applied sciences

MDPI

Article

Towards Better Sentiment Analysis in the Turkish Language: Dataset Improvements and Model Innovations

Kevser Büşra Zümberoğlu [1,2], Sümeyye Zülal Dik [1,2], Büşra Sinem Karadeniz [1,2] and Shaaban Sahmoud [1,2,*]

[1] Computer Engineering Department, Fatih Sultan Mehmet Vakif University, Istanbul 34015, Turkey; kbzumberoglu@fsm.edu.tr (K.B.Z.); szdik@fsm.edu.tr (S.Z.D.); busrasinem.karadeniz@stu.fsm.edu.tr (B.S.K.)
[2] Data Science Application and Research Center (VEBIM), Fatih Sultan Mehmet Vakif University, Istanbul 34015, Turkey
* Correspondence: ssahmoud@fsm.edu.tr

Abstract: Sentiment analysis in the Turkish language has gained increasing attention due to the growing availability of Turkish textual data across various domains. However, existing datasets often suffer from limitations such as insufficient size, lack of diversity, and annotation inconsistencies, which hinder the development of robust and accurate sentiment analysis models. In this study, we present a novel enhanced dataset specifically designed to address these challenges, providing a comprehensive and high-quality resource for Turkish sentiment analysis. We perform a comparative evaluation of previously proposed models using our dataset to assess their performance and limitations. Experimental findings demonstrate the effectiveness of the presented dataset and trained models, offering valuable insights for advancing sentiment analysis research in the Turkish language. These results underscore the critical role of the enhanced dataset in bridging the gap between existing datasets and the importance of training the modern sentiment analysis models on scalable, balanced, and curated datasets. This can offer valuable insights for advancing sentiment analysis research in the Turkish language. Furthermore, the experimental results represent an important step in overcoming the challenges associated with Turkish sentiment analysis and improving the performance of existing models.

Keywords: sentiment analysis; turkish language; turkish sentiment analysis; BERT

Received: 14 January 2025
Revised: 8 February 2025
Accepted: 9 February 2025
Published: 16 February 2025

Citation: 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. https://doi.org/10.3390/app15042062

1. Introduction

Sentiment analysis has emerged as a rapidly advancing research area in natural language processing (NLP), driven by the increasing need to analyze user-generated content [1]. In contemporary society, individuals exhibit a strong inclination to understand the emotions and thoughts of others who have had similar experiences, even when making relatively simple decisions. Studies indicate that 84% of internet users perceive online reviews as personal recommendations, and 68% form decisions after reading only one to six reviews. Consequently, the proliferation of the internet and social media has transformed user-generated online reviews into a highly valuable data source, characterized by its accessibility and diversity. These reviews play a critical role not only in guiding individual decisions but also in enabling companies to identify user preferences and trends. Moreover, sentiment analysis facilitates decision-making based on societal emotional trends by examining the rapidly expanding volume of user-generated content on digital platforms such as social media, blogs, and forums. Users' opinions about products and services provide businesses with opportunities to analyze this content and monitor market trends. Moreover,

organizations and governments can make strategic decisions informed by emotional trends derived from social media posts [2].

Sentiment analysis methods process texts to identify positive, negative, or neutral sentiments expressed in various forms, such as product reviews, movie critiques, or social media posts. This capability is particularly critical for e-commerce companies that aim to understand online trends, enhance product and service quality, and respond effectively to customer preferences. For example, sentiment analysis can determine which products are more popular, assess audience reactions to a movie, or identify trending music [3]. As the volume of data on the internet continues to grow, sentiment analysis has become a crucial tool for enhancing decision-making processes and gaining insights into societal emotional and cognitive trends. However, while significant progress has been achieved in sentiment analysis for widely spoken languages such as English, unique challenges arise when working with Turkish due to its linguistic characteristics. Turkish, as an agglutinative language, forms words by adding a variety of suffixes to root words. This morphological structure results in a vast range of word forms, creating complex linguistic patterns that can impact the accuracy of sentiment classification [4]. For instance, a single lemma in Turkish may have hundreds of surface forms due to inflectional and derivational morphology, significantly increasing vocabulary size and leading to data sparsity in statistical models. Additionally, Turkish exhibits extensive vowel harmony, consonant alternations, and long words formed by suffix concatenation, where a single misspelled word can disrupt sentence-level comprehension [5]. Additionally, the flexible syntax of Turkish introduces further complexities to text processing tasks such as sentiment analysis, making it more challenging to extract accurate insights from Turkish texts. The flexible syntax of Turkish introduces further complexities to text processing tasks such as sentiment analysis, as word order is not fixed and can vary significantly depending on context and emphasis [6]. Furthermore, the challenges of processing informal Turkish social media texts are magnified due to the prevalence of missing vowels, diacritics, emoticons, slang, and regionally influenced spelling errors.

Another notable limitation is the scarcity of large-scale labeled datasets for Turkish, in contrast to English and other widely spoken languages. This shortage poses a significant challenge for training robust models using machine learning and deep learning techniques. To address these challenges, this study introduces a balanced and extensive dataset to advance the field of Turkish sentiment analysis. The amount of digital data resources related to Turkish is quite limited. This highlights the need for more comprehensive efforts to expand the scope and diversity of datasets prepared in the Turkish language. Enriching Turkish digital data resources will lead the development of language-focused datasets and support advancements in the field.

The primary objective of this study is to provide resources for the development of prediction models specific to the Turkish language and to establish a new basis for evaluating the performance of existing models. This approach seeks to gain a deeper understanding of both the dataset's and the models' ability to capture and analyze the complex linguistic structure of Turkish. Additionally, finding neutral data in Turkish datasets presents a significant challenge, as many existing datasets often include non-sensical or irrelevant information labeled as neutral. To address this deficiency, the dataset we prepared includes many neutral labels sourced from categories such as tweets and product reviews. This ensures that the models are trained to distinguish neutral and diverse sentiments effectively, supported by high-quality and reliable data.

The proposed dataset contains 2333 positive, 2334 neutral, and 2333 negative sentences, ensuring a balanced structure that facilitates fair and comprehensive model training. Balanced datasets are critical for achieving accurate and consistent results across different

sentiment classes. Two data augmentation techniques, back-translation and synonym replacement were applied to enhance model performance and increase data diversity. Back-translation involves translating a sentence into another language and then back into the original language to introduce linguistic variations, while synonym replacement substitutes words with their synonyms to create alternative sentence structures while preserving meaning. To maintain the dataset's balance, a second version expanded dataset was created, comprising 15,853 examples: 5284 positive, 5206 neutral, and 5363 negative sentences.

Using the proposed dataset, a range of machine learning and deep learning models were trained and evaluated. This study employed state-of-the-art deep learning architectures specifically designed or fine-tuned for the Turkish language, including ELECTRA Base Turkish (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) [7], a pre-trained transformer model optimized for token-level replacement tasks, which improves training efficiency by focusing on detecting replaced tokens. BERT Base Turkish (Bidirectional Encoder Representations from Transformers) [8], a bidirectional transformer model pre-trained on Turkish text, is particularly effective in capturing contextual relationships through its transformer-based attention mechanisms. Additionally, TurkishBERTweet [9], tailored for Turkish social media data, effectively handles informal and colloquial text, and BERT Base Multilanguage [10], capable of processing over 100 languages including Turkish, provides robust cross-lingual capabilities. Turkish RoBERTa Base (A Robustly Optimized BERT Approach) [11] offers improved performance by fine-tuning on Turkish-specific datasets, while XLM-RoBERTa Base (Cross-lingual Language Model Robustly Optimized BERT Approach) [12] is a versatile model pre-trained on multiple languages, making it suitable for cross-lingual tasks.

Alongside these advanced models, traditional machine learning methods were also explored. For instance, the Linear Support Vector Classifier (SVC) [13] is highly effective in handling high-dimensional data by finding the optimal hyperplane for classification, while Logistic Regression [14] provides a probabilistic approach for binary or multi-class classification tasks. Naive Bayes [15], known for its simplicity and efficiency, performs well in text classification by utilizing probabilistic methods. K-Nearest Neighbors (KNN) [16] relies on distance-based similarity measures to classify data points, making it intuitive for small datasets. Lastly, Decision Tree [17] creates interpretable models by splitting data based on feature importance, facilitating straightforward decision-making processes. This comprehensive approach enabled an evaluation of the relative performance of these models in Turkish sentiment analysis. Figure 1 shows the main steps of the sentiment analysis task. Turkish sentiment analysis presents significant opportunities for companies, researchers, and organizations. The accurate sentiment analysis tools developed in this study can provide valuable and actionable insights in various domains, such as understanding societal opinions in Turkish-speaking regions, evaluating customer feedback, assessing political sentiments, or analyzing social media trends.

The rest of this paper is organized as follows: Section 2 summarizes the research work on sentiment analysis in general with a focus on the English and Turkish languages. Section 3 presents our new dataset and provides its characteristics and some statistics. The experimental settings and results are given in Section 4. Section 5 concludes this paper and summarizes this paper's findings.

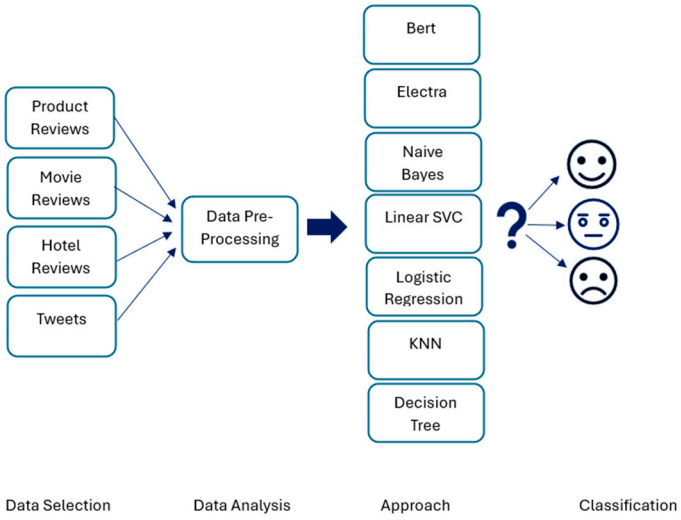

Figure 1. Sentiment analysis process flow.

2. Related Work

This section examines studies in the field of sentiment analysis in the current literature and the techniques employed in these studies. In recent years, numerous studies have focused on automatically detecting emotional tendencies in texts. These studies primarily aim to detect emotional tendencies in various data sources such as social media, customer feedback, and news content. Several studies in the literature have explored different approaches for sentiment analysis, including traditional machine learning methods such as Support Vector Machines (SVMs) [18] and Naive Bayes, deep learning models like Long Short-Term Memory (LSTM) [19] and Convolutional Neural Network (CNN) [20], and hybrid architectures that combine rule-based [21] and learning-based approaches [22]. In machine learning methods, text analysis is performed using word-based statistical models. Deep learning methods typically use powerful models that can capture more complex language patterns and contextual relationships. Hybrid models usually involve combining different models for classification purposes [23,24].

Regarding machine learning approaches, authors used classification algorithms such as Naive Bayes, Logistic Regression, SVM, KNN, Decision Tree, Extreme Gradient Boosting (XGBoost) [25], Adaptive Boosting (AdaBoost) [26], and Random Forest [27]. For deep learning and transformer models, authors used Multi-Layer Perceptron (MLP) [28], LSTM, CNN, Recurrent Neural Network (RNN) [29], BiLSTM (Bidirectional Long Short-Term Memory) [30], BERT, RoBERTa, Generative Pre-trained Transformer 3 (GPT-3) [31], Large Language Model Meta AI (Llama) [32], and hybrid models. Hybrid approaches, which combine multiple techniques for improved performance, have also gained attention in recent studies. For example, CNN-BiLSTM models integrate convolutional layers for feature extraction and BiLSTM layers for capturing long-range dependencies, enhancing sentiment classification accuracy [33]. Additionally, lexicon-based and machine learning hybrid approaches leverage rule-based sentiment scoring combined with supervised learning algorithms to improve robustness [34]. Ensemble-based hybrid models, such as those incorporating Random Forest and Gradient Boosting classifiers, have also been proposed to enhance sentiment classification reliability [35]. These approaches require two datasets for training and evaluation. The training dataset is used for the algorithm to learn the features of the relevant domain, while the evaluation dataset is used to validate the model created from the training dataset. Hyperparameter tuning in models impacts the performance of the classifier, so fine-tuning is observed in most models.

The success of a model depends on the effectiveness of the method used in the feature extraction phase. Among the most frequently used feature extraction methods are Bag of Words, Term Frequency-Inverse Document Frequency (TF-IDF), n-grams (unigrams, bigrams, and trigrams), POS tagging-based features, and dependency rule-based features [36–45].

In [24], a hybrid sentiment analysis model for Turkish texts was presented by combining lexicon-based methods and machine learning classifiers. Lexicon-based methods rely on predefined word lists containing words and synonyms that represent emotional states, followed by the calculation of a predicted sentiment score [36]. In this approach, the polarities of words in tweets were evaluated using the open-source Orange program, and positive, negative, or neutral labels were assigned [24]. Mladenovic, Djordje et al. [37] proposed a TF-IDF-based approach for improving the perception of malicious actions' emotions and contexts by offering a more robust, adaptable, and sustainable method. This study also employed various modern optimization methods, including a modified version of the Red Fox optimization algorithm, for hyperparameter tuning. The performance of four variations of Naive Bayes was thoroughly evaluated in the study conducted by Danyal et al. [38]. This research explored two vectorization techniques, namely TF-IDF and count vectorizer, using movie review datasets. The results highlighted that TF-IDF offered a slight improvement in classification performance compared to the count vectorizer, making it a more effective approach for text feature extraction.

Deep learning-based sentiment analysis has seen significant advancements. Alizadeh and Seilsepour [39] introduced a novel self-supervised sentiment classification method that employs semantic labeling based on contextual embeddings. By leveraging cosine similarity to evaluate semantic relationships in positive and negative words, they successfully trained a hybrid CNN-GRU model for sentiment classification, achieving remarkable accuracy in handling unseen data.

In the judicial domain, Abimbola et al. [40] proposed a machine learning framework combining LSTM and CNN algorithms to analyze Canadian maritime case law. Their model included two core components: one for text classification using word embeddings and another for event detection in time-series data via RNNs. This innovative approach improved the efficiency of legal document analysis.

Hybrid architectures have also gained traction in sentiment analysis. Ba Alawi and Bozkurt [41] developed a model integrating BERT, BiLSTM, and CNN for assessing sentiment and satisfaction in Turkish university-related tweets. Their approach leveraged multi-layered contextual embeddings to provide accurate predictions. Similarly, Jahin et al. [42] proposed the Transformer- and Attention-based Bidirectional LSTM (TRABSA) model. This hybrid framework combined RoBERTa-based transformers, attention mechanisms, and BiLSTM networks to improve sentiment analysis, particularly in aspect-based tasks. A sentiment analysis method for e-commerce product reviews, named Weighted Parallel Hybrid Deep Learning (WPHDL-SAEPR), has been introduced, utilizing the word2vec model for word embeddings and combining the Restricted Boltzmann Machine (RBM) and Singular Value Decomposition (SVD) for sentiment classification [43]. Similarly, researchers explored an ensemble model leveraging transformers and Large Language Models (LLMs) for sentiment analysis in foreign languages. This approach involved translating texts into English using LibreTranslate and Google Translate before conducting sentiment analysis, demonstrating that analyzing sentiments in English effectively captures sentiments from foreign languages [44].

In a different context, a study focused on feature extraction from social media and text data by employing methods such as Bag of Words, TF-IDF, Hashing Vectorizer, N-grams, and word embeddings. The Chi-Square feature selection technique was applied to enhance

model accuracy by eliminating irrelevant features [45]. Some researchers have taken an aspect-based sentiment analysis (ABSA) perspective, distinguishing sentiments related to specific aspects of a topic, product, or service. For instance, one study transitioned from word embedding to sentence embedding, leveraging the SBERT transformer model, while also innovating in label generation and topic modeling through Bayesian search clustering combined with Inverse Document Frequency (IDF) calculations [46]. To address the scarcity of labeled data in ABSA tasks, another approach utilized LLMs such as GPT-3.5-turbo (OpenAI, San Francisco, CA, USA) and Llama-3-70B (Meta Platforms, Inc., Menlo Park, CA, USA) to generate synthetic examples, thereby supporting model training in low-resource scenarios [47]. Further advancements include leveraging BERT's contextual understanding with affine attention mechanisms to define relationships between words in ABSA tasks, coupled with the Multi-Layered Enhanced Graph Convolutional Network (MLEGCN) model to improve relational complexity and word pair compatibility [48]. A distinct solution, the Deep Context-Aware Sentiment Analysis Model (DCASAM), integrates the Deep Bidirectional Long Short-Term Memory Network (DBiLSTM) and the Densely Connected Graph Convolutional Network (DGCN) to overcome the limitations of traditional sentiment analysis methods [49]. Another innovative proposal combines local semantic features extracted by the Local Semantic Feature Extractor (LSFE) with the global features of BERT/RoBERTa models, resulting in the PConvBERT and PConvRoBERTa models. This method incorporates adversarial training to improve robustness and Focal Loss to mitigate the effects of imbalanced datasets [50].

Going beyond traditional machine learning and pre-training requirements, some studies introduced mathematical optimization models and decision-making frameworks. One such model employs game theory to calculate performance scores by combining context scores, ratings, and sentiment scores, achieving Nash equilibrium through non-cooperative game theory [51]. Another study developed a smarter decision support system using the Sentiment Analysis-based Multi-person Multi-criteria Decision-Making (SA-MpMcDM) methodology, which combines natural language comments and numerical ratings with the DOC-ABSADeepL multi-task deep learning model for aspect detection and opinion separation [52]. Finally, innovative approaches such as the population game model have been proposed to bypass large dataset pre-training requirements, relying instead on lexicon-based context and sentiment scores to achieve high performance across multiple domains [53,54]. Additionally, the Ontology-Based Sentiment Analysis Process (OSAPS) represents domain knowledge through a hierarchical structure, incorporating entities, object properties, and their relationships to create a machine-readable ontology model [55]. Large Language Models (LLMs) have revolutionized sentiment analysis by offering a more nuanced and contextually aware approach compared to traditional methods. Their ability to understand complex linguistic patterns and subtle emotional cues allows for more accurate sentiment classification, especially in cases with sarcasm, irony, or implicit sentiment [56,57]. LLMs can be fine-tuned on specific datasets to improve performance in particular domains, or utilized in a zero-shot learning fashion, leveraging their pre-trained knowledge to perform sentiment analysis without explicit training data [58,59]. This adaptability makes LLMs highly versatile for various sentiment analysis tasks, from basic polarity detection to more complex aspect-based sentiment analysis and emotion recognition [60]. However, challenges remain in addressing biases present in training data and ensuring the robustness of LLMs in handling noisy or ambiguous text [61].

3. Dataset

In sentiment analysis research, open-source datasets containing positive, negative, and neutral categories are significantly limited compared to those excluding the neutral

class. Furthermore, sentiment categories (positive, negative, and neutral) can be classified as categorical or fine-grained polarities. In studies utilizing fine-grained sentiment polarity, data points slightly leaning toward positive or negative from neutral can be more precisely adjusted, facilitating classification. Conversely, in categorical sentiment polarity, the definitive classification of neutral sentiment poses greater challenges. These difficulties often lead to issues such as the mislabeling of neutral data during dataset preparation, which subsequently reduces the accuracy of models trained on these datasets. Consequently, many sentiment analysis studies rely on datasets that do not include a neutral category. Additionally, most existing open-source datasets are designed for widely spoken languages such as English, underscoring the need for a Turkish sentiment analysis dataset that includes neutral sentiment. To address this gap, the Fatih Sultan Mehmet Turkish Sentiment Analysis Dataset (FSMTSAD) was developed, offering a balanced and comprehensive resource for Turkish sentiment analysis studies.

A review of the existing literature and datasets revealed a dataset similar to this study: the BounTi [62] dataset, developed by Köksal et al. The BounTi dataset comprises Turkish tweets about universities in Turkey, manually labeled as positive, neutral, or negative. As presented in Table 1, this dataset includes 2348 positive, 4215 neutral, and 1401 negative entries, amounting to a total of 7964 data points.

Table 1. Distribution of data in the BounTi dataset.

Sentiment	Training	Validation	Test	Total
Positive	1691	188	469	2348
Neutral	3034	338	843	4215
Negative	1008	113	280	1401
Total	5733	639	1592	7964

Existing datasets often suffer from an imbalance in sentiment class distribution, with significantly fewer neutral examples, which exacerbates the difficulty of distinguishing the neutral class from positive and negative sentiments. This lack of balance not only hampers the ability to identify neutral sentiments but also impacts the overall robustness of models trained on such datasets. To address both the scarcity of neutral data and the broader class imbalance, FSMTSAD was meticulously constructed to achieve an equal representation of all sentiment classes. As shown in Table 2, FSMTSAD contains 2333 positive, 2334 neutral, and 2333 negative samples, ensuring that each class constitutes approximately 33.3% of the dataset. As shown in Table 3, the augmented version of FSMTSAD ensures that FSMTSAD maintains its balanced structure while providing a greater representation of neutral data compared to existing benchmarks. This balanced and augmented distribution enables models to effectively distinguish neutral sentiments while also mitigating the biases introduced by disproportionate class representations in other datasets.

Table 2. Distribution of data in the FSMTSA dataset.

Sentiment	Total
Positive	2333
Neutral	2334
Negative	2333
Total	7000

Table 3. Distribution of the augmented FSMTSA dataset.

Sentiment	Total
Positive	5284
Neutral	5206
Negative	5363
Total	15,853

The FSMTSAD dataset incorporates data from diverse real-world sources, as exemplified in Table 4. These sources include hotel, restaurant, movie, and e-commerce product reviews, alongside tweets covering various topics. Additionally, the dataset is augmented with texts generated by Large Language Models (LLMs) based on specific instructions to enhance its diversity and representativeness. For neutral data, preference was given to texts devoid of subjective judgments. Specifically, examples were selected where both positive and negative sentiments were present but equally balanced, or where the author refrained from expressing a definitive positive or negative opinion. Such criteria ensure the accurate representation of the neutral class.

Table 4. Some samples from the proposed the FSMTSA dataset.

Text	Type	Sentiment
Akşam 9 da kapanma olacak ya sanırım İstanbul'un trafik yoğunluğunun %50 si şuan Yeniköy 'de bu ne hal? https://t.co/ot	Tweet	−1
Vatandaşlar, oy kullanma hakkına sahiptirler, ulaşılabilirlik konusuna dikkat edilmektedir.	LLM-generated	0
İyi daha güzel yapabilir	Product comments	0
Kokusu güzel hafif diğer yumuşatıcılar gibi ağır yoğun bir kokusu yok. Bahar gibi kokuyor bahar aylarında tercih edilebilecek bir yumuşatıcı bence. Fiyatı çok uygundu indirimdeyken aldım. Hafif bahar kokuları sevenlere tavsiye ederim.	Product comments	1

In this study, two different data augmentation methods were employed to improve model performance and enhance data diversity. The first method involved the application of the back-translation technique, in which the original texts were translated into another language and then back into the original language, as illustrated in Figure 2. This process generated various textual variations.

Through these augmentation methods, the total number of data points was increased to 15,853, comprising 5284 positive, 5284 neutral, and 5363 negative examples, as detailed in Table 3. During the augmentation process, potential duplicates were systematically identified and removed to ensure the dataset's quality and uniqueness. While the original dataset maintains a perfectly balanced structure, the slight imbalance in the expanded dataset arises from the nature of data augmentation methods. Synonym replacement provides more variations for positive and negative words, while back-translation may introduce subtle shifts in neutrality. To preserve data quality, linguistic integrity was prioritized over exact numerical balance.

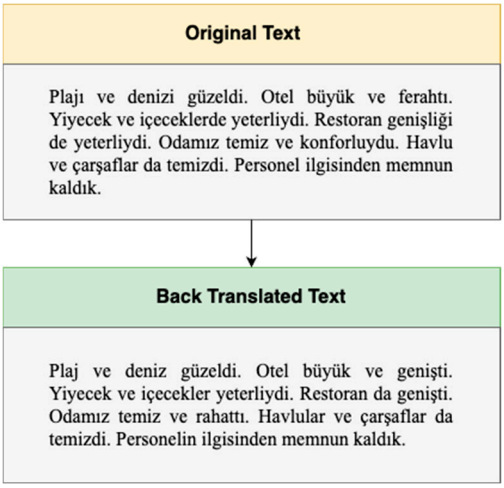

Figure 2. Sample of the back-translation method.

As a result of these enhancements, the proportions of positive, negative, and neutral classes in the expanded dataset were adjusted to 33.3%, 33.8%, and 32.8%, respectively.

The dataset annotation process was conducted manually. Each data sample was labeled by three different annotators. The majority votes for each data sample were reviewed by a supervisor. In cases of disagreement among the annotators, the responses were compared with the outputs of at least three different Large Language Models (LLMs), and the annotation for that sample was reperformed from scratch. This ensured that the data were accurately assigned to the correct class. The sentiment classes of the data were encoded as -1, 0, and 1, representing negative, neutral, and positive sentiments, respectively, as shown in Table 4.

Data preprocessing for this study involved several crucial steps to ensure data quality and suitability for analysis. For example, Twitter data were collected using the Twitter API by filtering based on relevant keywords and hashtags. Subsequently, both Twitter and review website data underwent a cleaning process, including removing irrelevant characters, HTML tags, URLs, and excessive whitespace. Special attention was paid to handling emojis and emoticons; these were either preserved for sentiment analysis or converted to their textual representations, depending on the specific task. Duplicate entries were identified and removed to avoid bias and inflate statistical significance. Text normalization techniques were employed for sentiment analysis, including lowercasing all text and handling contractions. Stop words, common words with little semantic value, were removed to focus on more informative terms. Finally, the text data were tokenized, splitting it into individual words or phrases for subsequent analysis. This comprehensive preprocessing pipeline aimed to minimize noise, standardize the data format, and enhance the signal-to-noise ratio for improved model performance.

The data were not pre-split into training, validation, and test sets; this division was made prior to model training. This ensured that different data samples were used each time to create the training, validation, and test sets, thereby preventing any potentially misleading effects of the selected data on the test results. Care was taken to ensure that word modifications in the added data did not change the sentiment polarity, so the annotations were made to ensure they had the same sentiment polarity as the original data from which they were generated. The histogram presented in Figure 3 was generated using Python's Matplotlib library. This visualization represents the distribution of word lengths in the dataset, where the x-axis shows the range of sentence lengths, and the y-axis indicates the number of words within each range.

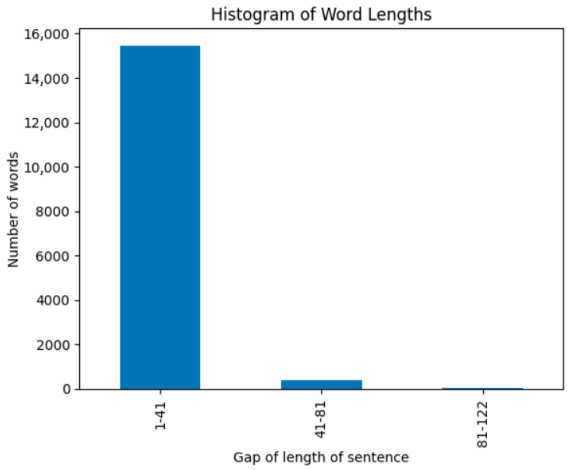

Figure 3. The histogram distribution of word lengths in the proposed FSMTSA dataset.

4. Experimental Results and Discussions

In this study, we conducted comprehensive experimental analyses to evaluate the performance and effectiveness of the FSMTSA dataset in Turkish Sentiment Analysis tasks. These analyses leveraged a combination of advanced deep learning-based language models and traditional machine learning algorithms to ensure a robust comparison. The deep learning models utilized in the experiments include BERT Base Turkish, XLM-RoBERTa Base, TurkishBERTweet, Turkish RoBERTa Base, and ELECTRA, representing state-of-the-art approaches in natural language processing. Meanwhile, the traditional machine learning algorithms assessed in the study encompass Naive Bayes, Linear SVC, Logistic Regression, KNN, and Decision Tree, providing a comprehensive baseline for performance evaluation.

The data were labeled as -1 for negative sentiment polarity, 0 for neutral polarity, and 1 for positive sentiment polarity. In the studies conducted with language models, before training the model, the negative polarity data were encoded as 0, the neutral ones as 1, and the positive ones as 2. The dataset was divided into a training set of 12,682 samples, a validation set of 1585 samples, and a test set of 1586 samples before the training process. Afterward, the text was converted into an appropriate input format for the language model or algorithm being used. In the first part of the experiment, training, validation, and testing phases were applied sequentially with language models. Among the language models, Electra Base Turkish, BERT Base Turkish, Turkish BERT Tweet, and Turkish RoBERTa Base performed better since they are pre-trained models for Turkish text data. BERT Base Multilingual and XLM-RoBERTa Base models, on the other hand, are trained on more than 100 languages, including Turkish. Therefore, they also performed well for Turkish, which has fewer resources. The training parameters for the language models were set as follows: a learning rate of 1×10^{-5}, training and evaluation batch size of 36, weight decay of 0.1, 10 training epochs, and 300 warmup steps. These parameters were chosen to ensure fair and consistent experimental conditions across datasets. Specifically, we adopted the same hyperparameter settings as those used in the BounTi dataset experiments to maintain comparability and prevent any bias in performance evaluation. This approach allowed for a robust and objective assessment of our model's effectiveness under identical training conditions.

To evaluate the performance of the models, standard classification metrics, including accuracy, precision, recall, and F1-score, were utilized. Accuracy measures the proportion of correctly classified instances across all predictions, providing an overall assessment of model correctness, and is computed using Equation (1). Precision, defined in Equation (2), quantifies how many of the instances predicted as positive are actually positive, ensuring that false positives are minimized. Recall, also referred to as sensitivity, evaluates the

model's ability to correctly identify actual positive instances, as shown in Equation (3), indicating how well it captures relevant cases. F1-score, which is the harmonic mean of Precision and Recall, is formulated in Equation (4) and offers a balanced measure by considering both false positives and false negatives, making it particularly useful when dealing with imbalanced data. These metrics were computed using standard formulas, where True Positives (TP) denote correctly classified positive instances, True Negatives (TN) represent correctly identified negative cases, False Positives (FP) occur when negative instances are misclassified as positive, and False Negatives (FN) refer to positive instances incorrectly predicted as negative. By incorporating these evaluation metrics, a comprehensive assessment of model performance was ensured.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{F}_1\text{Score} = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

In our tables, bold line indicates the best results obtained in the corresponding experiment instance. The validation results for the language models trained with these parameters are presented in Table 5, where the best result, 91.56%, belongs to Electra BERT Base. Following this, the BERT Base Turkish language model, which has high performance in Turkish language processing tasks, achieves a score of 91.56%.

Table 5. Validation results of the models compared after training with the FSMTSA dataset.

Validation Results				
Model	**Accuracy**	**Recall**	**Precision**	**F1**
Electra Base Turkish	**91.56%**	**91.57%**	**91.61%**	**91.58%**
Bert Base Turkish	91.16%	91.18%	91.17%	91.18%
TurkishBERTweet	88.48%	88.50%	88.53%	88.51%
Bert Base Multilang	85.96%	85.93%	86.17%	85.99%
Turkish RoBERTa base	85.72%	85.74%	85.83%	85.77%
XLM- RoBERTa Base	78.15%	78.16%	78.14%	78.13%

Looking at the test results in Table 6, it can be observed that language models specifically trained for Turkish outperform multilingual models. The F1 scores for Electra Base Turkish and BERT Base Turkish models are 90.64% and 90.21%, respectively. In contrast, the multilingual models, BERT Base Multilingual and XLM-RoBERTa Base, performed with lower F1 scores of 84.76% and 78.11%, respectively, on Turkish data.

The confusion matrices presented in Figure 4 were generated using Python, specifically with the Matplotlib and Seaborn libraries, based on the experimental results obtained in this study. Figure 4 shows the confusion matrices for each language model trained. Upon examining the confusion matrices, it can be seen that the models specifically trained for Turkish make fewer errors and demonstrate better class separation.

Table 6. Testing results of the models compared after training with the FSMTSA dataset.

	Test Results			
Model	**Accuracy**	**Recall**	**Precision**	**F1**
Electra Base Turkish	**90.64**	**90.66**	**90.63**	**90.64**
Bert Base Turkish	90.22	90.25	90.22	90.21
TurkishBERTweet	88.10	88.09	88.14	88.08
Bert Base Multilang	84.76	84.81	84.93	84.76
Turkish RoBERTa base	86.33	86.30	86.31	86.29
XLM-RoBERTa Base	78.13	78.18	78.27	78.11

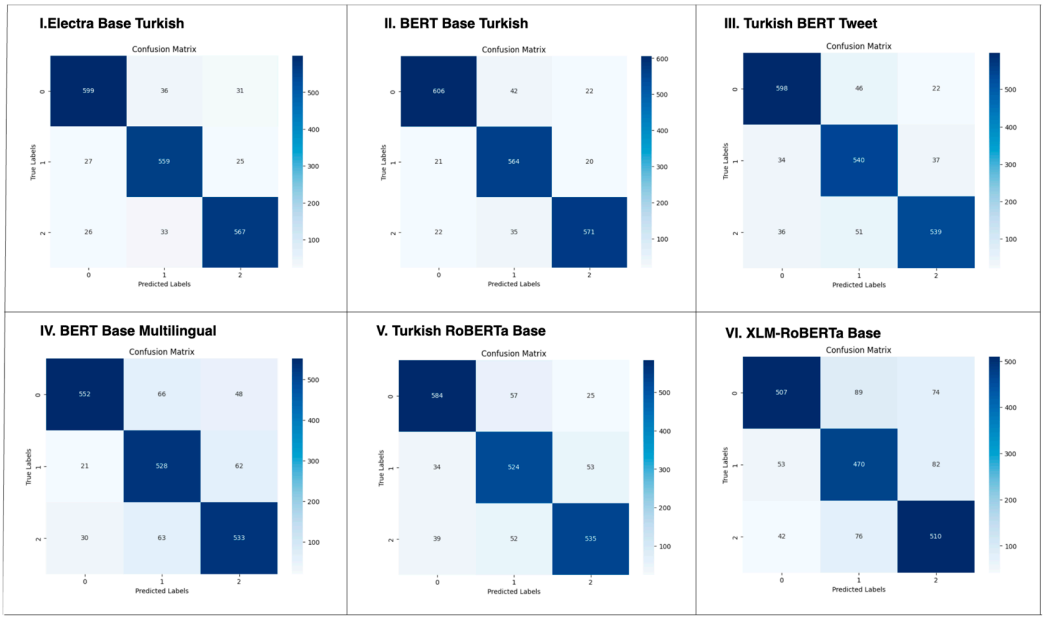

Figure 4. Confusion matrices of the trained language models using our proposed FSMTSA dataset.

The results shown in Table 7 reveal that deep learning-based language models outperform traditional machine learning algorithms. The highest accuracy rate of 88.36% is achieved with the Linear SVC (Support Vector Classifier) algorithm. Due to the linear separability assumption, the logistic regression algorithm may not fully model semantic relationships arising from context and suffixes in Turkish. If this model is not supported by morphological processing, the misclassification rate may increase due to word variations. Logistic Regression follows in second place with an accuracy rate of 87.44%. The other traditional methods such as Naive Bayes and KNN show lower performance, with accuracy rates of 85.77% and 81.05%, respectively. One reason for this result is that root-based feature extraction is not performed, the model cannot fully grasp the meaning changes arising from word context and suffixes in Turkish. This can decrease the model's performance. The KNN algorithm is in fourth place since it works directly at the word level and may struggle to capture the correct similarity between neighbors due to the agglutinative structure of Turkish. This could negatively impact the model's performance. Decision Tree follows in last place with the lowest accuracy rate of 76.03%. Due to word suffixes and syntactic flexibility in Turkish, the tree structure can become very complex. The branching structure may grow excessively, losing its ability to generalize. As a result, the overfitting level of the model may increase.

Table 7. The testing results of the models were compared after training with the FSMTSA dataset.

	Test Results			
Model	**Accuracy**	**Recall**	**Precision**	**F1**
Linear SVC	88.36	88.36	88.43	88.38
Logistic Regression	87.44	87.44	87.61	87.46
Naive Bayes	85.77	85.77	85.73	85.70
KNN	81.05	81.46	81.05	81.11
Decision Tree	76.03	76.03	76.08	76.04

Figure 5 presents the confusion matrices for each machine learning algorithm trained. Correct predictions are concentrated along the diagonal of the matrices, demonstrating that the predicted labels align closely with the true values. This alignment underscores the effectiveness of the proposed model in accurately capturing the patterns inherent in the FSMTSA dataset, which includes nuanced sentiment information specific to Turkish.

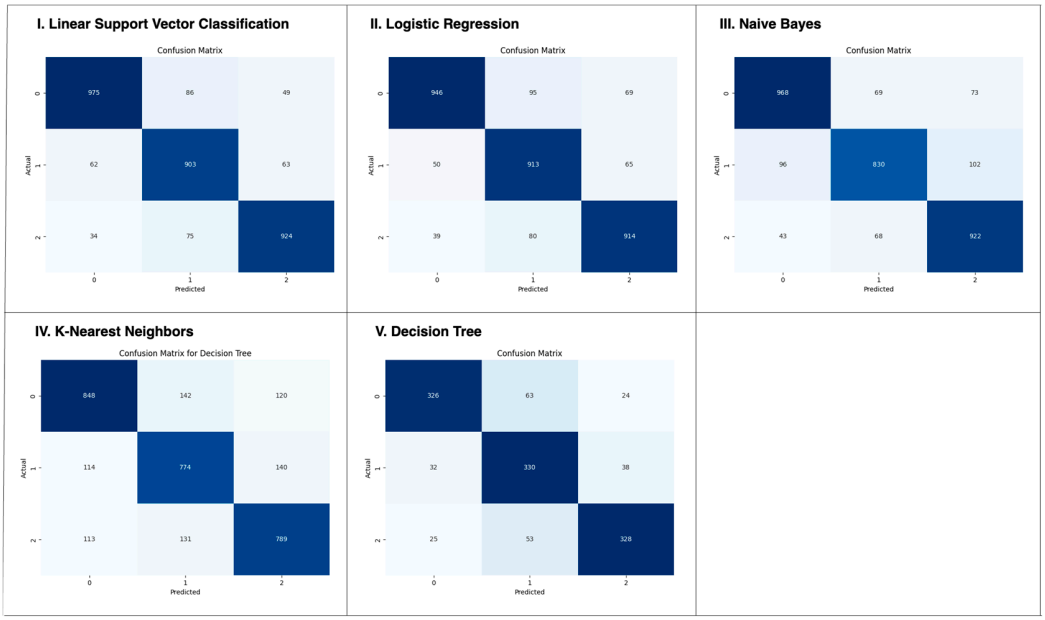

Figure 5. Confusion matrices for the results of the tested machine learning algorithms using our proposed FSMTSA dataset.

The performance comparison across different models on the FSMTSA and BounTi datasets (Tables 8 and 9) highlights significant variations influenced by dataset characteristics. Notably, the Electra Base Turkish model outperforms the other models in terms of accuracy, precision, recall, and F1 score on both datasets. Specifically, Electra achieved the highest validation accuracy of 91.56% and test accuracy of 90.64% on the FSMTSA dataset (Table 8). This superior performance may be attributed to Electra's replaced token detection mechanism, which potentially aligns well with Turkish's agglutinative structure and flexible syntax, enabling the model to better capture linguistic nuances. These results indicate the model's exceptional ability to capture semantic nuances within the FSMTSA data. Similarly, the precision, recall, and F1 metrics are consistently above 90%, reinforcing its robust performance across all evaluation dimensions. Conversely, the model's performance on the BounTi dataset drops considerably, with a validation accuracy of 72.52% and a test accuracy of 73.40% (Table 9). While Electra maintains a slight lead over other models for this dataset, its lower precision (69.81%) and F1 score (70.0%) on the test set suggest that

it struggles with the dataset's more diverse or potentially imbalanced distribution. This outcome may stem from the BounTi dataset containing more complex or noisier linguistic patterns compared to the FSMTSA dataset. Furthermore, the FSMTSA dataset's diverse and challenging structure ensures that the performance observed is not merely a result of larger datasets but reflects the model's ability to adapt and generalize across varied linguistic contexts. This adaptability indicates that Electra's architecture is not over-reliant on dataset size but rather leverages its structural advantages to effectively process nuanced language patterns, highlighting its practical applicability.

Table 8. Comparison of models' validation performance on FSMTSA dataset vs. BounTi dataset.

		Validation			
Dataset	**Model**	**Accuracy**	**Precision**	**Recall**	**F1**
FSMTSA	Electra Base Turkish	91.56	91.57	91.61	91.58
BounTi		72.52	68.70	69.80	69.20
FSMTSA	BERTurk	91.16	91.18	91.17	91.18
BounTi		70.60	73.00	71.50	74.50
FSMTSA	XLM-RoBERTa	85.96	85.93	86.17	85.99
BounTi		73.40	68.90	70.80	69.60
FSMTSA	Bert Multilanguage	78.15	78.16	78.14	78.13
BounTi		67.40	63.10	65.10	63.60

Table 9. Comparison of models' test performance on FSMTSA dataset vs. BounTi dataset.

		Test			
Dataset	**Model**	**Accuracy**	**Precision**	**Recall**	**F1**
FSMTSA	Electra Base Turkish	90.64	90.66	90.63	90.64
BounTi		73.40	69.81	70.36	70.00
FSMTSA	BERTurk	90.22	90.25	90.22	90.21
BounTi		69.20	72.90	70.10	72.30
FSMTSA	XLM-RoBERTa	78.13	78.18	78.27	78.11
BounTi		71.30	67.70	70.20	68.50
FSMTSA	Bert Multilanguage	84.76	84.81	84.93	84.76
BounTi		66.80	63.10	65.80	63.60

These findings underscore that while the FSMTSA dataset appears more conducive to yielding high classification performance, potentially due to cleaner or more homogeneous samples, the BounTi dataset presents challenges that impact the generalizability of even high-performing language models. The comparison with BERTurk, XLM- RoBERTa, and Bert Multilanguage further reinforces these observations. BERTurk, for instance, performs similarly well on FSMTSA (validation accuracy: 91.16%, test accuracy: 90.22%) but shows a larger performance decline on BounTi (validation accuracy: 70.6%, test accuracy: 69.2%). XLM- RoBERTa and Bert Multilanguage exhibit even more pronounced performance degradation on BounTi, with the latter scoring the lowest (validation accuracy: 67.4%, test accuracy: 66.8%). The Electra Base Turkish model's relatively consistent performance across datasets suggests that it may generalize better across varied data distributions compared to the other evaluated models. However, the significant performance gap under-

scores the necessity of dataset-specific model fine-tuning and highlights the importance of considering dataset variability in model evaluation.

This research presents a significant practical contribution to Turkish sentiment analysis by introducing FSMTSA, a high-quality, balanced dataset that enables better model training and evaluation. The dataset's diversity across multiple domains ensures that sentiment models trained on FSMTSA are applicable to a wide range of real-world scenarios, including customer review analysis, brand reputation monitoring, and public sentiment tracking. However, some limitations must be acknowledged. First, while FSMTSA is diverse, it does not yet cover all possible textual domains, such as legal or medical texts, which may require further expansion. Second, FSMTSA's annotation process, despite using multiple reviewers and language models for quality control, may still contain subjective biases, which could impact classification accuracy in edge cases (but it is still much better than automatically annotated datasets). Finally, this study primarily focuses on traditional deep learning models and does not include recent advancements in large-scale generative models (e.g., GPT-based architectures), which could be explored in future work. Despite these limitations, FSMTSA represents a crucial step forward in addressing the challenges of Turkish sentiment analysis, and it provides a strong foundation for further research in this field.

4.1. Dataset Generalizability and Robustness

4.1.1. Testing with Noisy Samples

In everyday life, texts often contain various sources of noise such as spelling mistakes, repeated characters, slang, emojis, and abbreviations. In particular, users on social media, messaging applications, or review platforms tend to write quickly and informally rather than using a formal style. Therefore, the model must be prepared to handle the diverse forms it may encounter in real-world contexts. As shown in Figure 6, the model effectively handled noisy elements across different examples.

```
=== Noisy Samples ===
Sentence: 'Bn bugn işten sonra o kafe'ye gittm, ortam süüppeeerr hoşmuş, bayıldm valla!'
Predicted Sentiment: POSITIVE (73.59%)

Sentence: 'Ay yine rezil oldum ya, sipariş hepsi yanlış gelmiş, kimse ilgilenmiyo, cidden sinir bozucu!!!'
Predicted Sentiment: NEGATIVE (79.06%)

Sentence: 'Kanka bu hafta sonu bulusuyoz mu yoksa iptal mi ???'
Predicted Sentiment: NEUTRAL (95.08%)
```

Figure 6. Testing the Electra model with positive, negative, and neutral noisy samples.

In the noisy positive example in Figure 6, "Bn bugn işten sonra o kafe'ye gittm, ortam süüppeeerr hoşmuş, bayıldm valla!", the model accurately interpreted missing letters and elongated words, detecting a positive sentiment with a confidence score of 73.59%. In the noisy negative example, "Ay yine rezil oldum ya, sipariş hepsi yanlış gelmiş, kimse ilgilenmiyo, cidden sinir bozucu!!!", despite spelling errors, the model produced a negative prediction with a high confidence score of 79.06%. Lastly, in the noisy neutral example, "Kanka bu hafta sonu bulusuyoz mu yoksa iptal mi???", the model classified the statement as neutral with 95.08% confidence, despite the presence of slang and abbreviations.

4.1.2. Testing with Domain-Specific Samples

A model trained on a particular dataset may produce unexpected errors when encountering specialized terminology in domains such as medicine, law, or gastronomy. Consequently, observing how the model responds to texts from different specialized fields is essential for evaluating its generalization capability.

As shown in Figure 7, the model effectively classified domain-specific texts despite the presence of technical terms. In the domain-specific positive example, "Onkoloji depart-

manında uygulanan yenilikçi immünoterapi protokolü, tümör boyutunu beklenenden hızlı küçülttü doktorlar bile bu iyileşmeye hayran kaldı!", which includes specialized terms such as "onkoloji" and "immünoterapi", the model classified the statement as positive with a confidence score of 69.97%. In the domain-specific negative example, "Rekonstrüksiyon ameliyatı sonrası beklenmeyen komplikasyonlar gelişince, hastanın durumu kötüleşti", the model recognized the negative emphasis in a medical context with a confidence score of 76.82%. Lastly, in the domain-specific neutral example, "İstinaf mahkemesi, dosyadaki eksik belgeler nedeniyle davanın yeniden değerlendirilmesine karar verdi.", the model identified the legal term "istinaf" and classified the sentence as neutral with a confidence score of 76.48%.

```
=== Domain-Specific Samples ===
Sentence: 'Onkoloji departmanında uygulanan yenilikçi immünoterapi protokolü, tümör boyutunu beklenenden hızlı küçülttü doktorl
ar bile bu iyileşmeye hayran kaldı!'
Predicted Sentiment: POSITIVE (69.97%)

Sentence: 'Rekonstrüksiyon ameliyatı sonrası beklenmeyen komplikasyonlar gelişince, hastanın durumu kötüleşti'
Predicted Sentiment: NEGATIVE (76.82%)

Sentence: 'İstinaf mahkemesi, dosyadaki eksik belgeler nedeniyle davanın yeniden değerlendirilmesine karar verdi.'
Predicted Sentiment: NEUTRAL (76.48%)
```

Figure 7. Testing the Electra model with positive, negative, and neutral domain-specific samples.

These results suggest that the model performs reasonably well in both noisy environments and domain-specific terminology. In future work, the aim is to improve the model in these respects by increasing data diversity and enhancing its resilience against various linguistic errors.

4.2. Model Interpretability and Explainability

Electra, the Transformer-based model that achieved the best performance in this study, was used to perform sentiment analysis in Turkish. To interpret the model's decision-making processes, Attention Heatmap [63] and SHapley Additive exPlanations (SHAP) [64] Analysis were applied. These methods were chosen to provide a transparent evaluation of the model's classification process by visualizing which words the model focuses on and how these words contribute to the predictions [1,2]. The tools Matplotlib, Seaborn, and SHAP were utilized for visualization and analysis. Below, positive, negative, and neutral sentences are analyzed in detail using both methods.

4.2.1. Attention Heatmap

Attention Heatmap is a method that visualizes the attention mechanism of Transformer-based models, showing which words the model focuses on when making predictions [1]. The attention mechanism generates a weight matrix to determine the importance of each word in the context of the sentence. Words with higher attention weights are considered more relevant to the classification task, while words with lower weights contribute less. Using Seaborn, these attention weights are visualized as heatmaps, where darker colors indicate higher attention, and lighter colors indicate lower attention.

Figure 8 presents the attention heatmap visualization for a positive sentence and shows that the model strongly focuses on the words "kusursuz" and "uyumu", which are the primary indicators of the positive sentiment in the sentence. In contrast, context-providing words such as "evin" and "dekorasyonunda" receive lower attention. This pattern suggests that the model effectively prioritizes meaningful words while correctly processing the contextual structure.

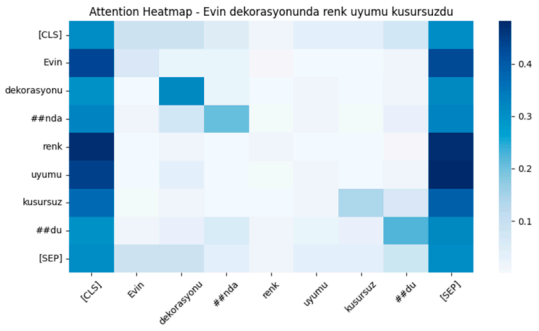

Figure 8. Attention heatmap visualization for a positive sentence.

Figure 9 illustrates the attention heatmap for a negative sentence. The figure reveals that the model assigns high attention weights to the words "bozulan" and "fiyasko", both of which convey strong negative sentiment. On the other hand, words such as "aldıktan" and "sonra", which provide temporal context but do not directly indicate sentiment, receive significantly lower attention. This distribution indicates that the model correctly identifies keywords that define the negative sentiment, ensuring a precise classification.

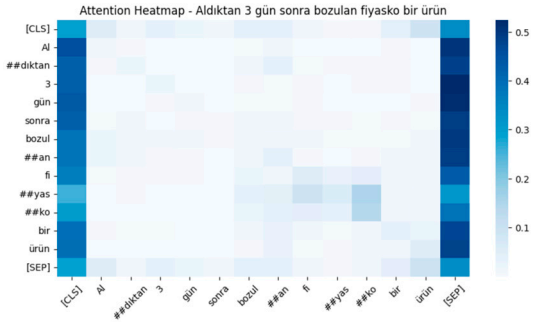

Figure 9. Attention heatmap visualization for a negative sentence.

Figure 10 depicts the attention heatmap for a neutral sentence. The figure demonstrates that the model primarily focuses on the phrases "belirtilen tarih" and "teslim edildi", as these words determine the neutral nature of the sentence. Meanwhile, lower emotional intensity words like "aralığında" and "adresime" receive less attention, aligning with the expected distribution for a neutral statement. This suggests that the model successfully applies attention weights to classify neutral expressions accurately.

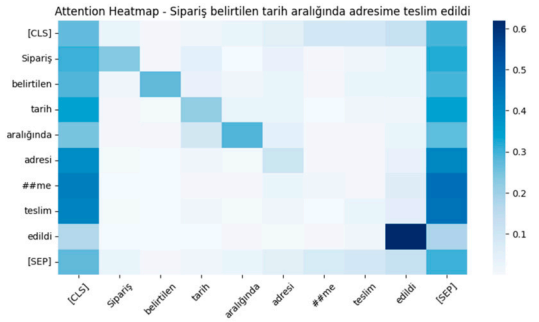

Figure 10. Attention heatmap visualization for a neutral sentence.

4.2.2. SHAP Analysis

SHAP (SHapley Additive exPlanations) analysis is a method that explains the contribution of each word in a sentence to the model's classification output by assigning numerical values and color-coded visualizations [2]. This method is based on Shapley values and quantifies how much each feature (word) contributes to the final prediction. The base

value represents the model's general tendency when no specific input is provided, serving as a reference point for classification. The f(output) value indicates the final prediction computed for a given input, incorporating the influence of individual words. The difference between base value and f(output) reflects the cumulative effect of the words in the input on the final classification. In SHAP visualizations, words contributing to a positive sentiment classification are highlighted in red/pink tones, whereas words pushing the classification towards a negative sentiment are shown in blue tones. The length of each bar represents the magnitude of a word's impact on the prediction, while the direction indicates whether it positively or negatively influences the classification. In this study, output 0 represents negative sentiment, output 1 represents neutral sentiment, and output 2 corresponds to positive sentiment.

In Figure 11, in the positive sentence "Evin dekorasyonunda renk uyumu kusursuzdu", the f(output) value (−0.605057) indicates that the model's prediction is directed toward the positive class. The word "kusursuz" makes a strong contribution to the positive sentiment, as indicated by the red tone and the length of the bar. Similarly, "uyumu" also provides a meaningful contribution. The difference between the base value (−1.74495) and f(output) demonstrates how the input words collectively influenced the model to classify the sentence as positive. Also, in the negative sentence "Aldıktan 3 gün sonra bozulan fiyasko bir ürün", the f(output) value (1.56536) suggests that the model has classified the sentence as negative. The word "bozulan" is the most influential in determining the negative classification, as shown by the blue color and the significant bar length. Additionally, "fiyasko" reinforces the negative sentiment prediction. The difference between the base value (−1.78928) and f(output) highlights the collective impact of these words in shifting the classification toward the negative class. Finally, in the neutral sentence "Sipariş belirtilen tarih aralığında adresime teslim edildi", the f(output) value (−1.51784) suggests that the model's classification remains close to neutral. The phrase "teslim edildi" contributes the most to the neutral classification, as indicated by the bar length. Additionally, "belirtilen tarih" plays a supporting role in maintaining the neutral prediction. The difference between the base value and f(output) quantifies the extent to which the words influence the sentence's classification toward neutrality.

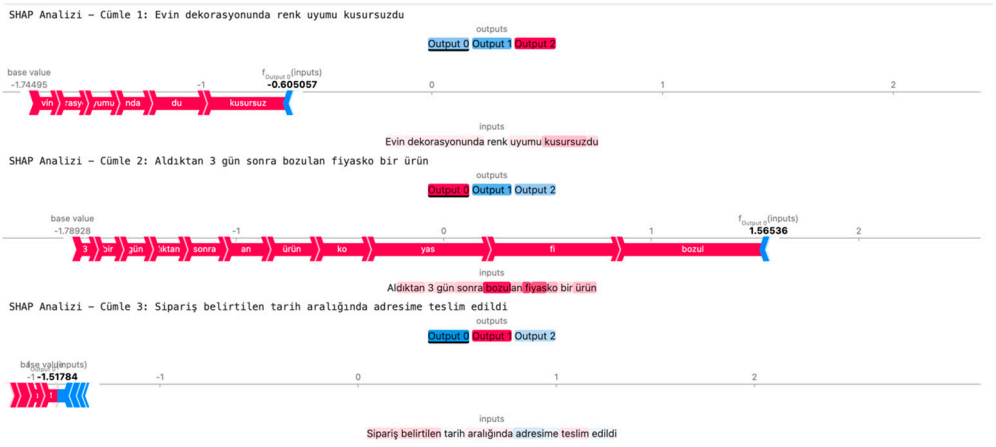

Figure 11. SHAP analysis visualization for positive, negative, and neutral sentences.

5. Conclusions

This study underscores the critical role of sentiment analysis in extracting actionable insights from text data, particularly for the Turkish language. While sentiment analysis has achieved remarkable progress for widely spoken languages like English, Turkish presents unique challenges due to its agglutinative morphology, flexible syntax, and the scarcity of large-scale labeled datasets. Addressing these obstacles is vital to developing accurate and

robust sentiment analysis tools for Turkish-speaking regions. To tackle these challenges, we introduced a balanced and extensive dataset for Turkish sentiment analysis, comprising 2333 positive, 2334 neutral, and 2333 negative sentences. To enhance diversity and maintain balance, an expanded dataset with 15,853 examples was created using data augmentation techniques such as back-translation and synonym replacement. These datasets provide a valuable foundation for training and evaluating sentiment analysis models in Turkish.

Using the proposed datasets, we trained and evaluated a comprehensive range of machine learning and state-of-the-art deep learning models. These included advanced architectures such as Electra Base Turkish, BERT Base Turkish, TurkishBERTweet, BERT Base Multilanguage, Turkish RoBERTa Base, and XLM-RoBERTa Base, alongside traditional methods like Linear SVC, Logistic Regression, Naive Bayes, KNN, and Decision Tree. This rigorous evaluation offered insights into the relative performance of various models, highlighting their potential for Turkish sentiment analysis.

The improvements in our dataset and model innovations have significant practical implications in various real-world applications. For instance, in customer service automation, our sentiment analysis model can enhance chatbot interactions by accurately understanding customer emotions, leading to more effective and empathetic responses. Similarly, in social media monitoring, the refined sentiment analysis capabilities can help businesses and policymakers track public sentiment more accurately, enabling better decision-making and crisis management. Furthermore, e-commerce platforms can leverage our model to analyze product reviews more effectively, improving recommendation systems and customer experience. These examples highlight the practical benefits of our study in addressing real-world challenges where sentiment analysis in the Turkish language plays a crucial role.

Author Contributions: Conceptualization, K.B.Z. and S.S.; methodology, S.S. and K.B.Z.; software, K.B.Z. and B.S.K.; validation, K.B.Z., S.Z.D. and B.S.K.; formal analysis, K.B.Z.; resources, K.B.Z., S.Z.D. and B.S.K.; data curation, K.B.Z., S.Z.D. and B.S.K.; writing—original draft preparation, K.B.Z., S.Z.D. and B.S.K.; writing—review and editing, S.S.; visualization, K.B.Z., S.Z.D. and B.S.K.; supervision, S.S.; project administration, S.S.; funding acquisition, S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original datasets presented and used in the study are openly available in the FSMTSA GitHub repository at (https://github.com/kevserbusrayildirim/FSMTSAD, accessed on 15 January 2025).

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

Abbreviations

The following abbreviations are used in this manuscript:

NLP	Natural Language Processing
FSMTSAD	Fatih Sultan Mehmet Turkish Sentiment Analysis Dataset
Electra	Efficiently Learning an Encoder that Classifies Token Replacements Accurately
BERT	Bidirectional Encoder Representations
RoBERTa	A Robustly Optimized BERT Pre-training Approach
SVC	Support Vector Classifier
KNN	K-Nearest Neighbors
XGBoost	eXtreme Gradient Boosting
MLP	Multi-Layer Perceptron

LSTM	Long Short-Term Memory
BiLSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
GRU	Gated Recurrent Unit
LLMs	Large Language Models
LLaMA	Large Language Model Meta AI
TF	Term Frequency
IDF	Inverse Document Frequency

References

1. Agarwal, B.; Mittal, N.; Bansal, P.; Garg, S. Sentiment analysis using common-sense and context information. *Comput. Intell. Neurosci.* **2015**, *2015*, 715730. [CrossRef]
2. Tuzcu, S. Çevrimiçi kullanıcı yorumlarının duygu analizi ile sınıflandırılması. *Eskişehir Türk Dünyası Uygul. Ve Araştırma Merk. Bilişim Derg.* **2020**, *1*, 1–5.
3. Agarwal, B.; Namita, M. Optimal feature selection for sentiment analysis. In Proceedings of the Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, 24–30 March 2013; Proceedings, Part II 14. Springer: Berlin/Heidelberg, Germany, 2013.
4. Oflazer, K.; Kuruöz, Ì. Tagging and morphological disambiguation of Turkish text. In Proceedings of the Fourth Conference, Stuttgart, Germany, 13–15 October 1994; pp. 144–149.
5. Torunoğlu-Selamet, D.; Gülşen, E. A cascaded approach for social media text normalization of Turkish. In Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM), Gothenburg, Sweden, 26–30 April 2014.
6. Göksel, A.; Celia, K. *Turkish: A Comprehensive Grammar*; Routledge: Abingdon, UK, 2004.
7. Dbmdz. *Electra-Base-Turkish-Cased-Discriminator* [Model]. Hugging Face. 2020. Available online: https://huggingface.co/dbmdz/electra-base-turkish-cased-discriminator (accessed on 13 January 2025).
8. Schweter, S. BERTurk-BERT models for Turkish. *Zenodo* **2020**, *2020*, 3770924.
9. Najafi, A.; Varol, O. TurkishBERTweet: Fast and reliable large language model for social media analysis. *Expert Syst. Appl.* **2024**, *255*, 124737. [CrossRef]
10. Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
11. Tas, N. RoBERTurk: Adjusting RoBERTa for Turkish. *arXiv* **2024**, arXiv:2401.03515.
12. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 8440–8451.
13. Hsu, C.-W. *A Practical Guide to Support Vector Classification*; Department of Computer Science, National Taiwan University: Taipei, Taiwan, 2003.
14. Kleinbaum, D.G.; Mitchel, K. *Logistic Regression*; Springer: New York, NY, USA, 2002.
15. Rish, I. An empirical study of the naive Bayes classifier. In Proceedings of the IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, Seattle, DC, USA, 4–6 August 2001.
16. Cover, T.; Peter, H. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [CrossRef]
17. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
18. Cortes, C. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
19. Hochreiter, S. *Long Short-Term Memory*; Neural Computation MIT-Press: Cambridge, MA, USA, 1997.
20. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
21. Taboada, M.; Julian, B.; Milan, T.; Kimberly, V.; Manfred, S. Lexicon-based methods for sentiment analysis. *Comput. Linguist.* **2011**, *37*, 267–307. [CrossRef]
22. Pang, B.; Lillian, L. Opinion mining and sentiment analysis. *Found. Trends.* **2008**, *2*, 1–135.
23. Paredes-Valverde, M.A.; Colomo-Palacios, R.; Salas-Zárate, M.D.P.; Valencia-García, R. Sentiment analysis in Spanish for improvement of products and services: A deep learning approach. *Sci. Program.* **2017**, *2017*, 1329281. [CrossRef]
24. Cam, H.; Cam, A.V.; Demirel, U.; Ahmed, S. Sentiment analysis of financial Twitter posts on Twitter with the machine learning classifiers. *Heliyon* **2024**, *10*, e23784. [CrossRef] [PubMed]
25. Chen, T. *Xgboost: Extreme Gradient Boosting, R Package Version 0.4–2 1.4*; The R Foundation for Statistical Computing: Vienna, Austria, 2015.
26. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [CrossRef]

27. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
28. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
29. Jeffrey, L.E. Finding structure in time. *Cogn. Sci.* **1990**, *14*, 179–211.
30. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* **2005**, *18*, 602–610. [CrossRef]
31. Floridi, L.; Massimo, C. GPT-3: Its nature, scope, limits, and consequences. *Minds Mach.* **2020**, *30*, 681–694. [CrossRef]
32. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. Llama: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971.
33. Beleveslis, D.; Tjortjis, C.; Psaradelis, D.; Nikoglou, D. A hybrid method for sentiment analysis of election related tweets. In Proceedings of the 2019 4th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM), Piraeus, Greece, 20–22 September 2019; pp. 1–6.
34. Kına, E.; Emre, B. Tweetlerin Duygu Analizi İçin Hibrit Bir Yaklaşım. *Doğu Fen Bilim. Derg.* **2023**, *6*, 57–68. [CrossRef]
35. Aydoğan, M.; Abdullah, Ş. Duygu Analizi Tabanli Yeni Bir Hibrit Tavsiyeci Sistem. *Euroasia J. Math. Eng. Nat. Med. Sci.* **2020**, *7*, 48–62.
36. Sebastiani, F.; Andrea, E. Sentiwordnet: A publicly available lexical resource for opinion mining. In Proceedings of the 5th International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA), Genoa, Italy, 20–25 May 2006.
37. Mladenovic, D.; Antonijevic, M.; Jovanovic, L.; Simic, V.; Zivkovic, M.; Bacanin, N.; Zivkovic, T.; Perisic, J. Sentiment classification for insider threat identification using metaheuristic optimized machine learning classifiers. *Sci. Rep.* **2024**, *14*, 1–39. [CrossRef]
38. Danyal, M.M.; Khan, S.S.; Khan, M.; Ullah, S.; Ghaffar, M.B.; Khan, W. Sentiment analysis of movie reviews based on NB approaches using TF–IDF and count vectorizer. *Soc. Netw. Anal. Min.* **2024**, *14*, 25731. [CrossRef]
39. Alizadeh, M.; Seilsepour, A. A novel self-supervised sentiment classification approach using semantic labeling based on contextual embeddings. *Multimed. Tools Appl.* **2024**, 1–26. [CrossRef]
40. Abimbola, B.; Tan, Q.; Marín, E.A.D.L.C. Sentiment analysis of Canadian maritime case law: A sentiment case law and deep learning approach. *Int. J. Inf. Technol.* **2024**, *16*, 3401–3409. [CrossRef]
41. Alawi, A.B.; Ferhat, B. A hybrid machine learning model for sentiment analysis and satisfaction assessment with Turkish universities using Twitter data. *Decis. Anal. J.* **2024**, *11*, 100473. [CrossRef]
42. Jahin, M.A.; Shovon, M.S.H.; Mridha, M.F.; Islam, M.R.; Watanobe, Y. A hybrid transformer and attention based recurrent neural network for robust and interpretable sentiment analysis of tweets. *Sci. Rep.* **2024**, *14*, 24882. [CrossRef]
43. Vijayaragavan, P.; Suresh, C.; Maheshwari, A.; Vijayalakshmi, K.; Narayanamoorthi, R.; Gono, M.; Novak, T. Sustainable sentiment analysis on E-commerce platforms using a weighted parallel hybrid deep learning approach for smart cities applications. *Sci. Rep.* **2024**, *14*, 26508. [CrossRef]
44. Miah, S.U.; Kabir, M.; Bin Sarwar, T.; Safran, M.; Alfarhood, S.; Mridha, M.F. A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. *Sci. Rep.* **2024**, *14*, 9603. [CrossRef] [PubMed]
45. Aliyeva, Ç.O.; Yağanoğlu, M. Deep learning approach to detect cyberbullying on twitter. *Multimed. Tools Appl.* **2024**, 1–24. [CrossRef]
46. Marutho, D.; Supriadi, R. Optimizing aspect-based sentiment analysis using sentence embedding transformer, bayesian search clustering, and sparse attention mechanism. *J. Open Innov. Technol. Mark. Complex.* **2024**, *10*, 100211. [CrossRef]
47. Hellwig, N.C.; Fehle, J.; Wolff, C. Exploring large language models for the generation of synthetic training samples for aspect-based sentiment analysis in low resource settings. *Expert Syst. Appl.* **2024**, *261*, 125514. [CrossRef]
48. Aziz, K.; Ji, D.; Chakrabarti, P.; Chakrabarti, T.; Iqbal, M.S.; Abbasi, R. Unifying aspect-based sentiment analysis BERT and multi-layered graph convolutional networks for comprehensive sentiment dissection. *Sci. Rep.* **2024**, *14*, 14646. [CrossRef] [PubMed]
49. Jiang, X.; Ren, B.; Wu, Q.; Wang, W.; Li, H. DCASAM: Advancing aspect-based sentiment analysis through a deep context-aware sentiment analysis model. *Complex Intell. Syst.* **2024**, *10*, 7907–7926. [CrossRef]
50. Feng, A.; Cai, J.; Gao, Z.; Li, X. Aspect-level sentiment classification with fused local and global context. *J. Big Data* **2023**, *10*, 176. [CrossRef]
51. Punetha, N.; Jain, G. Game theory and MCDM-based unsupervised sentiment analysis of restaurant reviews. *Appl. Intell.* **2023**, *53*, 20152–20173. [CrossRef]
52. Zuheros, C.; Martínez-Cámara, E.; Herrera-Viedma, E.; Herrera, F. Sentiment analysis based multi-person multi-criteria decision making methodology using natural language processing and deep. learning for smarter decision aid. Case study of restaurant choice using TripAdvisor reviews. *Inf. Fusion* **2021**, *68*, 22–36. [CrossRef]
53. Punetha, N.; Jain, G. Advancing sentiment classification through a population game model approach. *Sci. Rep.* **2024**, *14*, 20540. [CrossRef]

54. Phan, T.M. Sentiment-semantic word vectors: A new method to estimate management sentiment. *Swiss J. Econ. Stat.* **2024**, *160*, 9. [CrossRef]
55. Thakor, P.; Sasi, S. Ontology-based sentiment analysis process for social media content. *Procedia Comput. Sci.* **2015**, *53*, 199–207. [CrossRef]
56. Zhang, W.; Deng, Y.; Liu, B.; Pan, S.; Bing, L. Sentiment Analysis in the Era of Large Language Models: A Reality Check. In *Findings of the Association for Computational Linguistics: NAACL 2024*; Association for Computational Linguistics: Mexico City, Mexico, 2024; pp. 3881–3906.
57. Karabıyık, M.A.; Asım, S.Y.; Fatma, G.T. Yapay Zekâ Çağında Duygu Analizi: Büyük Dil Modellerinin Yükselişi ve Klasik Yaklaşımlarla Karşılaştırılması. *Afyon Kocatepe Üniversitesi Fen Ve Mühendislik Bilim. Derg.* **2024**, *24*, 1355–1363. [CrossRef]
58. Bhattarai, P.; Sharma, S.; Acharya, B. A Comprehensive Study of Sentiment Analysis Using Deep Learning Approaches. *arXiv* **2023**, arXiv:2305.15083.
59. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
60. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodel, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
61. 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.* **2022**, *35*, 11019–11038. [CrossRef]
62. Köksal, A.; Özgür, A. Twitter dataset and evaluation of transformers for Turkish sentiment analysis. In Proceedings of the 2021 29th Signal Processing and Communications Applications Conference (SIU), Istanbul, Turkey, 9–11 June 2021; pp. 1–4.
63. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762v7. [CrossRef]
64. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *arXiv* **2017**, arXiv:1705.07874. [CrossRef]