.75% and a precision of 91.22%. Özkan [21] conducted an ABSA study specifically for Turkish. In this study, Turkish reviews about smartphones from the web were utilised and goal-based sentiment analysis was performed on performance, price, and camera features. The results indicated that the highest precision, accuracy, and F-Score values were 93%, 94%, and 93%, respectively. Salur et al. [22] examined the SemEval 2016 ABSA Turkish restaurant data. Appearance extraction was carried out from lemma word, intermediate word, and raw word formats using different methods, and the results were evaluated by combining various strategies. It was reported that TF-IDF was the most successful method, achieving an F-Score of 60.07% for raw words. Girgin et al.

[23] provided a comprehensive review of sentiment analysis studies for Turkish.

In summary, past studies on Turkish ABSA have explored a range of techniques from frequency-based methods to custom algorithms. However, there has been a clear gap in leveraging the latest deep learning models for Turkish aspect-based sentiment analysis. Our work builds on this literature by introducing transformer-based models into the Turkish ABSA domain, aiming to substantially improve performance and contribute a new perspective to Turkish sentiment analysis research.

III. DATASETS

In the study, SemEval-2016 ABSA Restaurant Reviews-Turkish dataset was used for training and testing [12].

It consists of Turkish user reviews of restaurants, with each sentence annotated for aspect terms, aspect categories, and sentiment polarity towards each aspect. An example annotation from the dataset is illustrated in Figure 1 of the original paper, showing a sentence with tagged aspect term, its category, and the sentiment label.

There are 12 different 1 categories used in the category classification. The number of sentences containing these categories is shown in Table 1. For the sentiment analysis task in the training data, 820 of the 1535 sentences contain positive, 586 negative, and 129 neutral sentiments. This distribution shows that neutral sentiment is comparatively rare (only ~8% of training instances), which could pose a challenge for learning algorithms as they may be biased toward the more frequent positive/negative classes. We will see in our results that this class imbalance indeed affects the models' performance on the neutral class.

All text in the dataset is in Turkish. We performed minimal preprocessing, relying on the inherent text handling of the pre-trained models (which can handle Turkish input). We used the official train-test split without further subdivision; a small portion of the training set was set aside for validation during model fine-tuning if needed (for hyperparameter tuning and early stopping), though given the small dataset size, we mostly relied on the provided split for evaluation. No additional external data or augmentation was used. By using a standard dataset, we ensure our results are comparable to prior work and provide a clear benchmark for future Turkish ABSA studies.

TABLE I. DISTRIBUTION OF THE DATASET BY CATEGORY

Category	Train Data	Test Data
FOOD#QUALITY	446	77
AMBIENCE#GENERAL	277	31
SERVICE#GENERAL	235	25
RESTAURANT#GENERAL	228	22
FOOD#STYLE_OPTIONS	123	10
RESTAURANT#PRICES	65	9
DRINKS#QUALITY	51	6
LOCATION#GENERAL	39	4
DRINKS#STYLE_OPTIONS	30	1
FOOD#PRICES	31	4
DRINKS#PRICES	9	1
RESTAURANT#MISCELLANEOUS	1	1

```
<text>Manzara sahane evet ama servis rezalet.</text>
<Opinions>
   <Opinion target="servis" category="SERVICE#GENERAL"
   polarity="negative" from="24" to="30" />
   <Opinion target="Manzara" category="AMBIENCE#GENERAL"
   polarity="positive" from="0" to="7" />
```

Figure 1. Example of the dataset.

IV. METHODOLOGY

Our approach to aspect-based sentiment analysis in Turkish consists of two main stages: Aspect Term and Category Extraction using a token classification model, and Sentiment Classification for each extracted aspect using a sequential (sentence-level) classification model. We adopted this two-stage pipeline to break down the complex ABSA task into two more tractable subtasks, allowing the model to focus on one problem at a time. This is particularly beneficial given the morphological richness of Turkish: identifying aspect terms and their categories is a different challenge from determining sentiment, and treating them separately can lead to better performance than a single joint model.

A. Target Term Identification and Category Classification with Token Classification Model

In the first stage, we identify the aspect terms (also called target terms) present in a sentence and determine the category of each aspect term. We formulate this as a token classification problem, which is akin to a named-entity recognition task. Each token (word or sub-word unit) in the input sentence is assigned a label indicating whether it is part of an aspect term and, if so, which category it belongs to. For example, in the sentence "Servis çok yavaştı, ama yemekler lezzetliydi" ("The service was very slow, but the meals were delicious"), the token "Servis" would be labeled as part of a SERVICE#GENERAL aspect, "yavaştı" (slow) would likely be non-aspect (or considered part of the aspect phrase if we treat the whole phrase as aspect term), and "yemekler" (meals) would be labeled as FOOD#QUALITY. By labeling at the token level, the model can pinpoint exact spans of text that correspond to aspect mentions and simultaneously classify their category.

Token classification involves categorizing each token in a text sequence into predefined classes based on context. It is a fundamental technique in NLP with applications in tasks such as part-of-speech tagging, named entity recognition, and aspect extraction. Effective token classification requires capturing the context around a token to decide its label.

Token extraction is a crucial step in NLP preprocessing, performed before applying clustering, classification, or other

machine learning tasks. It includes tasks such as word segmentation, tokenization, word stopping, word stemming, and term frequency weighting [24]. It has various applications in NLP models, including phrase processing [25], hate speech detection [26], and identifying spurious correlations [27].

The present study uses token classification to identify target terms and categories. Prior to classification, we perform tokenization using the appropriate tokenizer for the pre-trained model (which, in the case of BERT and similar models, is typically a WordPiece or SentencePiece tokenizer that can handle Turkish morphology by breaking words into sub-word units as needed). After tokenization, our pipeline for aspect extraction follows these steps:

1) Tokenize the input sentence using the model's tokenizer. This splits the sentence into tokens (for example, "yemekler" might be split into yemekler or yemek@@ + ler, depending on the tokenizer's vocabulary).

2) Apply a fine-tuned token classification model to the tokenized sequence. We fine-tuned each transformer model on the training data with tokens labeled according to their aspect category or "O" (outside) if they are not part of any aspect term.

3) Identify aspect terms and categories from model output. The model outputs a label for each token indicating its category (or O for non-aspect). From these labels, we reconstruct aspect terms (contiguous tokens with the same category label form an aspect term). For instance, if tokens "çok" "lezzetli" are both labeled as FOOD#QUALITY, together they form one aspect term phrase describing the food quality.

4) Output the list of aspect terms with their categories for the sentence. If a sentence has no aspect terms, the model would label all tokens as non-aspect (O).

This token-level approach inherently handles sentences with multiple aspects: each token is considered independently with context, so multiple distinct aspect terms can be recognized in one pass. It also links each term to a category immediately. We opted to include the category in the token labeling (as opposed to first identifying aspect spans then classifying them separately) to allow the model to use context to decide the appropriate category, which is useful in Turkish where some aspect terms might be ambiguous without context. Token classification is facilitated by the rich contextual embeddings provided by the transformer models, which can capture long-range dependencies in Turkish (e.g., subject-object relationships or adjective-noun agreement that might signal aspect-category pair). We fine-tuned five pre-trained transformer models for this token classification task: BERT, ConvBERT, ELECTRA, DeBERTa, and DistilBERT. All models were used in a multilingual or Turkish-capable version so they could process Turkish text. In training, we used the cross-entropy loss on the token labels, and we evaluated performance using standard sequence labeling metrics: precision, recall, and F1-score for correctly identified aspect terms (with correct category), as well as overall accuracy of token classification. The models were trained for a number of epochs with early stopping based on validation F1 to prevent overfitting, given the small dataset size.

B. Sentiment Analysis with Sequential Classification Model

A sequential classification model is commonly used in NLP to leverage sequential information for predicting or classifying text data. After extracting aspect terms and categories from a sentence, the next step is to determine the sentiment expressed toward each aspect. A sequential classification model processes the entire text sequence, utilizing the order of words to understand contextual meaning. The model observes the complete sentence (or the sentence combined with the aspect term) and outputs a single sentiment label for the target aspect.

In NLP, sequential information refers to the order of words or tokens in a text. Sequential classification models have been successfully applied in various NLP tasks such as sentiment analysis, fake news detection, text classification, information extraction, and question answering [15], [28]–[30]. These models are capable of learning to classify text data based on sequential patterns and contextual dependencies present in the input. Once trained on labeled data, sequential models can generalize and classify previously unseen instances effectively.

Sequential Classification Background: In our pipeline, we feed the model both the sentence and an indication of the target aspect whose sentiment we want to classify. We experimented with a simple approach: appending the aspect term at the beginning of the sentence separated by a special token (e.g., "[ASP] aspect_term [SEP] sentence") so that the model is aware of which aspect to focus on. Alternatively, the aspect term could be replaced with a special placeholder or marked with tags; the key objective is to inform the model explicitly about the target. In training, each training instance for the sentiment model consists of a sentence, an aspect (either explicitly given or marked in the sentence), and a sentiment label. We generate these from the original training annotations: if a sentence has multiple aspects, it contributes multiple training instances (sentence+aspect1 → sentiment1, sentence+aspect2 → sentiment2, etc.). We fine-tuned the models to predict the sentiment label given this input.

Notably, the sentiment classification model is trained after the aspect extraction model, but during testing it relies on the aspect extractor's output. Any missed or incorrectly identified aspect in the first stage would mean the second stage might not get the correct input. This is an inherent challenge in pipeline approaches (error propagation). However, separating the tasks allows each model to specialize, and we can also evaluate sentiment classification independently using the true aspect annotations to understand its upper-bound performance.

This study uses sequential classification models to perform sentiment analysis on sentences.We compared a total of 4 different pre-trained models (BERT, DistilBERT, ConvBERT, ELECTRA) for sentiment analysis.We compare the

performance of the four pre-trained models on sentiment classification by overall accuracy as well as precision, recall, and F1 for each sentiment class. The models were fine-tuned using cross-entropy loss on the sentiment labels. Because the dataset is imbalanced (with fewer neutral examples), we monitored class-specific performance to ensure the model does not completely ignore the neutral class. Techniques like class weighting or oversampling neutrals were considered, but in practice the transformer models still managed to learn to some extent the neutral class distinctions.

C. Method

The study revealed two distinct classification techniques: token classification and sequential classification. The target terms in the sentence and the categories to which they correspond are identified using the token classification approach. Figure 2 depicts the steps used for token classification. The steps used to determine the categories are as follows:

1) With the aid of a tokenizer appropriate for the model being used, the sentence that was provided as input is broken up into tokens.

2) The pre-trained models receive the tokenized sentence.

3) The model identifies the target terms.

4) The model indicates the category to which the target terms belong. Which tokens are target terms and to which category they belong are determined as outputs.

Another method used in the study is sequential classification. The sequential classification method was used to identify different sentiments in the sentence and to determine to which category these sentiments belong. The steps followed for sequential classification are shown in Figure 3. Sentiment analysis is performed by applying the following steps respectively.

1) The target terms and sentences identified by the token classification are given as input to the sequential classification model.

2) The sequential classification models trained with the target terms, sentiment, and sentences determine the sentiment of the target term in the new sentence.

3) As a result of the prediction, the sentiment corresponding to the target term is determined.

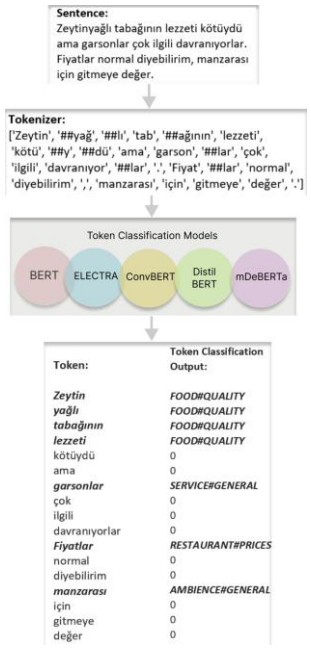

Figure 2. Category determination method with token classification.

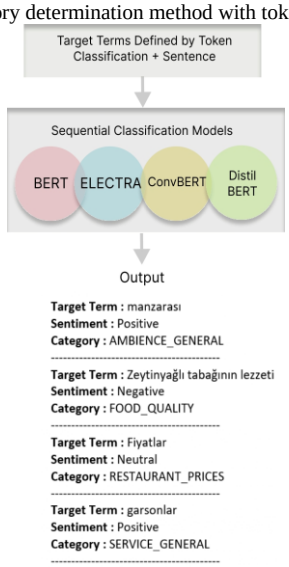

Figure 3. Sentiment analysis method with sequential classification.

V. RESULT AND DISCUSSION

This section presents the experimental results for both stages of the pipeline: aspect term & category extraction (token classification) and aspect-level sentiment classification (sequential classification). We report each model's performance on the test set and provide an analysis to interpret the results. All results are evaluated against the gold-standard annotations of the SemEval-2016 Turkish ABSA dataset described earlier.

A. Target Term Identification and Category Classification with Token Classification Model

The token classification models are compared based on various metrics, as shown in Table 2. The BERT model exhibits superior performance, with high precision (99.41%), recall (99.32%), and F-Score (99.37%). Additionally, it achieves a notably high training accuracy (99.89%), although its test accuracy (92.20%) is slightly lower than the other metrics, suggesting potential overfitting to the training data.

The ConvBERT model also performs well, with precision (94.35%), recall (94.51%), and F-Score (94.43%) all at competitive levels. Its training accuracy (98.87%) is slightly higher than its test accuracy (91.68%), indicating a well-balanced overall performance.

The ELECTRA model demonstrates high precision (94.40%) and F-Score (94.16%), though its recall (93.92%) and test accuracy (90.95%) are somewhat lower in comparison to other models. With a training accuracy of 98.71%, the model appears to be well-adapted to the test data.

The DeBERTa model shows strong results in terms of precision (92.22%) and F-Score (92.86%), while its recall (93.46%) and test accuracy (90.77%) are slightly lower than those of other models. The model's training accuracy (98.70%) suggests an overall balanced performance.

Lastly, the DistilBERT model demonstrates comparatively lower performance, with precision (90.59%), recall (90.78%), and F-Score (90.69%). Its training accuracy (98.01%) and test accuracy (89.95%) are among the lowest values observed, indicating a tendency to overfit the training data.

TABLE II. RESULTS OF TOKEN CLASSIFICATION (%)

Model	Precision	Recall	F-Score	Train Accuracy	Test Accuracy
BERT	99.41	99.32	99.37	99.89	92.20
ConvBERT	94.35	94.51	94.43	98.87	91.68
ELECTRA	94.40	93.92	94.16	98.71	90.95
DeBERTa	92.22	93.46	92.86	98.70	90.77
DistilBERT	90.59	90.78	90.69	98.01	89.95

Overall, all models achieved high performance on aspect term extraction, with BERT and ConvBERT leading the pack. The differences between 92.2% and 89.95% accuracy may correspond to just a few sentences in the test set where DistilBERT failed to identify an aspect or predicted an incorrect category while BERT succeeded. The near-ceiling precision/recall of BERT indicates that transformer models can very effectively learn the task of marking aspect terms in Turkish despite the language's complexity. Likely, the large pre-trained knowledge combined with fine-tuning allowed BERT to recognize even infrequent aspect terms or morphological variants. One minor point is the overfitting tendency: BERT's train vs test discrepancy suggests caution — with such a small dataset, it would be easy for a powerful model to memorize training examples. We mitigated this by early stopping; still, the model is almost too perfect on training data. Future work could use cross-validation or more data augmentation to ensure robustness.

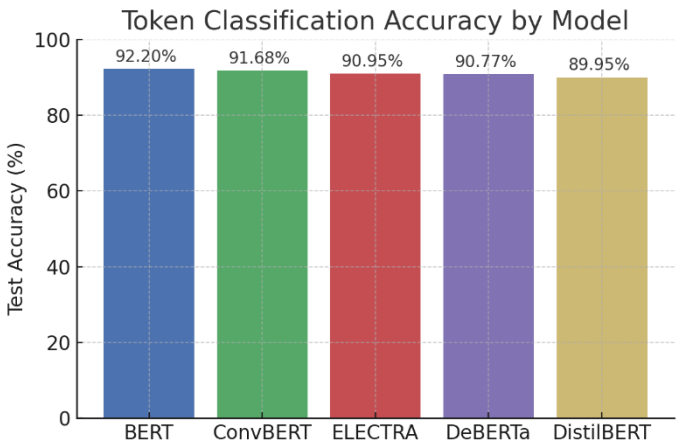

Figure 4. Comparison of token-level aspect extraction accuracy for each model on the Turkish review test set.

When Figure 4 is examined, all models show strong performance in appearance term and category identification, exceeding 89% accuracy. This is a promising outcome, as aspect extraction is often the first step in ABSA pipelines—missing an aspect would mean we cannot determine its sentiment later. In our case, the models' errors are minimal. Most of the few mistakes involved either predicting the wrong category for a correctly identified aspect term (for example, labeling a FOOD#STYLE_OPTIONS aspect as FOOD#QUALITY – a reasonable confusion if the model wasn't sure), or missing extremely subtle aspect mentions (such as very implicit aspects). We also note that DistilBERT, while about 2-3% less accurate, might be an attractive option when computational resources are limited, as it still achieves ~90% accuracy with a much smaller model size.

B. Sentiment Analysis with Sequential Classification Model

As a result of performing sentiment analysis through sequential classification using the BERT model, the BERT loss-epoch graph is illustrated in Figure 5. Additionally, the confusion matrix generated from the sentiment analysis is shown in Figure 6. Various metrics were employed to evaluate the model's performance. The training process yielded an F-Score of 67.74%, a recall of 73.22%, a precision of 65.63%, and an accuracy of 82.72% on the test data. The model's performance across the three sentiment classes (positive, negative, and neutral) is further analyzed using these metrics, with the detailed results presented in Table 3.

TABLE III. RESULTS OF BERT MODEL ON SENTIMENT CLASSES (%)

Metrics	Positive	Negative	Neutral	Accuracy
Precision	96.19	75.71	25.00	82.72
Recall	82.78	86.88	50.00	82.72
F-Score	88.98	80.91	33.33	82.72

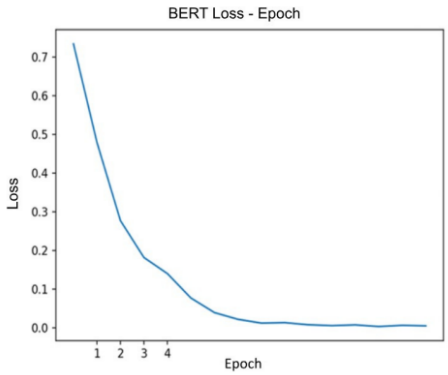

Figure 5. BERT loss - epoch graph.

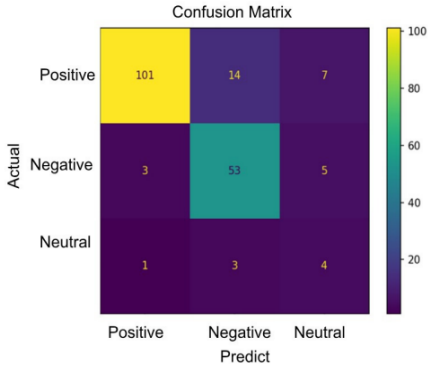

Figure 6. BERT model confusion matrix for sentiment analysis.

As a result of performing sentiment analysis with sequential classification with the ConvBERT model, the ConvBERT loss-epoch graph is shown in Figure 7.

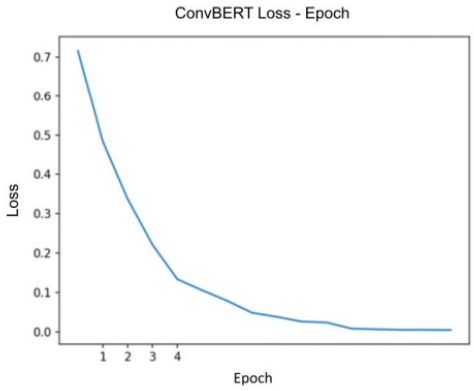

Figure 7. ConvBERT loss - epoch graph.

The results of sentiment analysis using sequential classification with the ConvBERT model are summarized, with the confusion matrix shown in Figure 8. The training process yielded an F-Score of 74.11%, a recall of 82.37%, a precision of 71.89%, and an accuracy of 86.91% on the test data. The performance across the three sentiment classes (positive, negative, and neutral) was evaluated in detail using various metrics, with the results presented in Table 4.

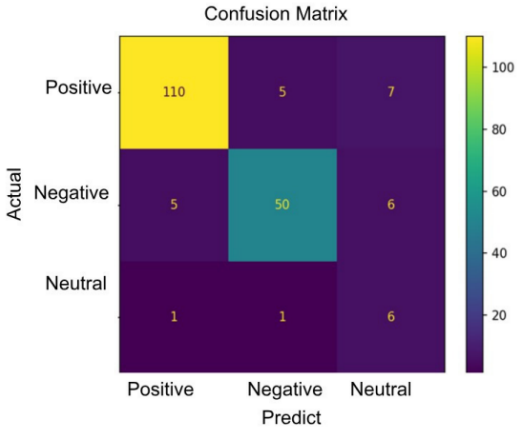

Figure 8. ConvBERT model confusion matrix for sentiment analysis.

TABLE IV. RESULTS OF CONVERT BERT MODEL ON SENTIMENT CLASSES (%)

Metrics	Positive	Negative	Neutral	Accuracy
Precision	94.82	89.28	31.57	86.91
Recall	90.16	81.96	75.02	86.91
F-Score	92.43	85.47	44.44	86.91

Similarly, sentiment analysis using the ELECTRA model with sequential classification is illustrated through the loss-epoch graph shown in Figure 9. Following training with the ELECTRA model, an F-Score of 69.38%, a recall of 73.22%, a precision of 68.54%, and an accuracy of 85.34% were achieved on the test data.

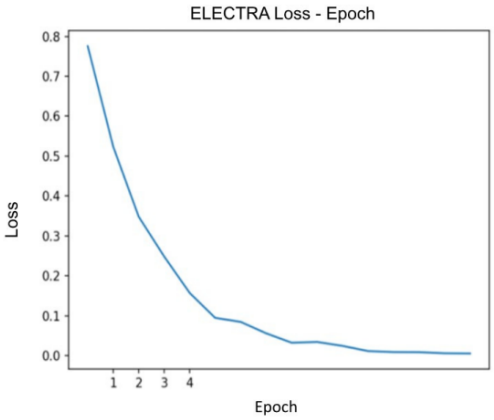

Figure 9. ELECTRA loss - epoch graph.

As a result of performing sentiment analysis with sequential classification with the ELECTRA model, the confusion matrix is presented in Figure 10. The success of the training, which was conducted in three classes: positive, negative, and neutral, was evaluated in detail with metrics and presented in Table 5.

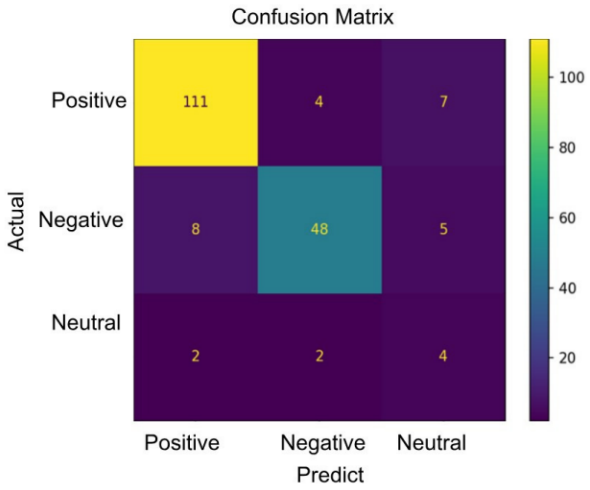

Figure 10. ELECTRA model confusion matrix for sentiment analysis.

TABLE V. RESULTS OF ELECTRA MODEL ON SENTIMENT CLASSES (%)

Metrics	Positive	Negative	Neutral	Accuracy
Precision	88.79	71.92	22.22	77.48
Recall	84.42	67.21	50.00	77.48
F-Score	86.55	69.49	30.76	77.48

The results of sentiment analysis using sequential classification with the DistilBERT model are presented, with the loss-epoch graph displayed in Figure 11. Following the sentiment analysis, the confusion matrix is shown in Figure 12. The training process yielded an F-Score of 62.27%, a recall of 67.21%, a precision of 60.98%, and an accuracy of 77.48% on the test data.

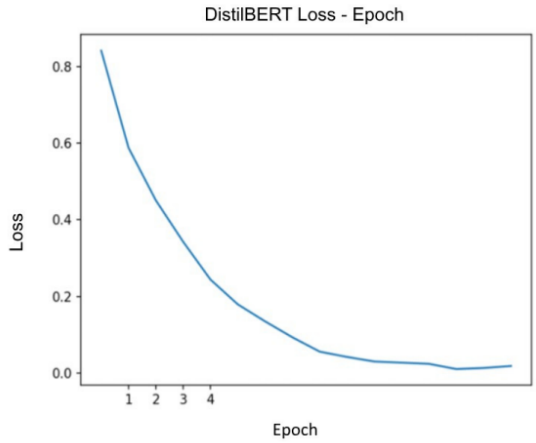

Figure 11. DistilBERT loss - epoch graph.

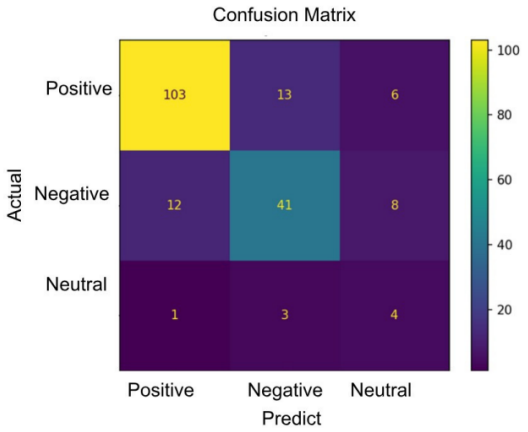

Figure 12. DistilBERT model confusion matrix for sentiment analysis.

The success of the training, which was carried out in three classes: positive, negative, and neutral, was evaluated in detail with metrics and presented in Table 6.

TABLE VI. RESULTS OF DISTILBERT MODEL ON SENTIMENT CLASSES (%)

Metrics	Positive	Negative	Neutral	Accuracy
Precision	88.79	71.92	22.22	77.48
Recall	84.42	67.21	50.00	77.48
F-Score	86.55	69.49	30.76	77.48

In this study, we employed two distinct methods: token classification and sequential classification models. For the token classification method, five different models were utilized: BERT, ConvBERT, ELECTRA, DeBERTa, and DistilBERT. The respective accuracy scores of these models on the test data were 92.2%, 91.68%, 90.95%, 90.77%, and 89.95%. Among these, the BERT model, which achieved the highest accuracy score, was selected as the preferred model for the study. This result highlights that the BERT model is the optimal choice for this research. To evaluate the performance of the models, metrics such as accuracy, recall, precision, and F-Score were used. A comparison of the performance metrics is shown in Figure 13, where it is evident that the BERT model outperforms the others, achieving the highest F-Score.

In addition, sequential classification models for different sentiment analysis tasks were compared, and the success of these models was assessed. The models used in this comparison were ConvBERT, ELECTRA, BERT, and DistilBERT, and their accuracy scores were compared. The accuracy scores were 86.91%, 85.34%, 82.72%, and 77.48%, respectively. Based on these results, the ConvBERT model demonstrated the highest success in the sequential classification task, with an accuracy score of 86.91%. Similar to the token classification method, the models' performances were measured using accuracy, recall, precision, and F-Score metrics.

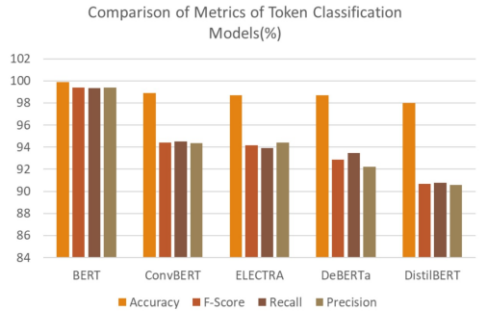

Figure 13. Comparison of metrics of token classification models (%).

The comparative analysis shows that the ConvBERT model has the highest accuracy score. Figure 14 shows the comparison of the metrics of the models' performance. In this graph, it is seen that the ConvBERT model achieves a higher F-score than the other models and performs better than the other models. Nevertheless, the overall performance is strong and a significant improvement over prior approaches in Turkish. Previously, one might have had to do sentiment analysis with lexicons or simpler classifiers, which likely would have much lower accuracy on such nuanced data. Our best model (ConvBERT) correctly classifies over 86% of aspect sentiments, which is quite high for a three-class problem in a morphologically rich language.

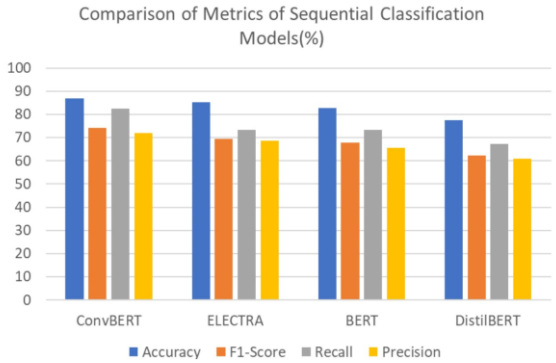

Figure 14. Comparison of metrics of sequential classification models(%).

VI. CONCLUSION

In this paper, we presented a two-stage aspect-based sentiment analysis approach for Turkish, leveraging token classification for aspect extraction and sequential classification for sentiment prediction. The study was motivated by the limited number of existing Turkish ABSA studies and the need for more accurate methods that account for the language's unique characteristics. Our approach can be applied across various domains (e.g., tourism, gastronomy, retail) to automatically analyze Turkish customer feedback. For instance, in a set of restaurant or hotel reviews, our model can identify aspects such as service, food, ambiance, and cleanliness, and determine the sentiment expressed about each. A single customer comment may convey multiple opinions (e.g., praising the food but criticizing the service); by identifying and linking these aspect-specific sentiments, our system can provide granular insights that help businesses target their improvements. Overall, the ability to break down a review into aspect-level sentiments offers valuable information for improving service quality and customer satisfaction, as well as developing effective marketing strategies focused on specific strengths or weaknesses. The experimental results with five different token classification models (BERT, ConvBERT, ELECTRA, DeBERTa, and DistilBERT) demonstrate that very high accuracy can be achieved on Turkish aspect term extraction—our best model (BERT) correctly identified over 92% of aspect terms and their categories on the test data. Furthermore, a comparative evaluation of four sequential sentiment classification models (ConvBERT, ELECTRA, BERT, and DistilBERT) shows that the ConvBERT model achieved the highest accuracy (about 87%) in determining aspect-level sentiment. This two-stage model substantially outperforms earlier approaches applied to Turkish, which were often constrained by simpler machine learning models or manual feature engineering. By fine-tuning pre-trained transformers, we harnessed a wealth of linguistic knowledge, enabling the models to overcome many challenges posed by Turkish (such as agglutinative morphology and free word order). Despite the strong results, there is room for improvement and further research. One limitation of our current work is the handling of neutral sentiment, which proved challenging due to its low frequency. Future studies could address this by balancing the training data or employing data augmentation for neutral examples, or by using advanced techniques like semi-supervised learning to take advantage of unlabeled data. Another possible extension is to integrate the two stages into a single joint model (for example, using a multi-task learning framework where a transformer model predicts aspect spans and sentiment labels together). This could potentially reduce error propagation, though it would require careful design to handle the complexity. Additionally, applying our approach to other Turkish datasets or other domains (such as product reviews, or social media posts) would test its generality. We anticipate that the models would maintain robust performance, given the fundamental language understanding captured by transformers, though domain-specific nuances might require further fine-tuning. In summary, this work contributes a novel transformer-based ABSA pipeline for the Turkish language and establishes strong baseline results for future research. The approach can directly benefit applications that need to automatically analyze Turkish text for fine-grained sentiment, enabling more scalable and detailed understanding of customer opinions in Turkish. We hope that our findings encourage further exploration of advanced NLP models for Turkish and other under-studied languages in the context of sentiment analysis.

REFERENCES

[1] B. Kama, M. Ozturk, P. Karagoz, I. H. Toroslu, and O. Ozay, "A web search enhanced feature extraction method for aspect-based sentiment analysis for Turkish informal texts," in Big Data Analytics and Knowledge Discovery: 18th International Conference, DaWaK 2016, Porto, Portugal, vol. 18, Springer, 2016, pp. 225–238.

[2] B. Kama, M. Ozturk, P. Karagoz, I. H. Toroslu, and M. Kalender, "Analyzing implicit aspects and aspect dependent sentiment polarity for aspect-based sentiment analysis on informal Turkish texts," in *Proceedings of the 9th International Conference on Management of Digital EcoSystems*, 2017, pp. 134–141.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv.org, Oct. 11, 2018. https://arxiv.org/abs/1810.04805

[4] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.

[5] Z. H. Jiang, W. Yu, D. Zhou, Y. Chen, J. Feng, and S. Yan, "ConvBERT: Improving BERT with span-based dynamic convolution," Advances in Neural Information Processing Systems, vol. 33, pp. 12837–12848, 2020.

[6] K. Clark, M. T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pretraining text encoders as discriminators rather than generators," arXiv preprint arXiv:2003.10555, 2020.

[7] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced BERT with disentangled attention," arXiv preprint arXiv:2006.03654, 2020.

[8] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 2005, pp. 347–354.

[9] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.

[10] Y. Wang, A. Sun, J. Han, Y. Liu, and X. Zhu, "Sentiment Analysis by Capsules," *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, 2018, doi: https://doi.org/10.1145/3178876.3186015.

[11] X. Hou, J. Huang, G. Wang, X. He, and B. Zhou, "Selective attention based graph convolutional networks for aspect-level sentiment classification," arXiv preprint arXiv:1910.10857, 2019.

[12] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "SemEval-2016 Task 5: Aspect-based sentiment analysis," in Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 19–30.

[13] D. Zhang, Z. Zhu, Q. Lu, H. Pei, W. Wu, and Q. Guo, "Multiple interactive attention networks for aspect-based sentiment classification," *Applied Sciences*, vol. 10, p. 2052, 2020.

[14] E. H. d. Silva and R. M. Marcacini, "Aspect-based sentiment analysis using BERT with disentangled attention," in Proceedings, 2021.

[15] H. Yang and K. Li, "Improving implicit sentiment learning via local sentiment aggregation," arXiv preprint arXiv:2110.08604, 2021.

[16] K. Scaria, H. Gupta, S. A. Sawant, S. Mishra, and C. Baral, "InstructABSA: Instruction learning for aspect based sentiment analysis," arXiv preprint arXiv:2302.08624, 2023.

[17] H. Türkmen, S. İ. Omurca, and E. Ekinci, "An aspect based sentiment analysis on Turkish hotel reviews," *Girne American University Journal of Social and Applied Sciences*, vol. 6, pp. 12–15, 2016.

[18] E. Ekinci, H. Türkmen, and S. İ. Omurca, "Multi-word aspect term extraction using Turkish user reviews," International Journal of Computer Engineering and Information Technology, vol. 9, p. 15, 2017.

[19] F. S. Çetin and G. Eryiğit, "Türkçe hedef tabanlı duygu analizi için alt görevlerin incelenmesi – hedef terim, hedef kategori ve duygu sınıfı belirleme," *Bilişim Teknolojileri Dergisi*, vol. 11, pp. 43–56, 2018, doi:10.17671/gazibtd.325865.

[20] P. Karagoz, B. Kama, M. Ozturk, I. H. Toroslu, and D. Canturk, "A framework for aspect based sentiment analysis on Turkish informal texts," Journal of Intelligent Information Systems, vol. 53, pp. 431–451, 2019.

[21] D. Ozkan, *Aspect-based sentiment analysis in Turkish*, M.S. thesis, Atılım University, 2020.

[22] M. U. Salur, İ. Aydın, and M. Jamous, "An ensemble approach for aspect term extraction in Turkish texts," *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, vol. 28, pp. 769–776, 2021.

[23] A. B. Girgin, G. Gümüşçekiççi, and N. C. Birdemir, "Turkish sentiment analysis: A comprehensive review," *Sigma Journal of Engineering and Natural Sciences*, vol. 42, no. 4, pp. 1292–1314, 2023.

[24] P. Meesad, "Thai fake news detection based on information retrieval, natural language processing and machine learning," *SN Computer Science*, vol. 2, p. 425, 2021.

[25] J. C. Pereira-Kohatsu, L. Quijano-Sánchez, F. Liberatore, and M. Camacho-Collados, "Detecting and monitoring hate speech in Twitter," *Sensors*, vol. 19, p. 4654, 2019.

[26] G. Salton, R. J. Ross, and J. D. Kelleher, "Idiom token classification using sentential distributed semantics," unpublished.

[27] T. Wang, R. Sridhar, D. Yang, and X. Wang, "Identifying and mitigating spurious correlations for improving robustness in NLP models," *arXiv preprint* arXiv:2110.07736, 2021.

[28] C. Li, X. Peng, H. Peng, J. Li, and L. Wang, "TextGTL: Graph-based transductive learning for semi-supervised text classification via structure-sensitive interpolation," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2021, pp. 2680–2686.

[29] P. L. Prasanna and D. R. Rao, "Text classification using artificial neural networks," *International Journal of Engineering & Technology*, vol. 7, pp. 603–606, 2018.

[30] K. Zhao, L. Huang, R. Song, Q. Shen, and H. Xu, "A sequential graph neural network for short text classification," *Algorithms*, vol. 14, p. 352, 2021.