From Data to Cluster: Enhancing Second-Hand Car Valuation with Descriptions

Emine Esin Yılmaz Onur Çalışkan Resul Erdem Arduç Tuğba Dalyan
Department of Computer Engineering, Istanbul Bilgi University,
Eski Silahtarağa Elektrik Santralı Kazım Karabekir Cad. No: 2/13 34060 Eyüpsultan, İstanbul

June 12, 2024

Abstract Second-hand car sites have recently been frequently preferred by people who want to buy a car. This study aims to investigate the effect of the description texts of each advertisement on second-hand vehicle sites on the prices in the relevant advertisements and to cluster the cars according to their quality by blending the ready-made features in the advertisements with the features extracted from the text.

1 Introduction

There has been a noteworthy trend in recent years to prioritize personal transportation, which has resulted in a considerable growth in the volume of vehicle transactions. Experts in the field [1] predict that e-commerce will continue to expand in the future at a consistent rate. A significant portion of this growth in volume, both domestically and internationally, can be attributable to pre-owned car sales. The popularity of online markets for the trade of second-hand automobiles has been substantially aided by the spread of the internet. However, despite its enormous size, this online marketplace frequently necessitates hours of work for both sellers and potential buyers to sift through an excessive amount of adverts.

From customers' perspective, the decision process requires a closer look at factors beyond brand, model, and mileage, even though they may already have ideas about the characteristics they want in a car. Examining the descriptive texts plays significant role thorough analysis in determining the vehicle's value. These descriptions usually include comprehensive details about the vehicle's history, such as previous repairs, owner usage patterns, and maintenance records. As a result of the necessity for a thorough analysis and comparison of all these intricate factors, clients may encounter a great deal of confusion and time commitment.

In this study, we aim to simplify the procedures for customers interested in buying second-hand cars while also promoting better decision-making. The project's initial phase will involve aggregating data from specific websites for used car sales, such as arabam.com, otokocikinciel.com, and vavacars.com. The data includes descriptive text that providing in-depth insights about the car in addition to standard vehicle specifications like model, year and mileage. Then, the proposed Natural Language Processing

(NLP) models extract both concrete characteristics as well as the opinions undertones from the descriptive texts. Then, the effect of the newly added features on the prices in the advertisements will be investigated through the description texts. Finally, an artificial intelligence model will be used to cluster cars using newly added features. The main contributions are:

- Extracting **useful keywords from text of the advertisement** and incorporating those deemed appropriate as features into dataset
- Detecting the overall **sentiment of the advertisement** using sentiment analysis, classify it as positive or negative, and provide score metrics.
- Investigating the **effect of features extracted from description texts** on price prediction
- Distinguishing cars with **data obtained through text analysis**

2 Related Works

Nowadays, second-hand car buyers frequently use web pages to reach car advertisements. These advertisements generally include the features of the car such as year, engine, color, and more. Additionally, some web pages contain detailed descriptions of the vehicle and information about its location. These are frequently more important than the vehicle's fundamental characteristics. However, these descriptive features are often disregarded in academic studies [2] investigating recommendation algorithms for second-hand car and real estate sales. The users have to read and evaluate the descriptions of each advertisement separately, users find it difficult to make effective choices as a result. Their decision-making process may be less effective as a result of this costly and difficult procedure.

In recent years, there are various studies to address the problem of estimation of vehicle for users. In the study by Kumar et al. (2020) [3], only the structural features of the cars, such as mileage, model year, fuel type, gear type, and horsepower, were extracted from the websites. This strategy did not take into account the descriptive text that was included with the ads, which might have left out important background information that could improve the relevance and accuracy of recommendation systems for the sale of second-hand cars.

Fafalios et al. (2022) [4] used machine learning models to solve a regression problem for vehicles with 0 kilometers on them, however they did not include the descriptive texts. Similarly, the author of the study by [5] used only vehicle information to predict prices, excluding text data. On the other hand, a study [6] proposed a recommendation system that relied on text processing techniques. The study [7] provides another illustration of text processing. The authors conducted sentiment analysis on texts about automobiles on Twitter and reached positive-neutral-negative assessments on automobiles through text processing techniques.

In this article [8], the NER model is used for Turkish texts. During the data training process, texts were selected from different fields such as unofficial texts written on the internet, customer comments, e-commerce, and movie reviews. It was made with pre-trained language models in the spaCy library. The paper also mentions that pre-trained models perform better than other NLP libraries. The results of the model are as in the table below 1.

In this article [9], the FBPSO model, one of the UFS methods, is introduced. In this article, 6 different data sets were studied and compared with MCFS, TRACK UFSACO, and UFSPSO models along with FBPSO. It has been stated that FBPSO gives better results than other models.

As we know, the descriptive texts of the cars and

Metrics	Values
NER Precision	0.914
NER Recall	0.913
NER F Score	0.913
TAG (XPOS) Accuracy	0.917
POS (UPOS) Accuracy	0.909

Table 1: Metrics and Values for [8]

the several features obtained from these texts are not included in the studies. The main contribution of our study is to use these text about second-hand car. Our main purpose to show that these descriptive texts have a significant weight in the machine-learning model planned to be created. The recommendation system will be created using named entity recognition and sentiment analysis methods, as well as the structural features of the tools. It is expected that the recommendations of the model trained with this hybrid method will be more efficient and consistent in the future. Within the scope of this project, it is aimed to integrate these explanation texts, which are critical for users who want to buy a second-hand vehicle, into the user suggestion algorithm. In this way, it will be ensured that the buyer can choose the most suitable advertisement without being exposed to information pollution and loss of time about the vehicle.

3 Methodology

The proposed methodology is roughly designed as collecting data at 3.1, and extracting description text features at 3.2 and 3.3. The effect of the features extracted from the texts on the price is presented in 3.4, and finally, clustering car ads with new features is presented in 3.5.

3.1 Dataset Preparation

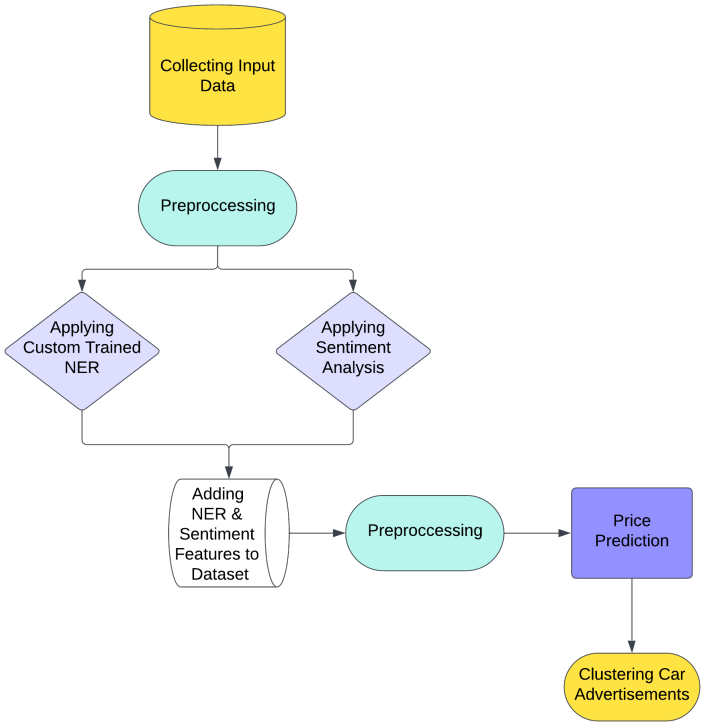

Figure 1: Methodology Diagram of the Project

Car features and descriptions will be performed by scraping data from websites using the Python language. The dataset has started to be created using the following libraries: BeautifulSoup, Sqlite3.

By sending HTTP requests to libraries and second-hand car sites [10] [11], the necessary description texts from front-end classes have begun to be collected. There is a similar study in [12], but the authors did not use the description texts to train the model. After all the data is collected, the main data set will be ready after processes such as data pre-processing methods and text processing, that is, extracting features from the text, and sentiment anal-

ysis, are completed to obtain better results from the model.

- Keyword and feature extraction have started to be performed with the NER model [13] [14].
- Sentiment analysis will be performed to classify the statements as positive-negative.

During the project, it is aimed to increase the data set, especially to test transformer-based NER models and to create a new model to obtain a more efficient output value. The development stages and general planning of the project after the preparation of the data set are stated in the following section.

3.2 Extracting Features by Using NER Models

As it was stated in the dataset preparation step, further analysis of the description texts were acquired by using NER models.

- In the first trial, Turkish pre-trained NER Model [15] were used in order to examine the Turkish NER Model's performance on the acquired dataset.
- After labeling the entities in the description texts using a pre-trained model, it is observed that the predefined labels were not sufficient for the further analysis of our dataset, concerning the specific nature of the dataset's domain.
- In order to customize the labels that will be detecting throughout the description text data, Custom Named Entity Recognition Model was trained using spacy format.

3.2.1 Training the Custom NER Model

To acquire more specific entity recognition throughout the text data, a Custom NER Model was trained.

- First, training data was annotated using an annotator tool.
- The annotated data was then transformed into a JSON format for further use in the training step.
- After the annotation of around 500 description texts, the JSON formatted data was transformed, and the Custom Model was trained using these 500 instances, learning the patterns for detecting further labels and entities.

After the model was obtained, other data instances were passed down to the model.

3.3 Sentiment Analysis

The positive-negative overall tone classification of seller description texts has a significant impact on the evaluation of the ad for the user. Therefore, it is planned to perform sentiment analysis on the texts. At this stage, studies on different models was carried out using the digitized and weighted variables from the previous stage, 3.4.1.

Pre-trained NER models and pre-trained sentiment analysis models on Turkish language models are also available. These pre-trained models can be fine-tuned to suit the project's dataset and the models can be used on the data again.

3.4 The Effect of Text Features on Price Prediction

Determining vehicle prices in the second-hand car market is a very complex process with the interaction of many parameters. For this reason, Random Forest Regressor, XGBoost Regressor and LGBM Regressor that are the supervised learning methods, was used to accurately predict vehicle prices.

- Regression models estimate the linear relationship between the independent variables and the dependent variable. These models produce clear and understandable results. Each coefficient in the model directly reflects the impact of the relevant feature on price.

- Regression models are relatively fast and computationally efficient. It provides highly effective results even when processing large data sets.

In this study, the main purpose of using price prediction is to demonstrate the positive impact of the new features created from the NER and Sentiment analysis results applied to the description texts of the advertisements in previous stages, firstly performed price prediction using a dataset that did not include the NER and sentiment scores and calculated the accuracy. Then, performed price prediction using a new dataset that included the features derived from the NER analysis and the Sentiment scores, and compared the consistency between the two predictions.

3.5 Clustering Advertisements

Clustering method was used to determine and group the most suitable advertisements. The choice of this method was made for the following reasons:

- Grouping second-hand car ads with similar features helps users more easily find ads that meet their needs and expectations. Clustering separates advertisements according to similar characteristics and ensures that the best advertisements are highlighted within these groups.

- Clustering algorithms present large and complex data sets into simpler and more meaningful groups. This allows the data to be visualized and understood more easily by users.

Second-hand car ads are multidimensional data (km, model, price, year, sentiment analysis result). Clustering algorithms such as K-means analyze this multidimensional data and enable the creation of groups with similar characteristics.

3.5.1 Customized and Meaningful Suggestions

When clustering algorithms are used to group vehicles with similar characteristics, more specific and engaging recommendations can be made to users. For example, grouping vehicles in a certain price range or a specific make and model in the same cluster highlights vehicles that may be of interest to the user. This, in turn, can increase user satisfaction [16].

3.5.2 Scalability and Efficiency

Clustering algorithms can work efficiently on large data sets and provide scalability. This means the ability to provide quick and effective recommendations even in a large database of listings [17]. This is a significant advantage, especially when working with large volumes of data.

3.5.3 K-Means Algorithm

In areas with large data sets, such as car listings, offering suggestions based on users' interests is critical to increasing user satisfaction and accelerating sales. In this context, clustering techniques such as the k-means algorithm stand out as an effective tool to provide more relevant suggestions to users by grouping ads with similar features.

3.5.4 Fundamentals of K-Means Algorithm

K-means algorithm is a clustering algorithm that aims to divide the data set into k clusters or groups.

Each cluster consists of data points with similar characteristics, and each data point is assigned to a cluster closest to a reference point called the centroid. The algorithm initially starts with k randomly selected centroids and reconstructs the clusters by iteratively updating the centroids. This process continues until the position of the center points becomes stable.

In the process of recommending car ads, using the k-means algorithm provides several advantages:

Car advertisements have a wide variety of features such as model, brand,price,mileage, fuel type. By analyzing this multidimensional dataset, the K-means algorithm can group vehicles with similar characteristics.Thus, it becomes possible to recommend tools that best suit users' search criteria.

3.5.5 Application Process

Data Preparation: In the first step, the characteristics of car advertisements are determined and appropriately standardized. Numerical data such as price and mileage are normalized; Categorical data such as brand and model are converted into numerical form using methods such as categorical encoding. Same normalization methods used on Price Prediction step.

Determining the Optimum K Value: The performance of the K-means algorithm depends on the number of k clusters. Silhouette score was used as basis to determine the optimum k value.

After all these steps clustering algorithm will be assign all advertisements to their clusters.

#	Column	Non-Null Count	Dtype
0	Fiyat	10224 non-null	int64
1	İlan No	10224 non-null	int64
2	Marka	10224 non-null	object
3	Seri	10224 non-null	object
4	Model	10224 non-null	object
5	Yıl	10224 non-null	int64
6	Km	10224 non-null	int64
7	Vites Tipi	10224 non-null	object
8	Yakıt Tipi	10224 non-null	object
9	Kasa Tipi	10224 non-null	object
10	Renk	10224 non-null	object
11	Motor Hacmi	10224 non-null	int64
12	Motor Gücü	10224 non-null	int64
13	Çekiş	10224 non-null	object
14	Ort. Yakıt Tüketimi	10224 non-null	float64
15	Yakıt Deposu	10224 non-null	int64
16	Boya-değişen	10224 non-null	object
17	Değişen	10224 non-null	object
18	Boya	10224 non-null	object
19	Kimden	10224 non-null	object
20	Açıklama	10224 non-null	object
21	Processed Açıklama Metni	10224 non-null	object

Figure 2: Dataset Info

4 Dataset

At the beginning of the project, different second-hand car sales sites were examined to obtain a data set. After reviewing the sites, it was decided to use [10] to provide English advertisements, and [11] to provide Turkish data. During the research process, observing that there were fewer studies conducted in Turkish and that studies by examining and adding Turkish texts would be more beneficial, the project began to be developed on the Turkish data set and Turkish models.

To obtain data from [11], web scraping was done by using BeautifulSoup library.

In addition to features such as the vehicle's price, mileage, and engine power, the description texts written by the advertiser were also obtained from [11]. The obtained data is converted to an

excel file with columns as 2

While preparing the data, the scraping was done a couple of times by every team member to mix the advertisements. After the scrapings, the redundant rows were removed and the dataset have been finalized around 10 thousand rows.

After the dataset was obtained, preprocessing steps began.

- All the text was converted to lowercase. Then punctuation marks were removed and replaced with spaces. As a result, more than one space appeared between some words. Since this situation was not desired, sections with more than one space were converted into single spaces.
- Since the data was captured using the web scraping method, HTML tags were removed to avoid being affected by possible HTML tags. And the maintaining emojis and symbols were removed.
- NaN values were examined and filled using methods such as mean, median and mode imputations were tried and the the decision was made to use mode imputation considering the similar car brand, model,and series.
- In the first sample of the dataset, there were features obtained together as one column. For example ”Boya-degisen”. In order to get better results and assumptions these combined columns were separated into two or three columns. Such as ”boya”, ”degisen”.
- Features such as ”Motor Gücü”,”Yakıt Deposu”, ”Yakıt Tüketimi”, ”Yakıt Hacmi”,”Fiyat”, ”Km” etc. were extracted from their currencies and finalized containing only numbers.
- A sample tokenization process was done as a separate file to examine the word frequencies and will be continued to develop in order to improve model performance.

	Fiyat	Ilan No	Marka	Seri	Model	Yıl	Km	Vites Tipi	Yakıt Tipi	Kasa Tipi	...	Motor Gücü	Çekiş	Ort. Yakıt Tüketimi	Yakıt Deposu	Boya-değişen	Değişen	Boya	Kimden	Açıklama	Processed Açıklama Metni
0	285000	24109801	Opel	Vectra	2.0 GLS	1996	256000	Düz	LPG & Benzin	Sedan	...	136	Önden Çekiş	8.7	61	Belirtilmemiş	Belirtilmemiş	Belirtilmemiş	Sahibinden	Açıklama 0 5 3 4 7 3 0 3 5 6 6...	0 5 3 4 7 3 0 3 5 6 6 orjinal sevenden orjinal...
1	515000	24108567	Toyota	Corolla	1.6 Comfort	2007	173100	Düz	Benzin	Sedan	...	124	Önden Çekiş	6.9	55	5 boyalı	Yok	5 boyalı	Sahibinden	Açıklama 0 ARAÇ ALICAGIM İÇİN SATIYORUM FİYATI...	0 arac alicagim icin satiyorum fiyatini o yuzd...
2	800000	24150269	Toyota	Corolla	1.6 Touch	2018	100000	Otomatik	Benzin	Sedan	...	132	Önden Çekiş	6.1	55	1 değişen	1 değişen	Yok	Sahibinden	Açıklama 0 araç satın alacağım için satılık. D...	0 arac satın alacagım icin satılık detaylı bil...
3	357500	24116791	Ford	Fiesta	1.4 TDCi Comfort	2004	234500	Düz	Dizel	Hatchback/5	...	68	Önden Çekiş	4.3	45	Tamamı orjinal	Yok	Yok	Sahibinden	Açıklama 01 aralık 2023 tarihinde taşıt muayen...	01 aralık 2023 tarihinde tasıt muayene ve egzo...
4	265000	24115171	Tofaş	Kartal	1.6 ie	2001	172000	Düz	LPG & Benzin	Station wagon	...	83	Arkadan İtiş	6.5	52	3 boyalı	Yok	3 boyalı	Sahibinden	Açıklama 02.04.2024 Muanesine var Kaportada bi...	02 04 2024 muanesine var kaportada bi kac sıkı...

Figure 3: Information Table of the Dataset

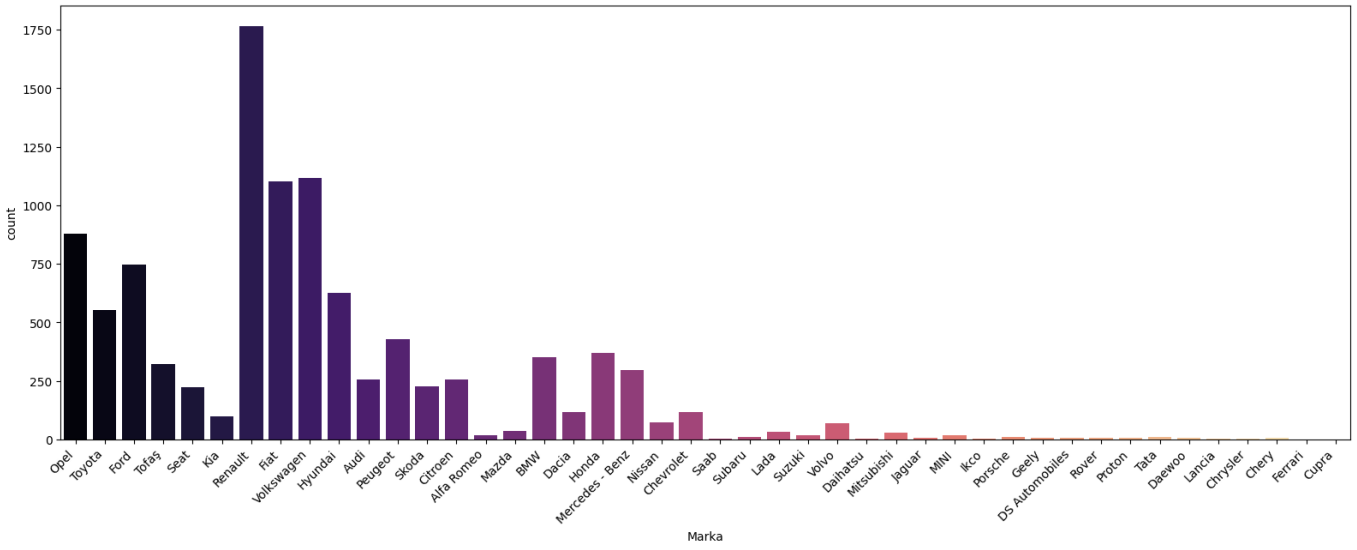

Figure 4: Distribution of Brands in the Dataset

5 Results & Discussions

After the dataset had been obtained and organized by using the pre-processing techniques mentioned, the project's main focus shifted to researching relevant models and techniques to further process the description texts in detail. As the project aims to provide analyses and investigate for additional features that can be extracted, the contributions were made to these parts before providing data for the models to be used in the project.

5.1 Analyses with NER Model

Ensuring effective communication between buyers and sellers in the second-hand car market and supporting transparent and reliable trade is a critical issue. In this context, the description texts presented by sellers about their vehicles contain valuable information for potential buyers.

Effectively extracting and making sense of this information can help buyers make the right decisions. At this point, the use of the NER model aims to increase information access by automatically identifying important entity information in the description texts.

5.1.1 NER Model and Second Hand Car Advertisements

In used car listings, sellers often provide detailed information about their vehicles in the description texts. However, these texts are often information clutter, which can make it difficult for recipients to quickly extract the information they want. At this point the NER model comes into play; By identifying important asset information (brand, model, year, mileage, etc.) within the text, it allows buyers to quickly and accurately focus on the information they are interested in.

5.1.2 Results of NER Analysis

The Spacy library in Python was used to obtain important information from the texts. Thanks to this library, features such as good features, bad features, and tramer records were extracted from the text.

In ready-made pre-trained NER models, features such as location, numbers and people were obtained, and it was determined that these features were not the necessary information for models to be used in the project. For this reason, the NER model was created from scratch. The new NER model created managed to extract the features of the cars such as good features, bad features and tramer records in the text written by the seller.

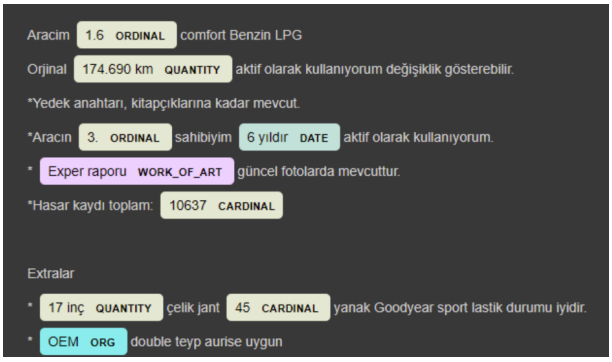

Figure 5: Results of Pre Trained NER Model

While training the model, the training data was created by randomly selecting 500 texts in the advertisements. It was easily organized with a public annotator platform [18] for labeling features in selected texts. After editing, training data was prepared in JSON data format. In addition, some of the features which is the result of NER application, come as string format has transformed into numeric format

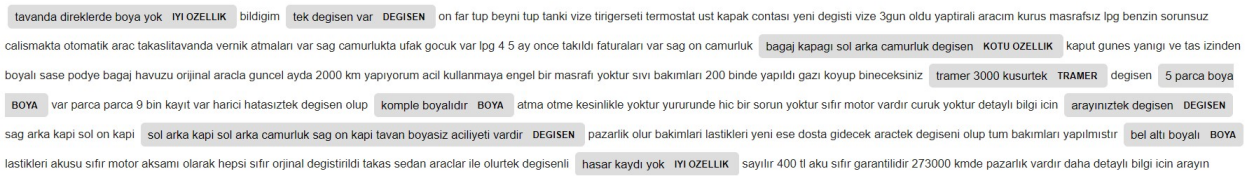

Figure 6: Example Result of Entity Labeling with Custom NER

with regular expressions and conditions. For example '4 parca boyalı' transformed into '4'. The NER model created for this project was applied on a sample advertisement text. According to this text, the revealed features of the car in the advertisement are visualized at Figure 6.
After the final steps; 'sentiment', 'sent-scores', 'IYI OZELLIK NUMERIC (numeric good features)', 'KOTU OZELLIK NUMERIC (numeric bad features)', 'Tramer Numeric (value of accident report)', 'BAKIM DURUMU BOOL (maintenance status)', 'YEDEK ANAHTAR BOOL (spare key availability)', 'BOYA SAYISI NUMERIC (amount of painted parts)', 'DEGISEN SAYISI NUMERIC (amount of changed parts)' columns were added to the dataset thanks to the NER model. In some advertisements, numbers of paint and changing parts are not specified. For this reason, these columns in unspecified cars were replaced with the values in the 'BOYA SAYISI NUMERIC (amount of painted parts)' and 'DEGISEN SAYISI NUMERIC (amount of changed parts)' columns obtained thanks to NER. Other obtained features were added to the dataset to be given to the models.

5.2 Sentiment Analysis

As the project works with the texts that are written by the owners in their way of expression the texts are very open to interpretation. After that conclusion was made, it was decided to apply sentiment analysis and label the text with their overall emotions (positive-negative).
Since the technical requirements to train a model to obtain the sentiments and also the lack of labeled data, (the data could be labeled but it is prone to interpretation and needs expert opinion/labeling) the team had to progress the analyses with pre-trained models.

For the sentiment analysis, Turkish-Bert-NLP-Pipeline [19] was used. Model aims to build a Bert-base NLP pipeline for Turkish; Named Entity Recognition (NER), Sentiment Analysis, Question Answering, Summarization, and Text Categorization. Model is fined tuned based on Turkish-Bert model [20]

After the use of the pre-trained BERT model, the sentiment labels were added to the data as a separate file, with adding another column as 'Sentiment'. Since the scores were also considered to be informative, the sentiment scores of the texts were also included in the remaining works. The table with the results is below

sentiment	sent-scores
negative	0,75800556
positive	0,78415078
negative	0,94270873
positive	0,62236649
negative	0,67087656
positive	0,64798599
positive	0,77817589
positive	0,8977803
positive	0,90155727

Figure 7: Sentiment Scores Column

Positive	5545
Negative	4679
Total	10224

Table 2: Positive and Negative Counts

5.3 Price Prediction Results

The results obtained in the above sections and the importance of using the price prediction algorithm to calculate the effect and effectiveness of the new features were mentioned in section 3.4.

```
['Model', 'Yıl', 'Km', 'Vites Tipi', 'Yakıt Tipi', 'Kasa Tipi',
   'Renk', 'Motor Hacmi', 'Motor Gücü', 'Çekiş', 'Ort. Yakıt Tüketimi',
   'Yakıt Deposu', 'Değişen', 'Boya'],
```

Figure 8: Standart Dataset Features used on Price Prediction

```
['Model', 'Yıl', 'Km', 'Vites Tipi', 'Yakıt Tipi', 'Kasa Tipi',
   'Renk', 'Motor Hacmi', 'Motor Gücü', 'Ort. Yakıt Tüketimi',
   'Yakıt Deposu', 'Değişen', 'Boya', 'sentiment',
   'IYI_OZELLIK_NUMERIC', 'TRAMER NUMERİK'],
```

Figure 9: NER and Sentiment features used on Price Prediction

In addition, Figure 9 shows the features from the NER analysis, and Figure 8 does not include the features from the NER analysis. With the application of the NER analysis, there are many features have created such as ”ıyı ozellık numerıc”,”yedek anahtar bool”,”tramer numerık”,”boyalı sayısı”,”degısen sayısı”. By adding these features to the dataset one by one, the price prediction was tried again to reach the best result.

5.3.1 Performance Comparison:

Three models were used to investigate the effect of new features obtained from description texts on the price in advertisements. These are Random Forest Regressor, XGBoost Regressor and LGBM Regressor. The effects of new features were analyzed separately for each model. MAE, RMSE, R^2 and MAPE metrics were used when evaluating the models.

	MAE	**RMSE**	**R^2**	**MAPE**
Without new features	0.1750	0.2368	0.9426	0.5923
All new features	0.1768	0.2373	0.9423	0.5835
The best features	0.1745	0.2340	0.9440	0.5841

Table 3: Random Forest Results

In Table 3 shows the effect of the newly added features on the random forest model. As can be seen, the new features have a positive effect on the training of the model. The best result was obtained by selecting the sentiment, good feature and tramer (accident report) columns.

	MAE	**RMSE**	**R^2**	**MAPE**
Without new features	0.1728	0.2345	0.9437	0.5929
All new features	0.1731	0.2309	0.9454	0.5918
The best features	0.1709	0.2292	0.9462	0.6138

Table 4: XGBoost Results

The performance of the XGBoost model is shown in Table 4. According to this table, the features ex-

tracted from the description texts strengthen the price prediction model. To get the best results, sentiment, good feature, bad feature and tramer (accident report) columns were selected. In this way, the best result was achieved in the XGBoost model with the newly added data.

	MAE	RMSE	R^2	MAPE
Without new features	0.1661	0.2255	0.9479	0.5281
All new features	0.1660	0.2251	0.9481	0.5211
The best features	0.1640	0.2227	0.9492	0.5086

Table 5: LGBM Results

Finally, the effect of the added data on the price was investigated on the LGBM model. It can be seen that the features added to this model create a positive prediction model. In the research conducted on the LGBM model, the best result was developed by selecting only the Tramer (accident report) column.

With the addition of new features, the MAE has decreased, meaning that the estimates are closer to the actual values.

The R^2 value has increased by adding new features. This indicates that the percentage of explanation of the variance of the independent variables over the dependent variable of the model has increased, that is, it indicates that the overall predictive performance of the model has improved.

With the addition of new features, MAPE decreased, indicating that the predictions were more accurate in percentage terms.

These results suggest that new features from NER and sentiment analysis improve price prediction performance. Lower MAE, RMSE, and MAPE values indicate that the estimates are more accurate and closer to the actual data. Furthermore, a high R^2 value indicates that with the addition of new features, the model better explains the effect of independent variables on price. These findings suggest that NER and sentiment analysis can improve the performance of the model used in used car listing price prediction and provide more accurate and reliable results.

5.4 Results of Clustering

After the price prediction step, the last step was to clustering the advertisements and examine the clusters to determine which of the car advertisements had the best price/performance. Clustering was tried on a dataset consisting of all brands and models, which included 10,000 data. It was also applied to a dataset consisting of Renault brand vehicles, which are a single brand and consist of only 4000 advertisements due to their large number in the dataset. The dataset which only contains Renault car gave better results and it's reason explained on section 5.4.1. So, 3 clusters were assigned as good, medium and bad advertisements according to their feature rankings. Cluster 1 has best advertisements, 0 has medium ones and 2 has the worst ones. Why Cluster 1 is best and the others are different has explained on Chapter 5.4.2. Also The weight and distance values of Features coming from dataset after Clustering process are given in the Table 7.

5.4.1 Optimized Results in Renault Vehicle Data

Clustering is a pivotal unsupervised learning technique used to identify natural groupings within a dataset. The effectiveness of clustering is often evaluated using metrics such as the silhouette score, which measures the compactness and separation of clusters. When clustering 4000 instances, all featuring the Renault brand, the silhouette score was notably higher than when clustering 10000 instances from a dataset comprising multiple car

brands. This section explores the reasons behind this phenomenon and its implications.

The dataset containing 4000 instances is inherently more homogeneous, focusing exclusively on Renault vehicles. This homogeneity contributes to more distinct clustering because the variations within the dataset are specific and limited to one brand. Also Common characteristics and fewer inter-brand variations make it easier to identify natural groupings within the data. In addition variations among different brands (e.g., differences in design philosophy, manufacturing standards, market segments) increase the complexity of identifying natural clusters while maintaining high inter-cluster separation.

Silhoutte Score:

Higher silhouette scores indicate better-defined clusters, with values close to 1 indicating points that are well-clustered.

Silhoutte Score table below shows results of the clustering of study

Metric	Score
Silhouette Score	0.5983
Davies-Bouldin Score	0.5539

Table 6: Clustering Evaluation Metrics

Feature	Cluster 0	Cluster 1	Cluster 2
Fiyat (Price)	0.679	-0.814	0.311
Km	-0.824	0.656	0.026
Değişen (Changed Part)	-0.263	0.095	0.087
Boya (Paint)	0.559	0.255	0.497
Kötü Özellik (Bad Features)	0.154	-0.204	0.093
Tramer (Accident Report)	0.033	0.255	0.009

Table 7: Table of Cluster's Centers

Silhouette Score of Clustering

5.4.2 Explanation of Cluster 0,1,2: How to Determine Which Cluster Has Best Advertisements

In this part of the research, it will be explained why the best advertisements are in cluster 1 and why the other advertisements are in different clusters by looking at the weights of the features on which the clusters are formed.

Price Impact:

- **Cluster 0**: 0.679 (high price, negative influence)
- **Cluster 1**: -0.814 (low price, positive influence)
- **Cluster 2**: 0.311 (moderate price, neutral influence)

A high price is generally worse for price-performance evaluation. Thus, Cluster 1, with a low price, is the best, followed by Cluster 2 with a moderate price, and Cluster 0 is the worst due to the high price.

Kilometers (Km) Impact :

- **Cluster 0**: -0.824 (low km, positive influence)
- **Cluster 1**: 0.656 (high km, negative influence)
- **Cluster 2**: 0.026 (moderate km, neutral influence)

Low mileage is preferable as it indicates the car has been less used. Therefore, Cluster 0 being the best, Cluster 2 moderate, and Cluster 1, with the highest km, being the worst in this regard.

Changed Parts (Değişen) Impact:

- **Cluster 0**: -0.263 (fewer changes, positive influence)
- **Cluster 1**: 0.095 (more changes, negative influence)
- **Cluster 2**: 0.087 (similar to Cluster 1 in influence)

Fewer changes indicate better condition, so Cluster 0 is slightly favored, whereas Cluster 1 and Cluster 2 are not much different here but are worse than Cluster 0.

Paint (Boya) Impact:

- **Cluster 0**: 0.559 (high, negative influence)
- **Cluster 1**: 0.255 (moderate, negative influence)
- **Cluster 2**: 0.497 (high, negative influence)

High amounts of paintwork suggest possible repairs, reducing desirability. Cluster 1, having the least, is preferable, whereas Clusters 0 and 2 are similar but less desirable.

Bad Features (Kötü Özellik) Impact:

- **Cluster 0**: 0.154 (more bad features, negative influence)
- **Cluster 1**: -0.204 (fewer bad features, positive influence)
- **Cluster 2**: 0.093 (some bad features, negative influence)

Fewer bad features (Cluster 1) is preferable, followed by Cluster 2, and Cluster 0 is the worst.

Accident Report (Tramer) Impact:

- **Cluster 0**: 0.033 (low, positive influence)
- **Cluster 1**: 0.255 (moderate, negative influence)
- **Cluster 2**: 0.009 (low, positive influence)

A lower accident report is better, shows how many times the car was involved in an accident and the cost of repairing. Here, Clusters 0 and 2 are better with low Tramer records compared to Cluster 1.

5.5 Synthesis and Conclusion of Results:

Cluster 1 (Best): Lowest price, which is a significant positive. However, high km, high number of changes, and moderate bad features slightly reduce desirability. But due to its low price and fewer bad features, it stands out as the best value in price-performance.

Cluster 0 (Medium): High price reduces attractiveness. Very low km and fewer changed parts are positives. Mixed performance in other variables does not make it the worst, thus keeping it in the middle.

Cluster 2 (Worst): Moderate price, which is not as competitive. Moderate km, indicating average usage which is neither a significant positive nor negative. High paintwork and slightly worse than average performance in other factors such as changes and bad features make it less desirable.

In summary, Cluster 1's combination of the lowest price and fewer bad features results in the best price-performance ratio. Cluster 0, with lower km and fewer changed parts but high price, forms the median. Cluster 2, with moderate values but not excelling significantly in any critical area, falls to be the least desirable for the best price-performance.

6 Conclusion and Future Works

The main aims of the project are to find effect of description texts features on prices on advertisements and to create a model that clusters cars according to their quality for people who want to buy second-hand vehicles without the need to drown in false and long information by training machine learning models with the addition of description texts.

Since the team started to find an answer to these questions of how description texts affect the prices of cars and building a machine learning model that clusters cars based on their quality.

Previous works:

- After the research was done and the methodology for the project was determined, data was taken from the relevant sites by web scraping method, NER model was applied to data to analyze and create new features BERT model was used to get the sentiment labels and scores from the texts.
- The effect of the features obtained from the description texts on the price was investigated.
- Quality cars were clustered using clustering models with newly added features

Within the scope of the project, NER and sentiment features were added to the dataset and their effect on price was examined. In the research conducted with Random Forest Regressor, XGBoost Regressor and LGBM Regressor models, a slight positive effect of description texts on price prediction was observed. Each model gave different results with different NER and sentiment features. The best results were observed with the selected features.

After investigating the impact of new features on price, the cars were clustered as good, average and bad with the KMeans model. The model was concluded with a silhouette score of 0.60 and a davies bouldin score of 0.55. Then the model was saved to be used in other projects.

Future works:

- A sentiment analysis model specific to this project will be customly trained like NER model.
- The NER model will be trained with more data and labeled with more specific labels, so that the features in the texts will be extracted with better results.
- Models used with more functional sentiment analysis and NER model will give better results.
- New datasets will be created for each car model and the clustering algorithm will be run separately for all of them.
- A user interface for using the models or a plugin for the relevant websites will be created

References

[1] B. Lal and C. R. Chavan, "A road map: e-commerce to world wide web growth of business world," *Global Journal of Management and Business Research*, vol. 19, no. A11, pp. 33–42, 2019.

[2] V. Thomas, L. B., "The design of web based car recommendation system using hybrid recommender algorithm," International Journal of Engineering Technology, 2018.

[3] K. R. A. Samruddhi, K., "sed car price prediction using k-nearest neighbor based model," pp. 629–632, . Int. J. Innov. Res. Appl. Sci. Eng.(IJIRASE), 2020.

[4] F. S. Tsagris, M., "Advanced car price modelling and prediction," Springer, 2022.

[5] C. X. huancan Chen, Lulu Hao, "Comparative analysis of used car price evaluation models," AIP, 2017.

[6] A. A. Jalal, "Text mining: Design of interactive search en- gine based regular expressions of online automobile advertise- ments," pp. 35–48, 2020.

[7] A. T. A. I. T. S. N. S. K. S. Asghar, Z., "Sentiment analysis on automobile brands us- ing twitter data.," Springer, 2020.

[8] D. Altinok, "A diverse set of freely available linguistic resources for Turkish," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (A. Rogers, J. Boyd-Graber, and N. Okazaki, eds.), (Toronto, Canada), pp. 13739–13750, Association for Computational Linguistics, July 2023.

[9] Y. Zhang, H. G. Li, Q. Wang, and et al., "A filter-based bare-bone particle swarm optimization algorithm for unsupervised feature selection," *Applied Intelligence*, vol. 49, no. 7, pp. 2889–2898, 2019.

[10] "Cars.com," 2023.

[11] "arabam.com," 2023.

[12] P. Cihan and E. Cerrahoglu, "Web scraping and machine learning techniques to prediction of secondhand car prices," *Journal Name*, 2022.

[13] T. N. Suite, "Turkish spacy models," Publication year, e.g., 2022.

[14] G. D. Experts, "Brand new spacy turkish models," Publication year, e.g., 2022.

[15] S. Yıldırım, "Bert base turkish ner cased." `https://huggingface.co/savasy/bert-base-turkish-ner-cased`, 2021. Hugging Face.

[16] R. Xu and D. Wunsch, "Survey of clustering algorithms," *Neural Networks, IEEE Transactions on*, vol. 16, pp. 645 – 678, 06 2005.

[17] P. Berkhin, "A survey of clustering data mining techniques," *Grouping Multidimensional Data*, pp. 25–71, 2006.

[18] Tecoholic, "Ner annotator," n.d. Accessed: 2023-10-09.

[19] S. Yildirim, "Turkish BERT NLP Pipeline." `https://github.com/savasy/Turkish-Bert-NLP-Pipeline`, Year.

[20] S. Schweter, "Turkish BERT: Pre-trained Turkish Language Model." `https://github.com/stefan-it/turkish-bert`, Year.