ALMA MATER STUDIORUM – UNIVERSITA' DI BOLOGNA

DIPARTIMENTO DI SCIENZE STATISTICHE "PAOLO FORTUNATI"

Corso di Laurea in Scienze Statistiche

Curriculum Economia ed Impresa

Analisi testuale e sentimentale sul mondo delle autovetture

Presentata da:

Tommaso Possenti
Trapin Matricola: 001002373

Relatore:

Prof. Luca

APPELLO III

ANNO ACCADEMICO 2022/2023

Index

1. Introduction

The aim of this paper is to analyze and represent the car market through the eyes of consumers, thanks to the reviews released on the Edmunds website.

This website contains information and reviews on numerous car models available for sale in the United States.

The analysis and representation of the automobile market are realized within three main questions that contain some of the most relevant aspects of the Market Analysis.

Specifically, we want to verify whether it is possible, through text and sentiment analysis:

1. Identify the main positive and negative characteristics that consumers associate with different brand products.
2. Quantify the impact that perceptions of certain attributes have on the overall evaluation of a product.
3. Trace the characteristics that have determined the trend of product ratings over the years.

This thesis aims to present the results obtained by applying text mining and sentiment analysis techniques on a dataset consisting of reviews and ratings belonging to the automotive market world.

The analysis of this market is particularly interesting becauseAutomobiles and their purchase represent a central part of the socioeconomic sphere of modern man.

The automobile is a good whose choice represents the result of a reasoned and in-depth research by the consumer, being an infrequent purchase and with a high cost.

The purchase of a model is therefore subsequent to a phase of logical and objective evaluation by the consumer; therefore, it will be the characteristics and benefits associated with the car that will define its success in terms of sales.

Furthermore, this market appears to be ideal for the analysis that will be

developed given its size in terms of: consumers, available car models and levels of attributes of which each model has different combinations.

The considered dataset contains reviews and ratings expressed by consumers on numerous brands and models commercially available between the years 2000 and 2020.
It was decided to use a dataset consisting mainly of comments since these represent the main feedback that a company receives from its consumers and carry large amounts of information.
However, the comments, being developed directly by the consumer, contain information that is latent and therefore to extract them it is necessary to work on the texts in order to standardize them and make them analyzable.
By using text mining tools and combining them with sentiment analysis functions, it was possible to detect consumer perceptions of cars and their main attributes.

The paper includes three key questions that represent some of the most relevant aspects within a market analysis.

In the first question: “Is it possible to identify positive and negative characteristics associated with car models through a textual analysis performed on reviews?” We asked ourselves whether it was possible, using the comments as input, to obtain useful information for developing a first overview of the car market.
More specifically, we asked ourselves whether it was possible to identify the strengths and weaknesses of the models in the eyes of the customer, as well as any defects and problems encountered by consumers; all fundamental information in terms of marketing and product improvement.
To do this, the frequencies of single words and pairs of expressions within the comments were evaluated, with the aim of finding the most used expressions in the description of each model.
This question aims to detect consumers' perceptions associated with the characteristics of the various models. We have researched the characteristics of a car most associated with certain terms, both positive and negative, such

as: "good", "bad", "problem".

For example, the expression “problem” was considered and the algorithm identified the characteristics most associated with this term within the comments.

This procedure allowed us to obtain an overview of the perceptions that the consumer has developed towards a model and its attributes.

In the second question: "Is it possible to evaluate how much consumer perceptions of a car's features impact the overall rating starting from comments?" We evaluated how much consumers' positive or negative perceptions of some features of a model impacted the final score released in the reviews.

This analysis used the information collected in the frequency analysis to define four main characteristics: “look”, “drive”, “Comfort” and “transmission”. Starting from these four characteristics, a multiple regression analysis was set up, in which the characteristics considered represent the independent variables.

While, the dependent variable is represented from the final score released by consumers within each review.

This question aims to evaluate the impact that the consumer's perception, with respect to four characteristics of the car, has on the final evaluation, with the aim of verifying whether, starting from comments, it is possible to develop an analysis capable of identifying the most important characteristics for the consumer and quantifying their relevance.

In the third question: “Using text mining and sentiment analysis functions, is it possible to understand the causes of the trend of the average scores of a model over the years?”

The historical series of average annual scores released by consumers for the Ford brand Focus model is graphically represented.

Subsequently, the historical series of average annual sentimental scores detected in the comments for the same model was also evaluated.

The most used positive and negative expressions in each year were then detected, thus rediscovering the causes of the trend of the historical series in the years.

Finally, the commercials of the years considered were found in order to evaluate the marketing activity used by the Ford brand with respect to the trend of the historical series.
This question aims to evaluate the evolution of the Ford Focus model over the years, through the eyes of the consumer and study the marketing strategies used.

The thesis is structured in the following manner:
Chapter two will explore the techniques used to initiate the research in terms of exploration and preparation of data for subsequent analyses.
In chapter three, the concepts of text mining and sentiment analysis will be explained, examining in depth the algorithms and functions used in the empirical part of the analysis.
In chapter four you will find the questions described above along with the software outputs reported in the form of graphs or tables used during the information extraction.
The thesis ends with chapters five and six containing final conclusions and bibliography/sitography respectively.

2. Preliminary data description and analysis

2.1 Dataset

The starting dataset was extracted from the KAGGLE website, a competition and collaboration platform in the field of “data science” and “machine learning”.

The dataset identified was initially called: “Edmund car review”, as the name itself suggests the data within it comes from the digital archive and website: “Edmund”, which constitutes an important digital resource for the automotive industry, both in the form of an archive of physical characteristics of cars put into circulation and of particularly extensive reviews and comments on the cars themselves.

The dataset was created through a web-scraping operation structured around 46 car brands such as BMW, Alfa-Romeo, Honda, Fiat, for a total of 897 models put on the market in the time frame from 2000 to 2019.

The dataset consists of eight columns: Company, Model, Year, Reviewer, Date, Title, Score, Review; and 299,045 rows, each containing information associated with a single comment.

As for the score column, it contains an evaluation on a scale from zero to five in which consumers report an overall rating relating to the model considered.

While the review column contains the comments left by each user or consumer.

2.2 Initial procedure

2.2.1 First interaction with data

Once the starting dataset was identified and imported into the “RStudio” software, an initial phase of familiarization with the data and exploration began.within the dataset itself in order to evaluate the overall formal correctness of the content and the presence of comments or other empty fields (“NA”).

The dataset was overall free of incorrect or absent forms.

This was followed by a random evaluation phase of the correctness and quality of the comments, in order to find any comments written in languages other than English and to be able to establish whether the comments

contained sufficient information to carry out the entire analysis.

2.2.2 Data cleansing and reorganization

Once the overall validity and usefulness of the data had been verified, we moved on to drafting the first part of the R code called “Skimming Cleaning”, which in fact sees two different phases presented.

In the first part of the code, the dataset was reduced with respect to the variables: “Year”, “Model” and “Brand”; in the first part of the analysis, in fact, they were considered for car models, Ford Focus, Hyundai Tucson, Toyota Prius, Toyota Highlander, within the year 2016, all this in order to speed up and optimize the analysis time needed by the system to process the data in the subsequent phases.

In the second part of the code, functions and algorithms belonging to the world of text analysis, or Text Mining, were introduced.

Specifically, it introduced a text cleaning phase, in which contents that were not useful for analysis were eliminated from the comments, such as: numbers, punctuation, uppercase characters and symbols.

Subsequently, expressions that were not very useful due to the low level of information they conveyed and their too frequent use were eliminated, such as: “macchina”, “comprare” and “avere”.

After cleaning up the texts, we moved on to transforming them into a structured form and standardized, all expressions were operationally brought back to their basic form in order to obtain homogeneous texts.

Expressions such as verbs, adjectives and nouns have been stripped of conjugations and endings and traced back to their roots.

3. Text Mining and Sentiment Analysis

Text mining is a process of extracting latent information and hidden patterns contained within lines of text through automated analysis.

Text mining is also the basis of key services that companies offer to their consumers, such as customer service that has become completely automated thanks to chat bots.

Such software or bots detect and understand the problems expressed from consumers through text analysis.

Specifically, text mining is a set of functions capable of cleaning the text from useless or irrelevant terms for the analysis and of standardizing the expressions in order to allow the development of further analyses such as the sentimental one.

Operationally, starting from comments, we proceed to clean them from any errors or unwanted elements, we perform a modification of some expressions that in certain cases are removed or are identified as compound forms, such as: “road noise” or “mile per gallon”.

Text mining can be compared to the reverse process of building a brick wall.

The starting text represents the complete and finished wall, through algorithms and functions we aim to analyze the individual bricks of the wall, that is, the individual expressions or words of the text.

The individual bricks are extracted and evaluated in order to understand each of their characteristics and therefore to be able to fully understand the characteristics of the brick wall itself and be able to extract useful information for the analysis.

Within the textual world, on which text mining is based, we find three different levels:

a) Macro level: the corpus, which can be imagined as a book or in our case the quasimillion reviews available
b) Intermediate level: the texts,that we can think of as chapters and therefore, in the current case, the single review
c) Levelmicro: the words

The analysis and development of algorithms starts from the intermediate level in which each text is cleaned through various phases that have the task of eliminating any unwanted elements or those considered irrelevant for the analysis, such as: hashtags, numbers, mentions.
Subsequently, the algorithm proceeds to eliminate from the text any expressions that lack useful information for the analysis given their logical-grammatical nature, expressions called “stopwords”.
Expressions that fall into this category are automatically identified by the algorithm through the use of specific dictionaries for the language of the text being analyzed.
In the dataset considered in this paper, stopwords such as: “I”, “the”, “if”, “then” were found.

To these expressions detected through dictionaries the user can add words that he considers irrelevant or too frequent and therefore useless. Among the words thus identified are: “car”, “vehicle”, “honda” and “toyota”.
This cleaning phase is followed by a text reorganization phase, which aims to identify and correct compound forms, i.e. combinations of multiple words that form a single, self-contained expression.
For example, the expression “spare wheel” even though it is made up of several words can be considered as a single expression and therefore the algorithm will change it to “spare_wheel”.
This operation aims to preserve the information carried by the various expressions so as to simplify and make the analysis more reliable.
Even in this phase, both digital dictionaries and expressions introduced directly by the user are used, such as: “steering wheel” which is rewritten as “steering_wheel”.
The transformation of compound forms allows us to facilitate the

interpretation of the text and the search for the logical meaning of the sentence.

The cleaning and reorganization of the text is then followed by a phase called lemmatization. In this phase each word of the text is traced back to its lemma, where lemma means the combination of word and grammatical category.

This operation allows us to normalize different conjugations of verbs or the declension of nouns, adjectives, pronouns, for example: in front of the words “cars”, “drove” the output after lemmatization will be “car, NOUN”, “drive, VERB”.

Once the Text mining phase is completed, we proceed to introduce the sentimental analysis functions.

Through Sentiment Analysis it has become possible to identify the emotion contained within a text, that is, it includes the tone with which a consumer has left a certain review.
The primary goal of this analysis is to identify the text and the expressions within it as positive, negative or neutral.
To do this, digital dictionaries are used, through which the software is able to identify expressions that can be associated with moods such as anger, joy, surprise, frustration, within comments;
Each emotion detected in the comments has its own sentimental score which can be greater or less than zero and therefore a general meaning positive or negative.
The sum of the sentimental scores of the expressions that form a comment defines the overall sentiment score of the review.
It is therefore possible to develop a study of emotions and feelings reported within the comments.
From a procedural point of view, two starting datasets were considered: one resulting from text cleaning, the other from the lemmatization of the cleaned texts.
The first of these datasets contains on each line a comment cleansed of symbols, punctuation, capitalization, compound forms and stop-words.
While in the second dataset, on each row we find the individual expressions that

constitute each comment, each labeled according to the comment to which it belongs.
In this second dataset, each expression is accompanied by the morphological nature of the expression, such as noun, verb, pronoun or adjective.

Using the first dataset, an analysis was performed that led to the definition of the sentimental score associated with each comment.
This operation was performed using the R function “get_nrc_sentiment” from the “syuhet” package.
The algorithm proceeds to analyze the text and provides as output a dataset in which each row corresponds to a text and the columns are named as follows: "anger", "anticipation", "disgust", "fear", "joy", "sadness", "surprise", "trust", "negative", "positive".
The algorithm reports under each column the number of expressions, detected in the comments, that correspond to the emotion title of the column.

In practice, this function reports how many times each emotion was used in a comment.
Based on the frequency of each emotion, the algorithm derives the number of positive and negative emotions reported in each comment, this within the last two columns of the dataset.

By adding the opposite of the negative emotion score to the positive emotion score, we obtain the “valence” of the entire text, or the sentimental score of the entire comment.
By doing this, you can understand the overall tone in which each comment was developed.
The relationship between the average of the scores released directly in the evaluations (x-axis) and the average sentimental scores (y-axis) was then evaluated by using a scatter-plot graphic representation on the data relating to sixteen models in the year 2014.

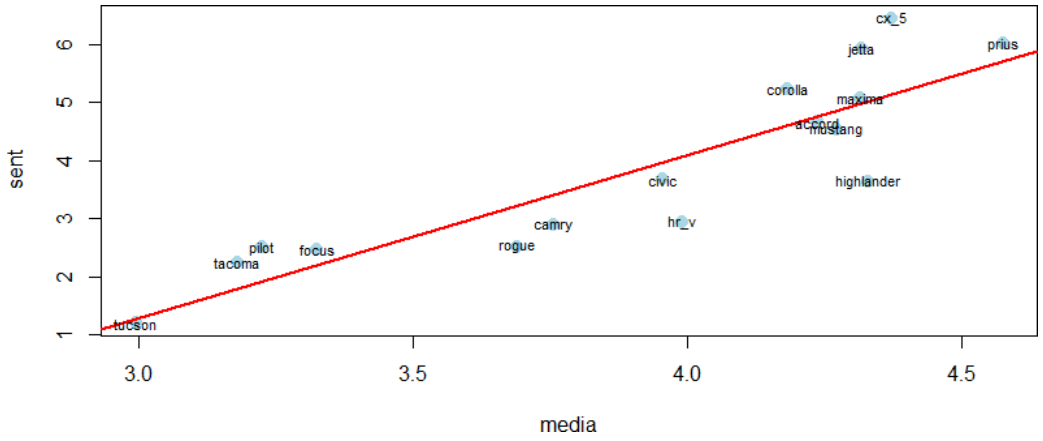

Figure 1, Scatter plot of 16 car models. The x-axis shows the average scores calculated for each model in 2014. The y-axis shows the average sentimental scores found in the comments for each model in 2014.

In Figure 1, note the red line, which represents the ideal ratio of average released scores to average sentimental scores.

It is immediately noticeable how the cloud of points formed by the models is predominantly compact and adherent to the line, confirming the logical connection between the scores released and the sentimental scores calculated.

However, you can see some models that are further away from the straight line than others, such as the Mazdacx_5, which has an average score of around 4.4 which should correspond to a sentimental score of around 5, but which in reality turns out to be higher than 6.

This result highlights the importance of sentiment analysis in terms of the additional information that can be extracted from comments.

As for the second dataset, this was exploited to perform sentiment analysis on subtexts containing specific expressions.

In fact, parts of the text containing expressions such as “look”, “Comfort” and “transmission” were identified and the emotions most commonly reported within these subtexts were evaluated.

Through this procedure it was possible to develop a research of the sentimental score associated with specific characteristics of a car and therefore an analysis was carried out which had the objective of evaluating the perceived importance of some characteristics of the car.

4. Empirical analysis

4.1 Is it possible through a textual analysis performed on reviews to identify positive and negative characteristics associated with car models?

This question aims to identify the positive and negative characteristics perceived by consumers that most characterize the various models.
Specifically, we want to develop a table that summarizes the most frequently cited features in the reviews of four different car models: Ford Focus, Hyundai Tucson, Toyota Prius and Toyota Highlander in the year 2014.
These four models were identified on the basis of two criteria: the number of comments and the position within the graph in Figure 1; with the aim of considering different and varied models in terms of average scores released and possibly of associated positive and negative characteristics.
The analysis of this paragraph is based on the evaluation of the frequencies of the expressions contained in the comments.
During the analysis, different methods of frequency detection will be used and for each of these methods a table of the highest frequency terms will be developed.
The final table, resulting from the refinement of the analysis, will represent an overview summary of the car market from the consumer's point of view.
This table will report: the expressions most associated with each model, the most frequently used positive and negative characteristics, and the expressions most frequently used alongside terms such as “problem” or “issue”.
The purpose of this table is to represent the perceptions of the models through the eyes and experiences of the customer.
The purpose of this analysis is to identify the strengths and weaknesses of the various models, identifying possible improvements to the product that satisfy the customer's needs.

The first step in this analysis was to evaluate the number of comments available for each model in the year 2014, which was found to be approximately two hundred for each model considered.
Starting from the dataset consisting of these comments, we proceeded to use a

“filter” function to extrapolate the data relating to the models chosen in the year considered. Before developing the frequency analysis, it was necessary to perform four text mining operations aimed at eliminating or at least limiting as many main problems.

As a first step, the texts were traced back to the same encoding code, in this case UTF-8 encoding was used.
The reason for this operation is to make the texts uniform and avoid that comments with particular encodings are not properly displayed by the software. In this phase, the texts were also cleaned from elements such as numbers, symbols and capital letters.

As a second step, it was necessary to standardize the expressions within the texts; each expression was brought back to its basic form in order to eliminate any conjugations or endings that would otherwise have distorted the frequency evaluation.
Through this process the texts are transformed in such a way as to make them comparable and uniform.
For example, the expressions “drive”, “drove”, “driven” which are different from each other but have a similar meaning, are brought back to the base form of the verb and are therefore all transformed into “drive”.

The third step was to eliminate from the texts the expressions called “stop-words”. These expressions can be pronouns, articles or expressions that are frequently used but convey little or no information in the context of a sentimental or textual analysis.
The task of these expressions in spoken and written language is to connect the various elements of a sentence, to make its meaning complete.
These terms were removed using a dedicated dictionary of English stop words, which was fed into the algorithm used to clean the texts.
These expressions can also distort the frequency detection since they can easily obtain very high frequencies.
It is important to note that under the stop-word category, the following have been manually inserted:

of expressions specific to the automotive world, such as the terms “car” or “buy”, expressions that do not convey any additional information for analysis.

The fourth step was to identify and mark compound expressions,that is, combinations of two, three or even four words capable of forming an expression with a complete and self-contained meaning.
Examples of these expressions could be “spare tire” or “passenger seat”. Here too, dictionaries and manual interactions were used to detect such expressions and mark them as unique expressions and not as separate words.
This is to avoid the loss of information carried within such composite forms.

Perform theseThe first steps were to detect frequencies.
Different methods were used in evaluating the frequencies of expressions for assigning weights.
Initially, weighting was used.
This first method was used to evaluate the “term-frequency”, that is, the frequency with which certain expressions appear within the comments, that is, the number of times a term appears in all the comments.

This first weighting method, however, had two main problems.

Excessive importance is given to expressions that cannot be categorized as stop words, but which nevertheless present high frequencies, as in the case of the term “get” which is used in most texts but does not convey any relevant information.

The second problem is the loss of information due to the failure to divide comments into positive and negative ones.
This problem is based on the inability to understand why a certain expression is quoted so many times in the comments.
In fact, without a division between positive and negative comments, it is not possible to understand the true meaning of the frequencies or, better, it is not possible to trace them back.

to the cause of high use of a certain term.
For example, there has often been a situation where the expression “transmission” was used extensively without being able to say whether the transmission was often cited due to excellent mechanical quality or due to technical problems and defects associated with it.
In solving these problems two different methods have been identified.

1. Frequency of expressions with weighting tf-idf

   To putTo remedy the shortcomings and defects of the TF method, a new weighting method, the TF-IDF method, was used.

   Where Tf stands for “term frequency” while idf represents the acronym for “inverse document frequency”, that is, the inverse of the numbers of documents in which a certain word appears.

   This new weighting gives a higher score to terms that do not appear in too many comments.

   This method rewards expressions that have a greater value in terms of information conveyed, despite their low frequency.
   While it limits the importance of terms with a high frequency but that do not have a significant semantic value, as in the case of the expression "get". In practice, this method reduces the frequency associated with terms that are found in too many comments.

   The new tf-idf weighting method was then associated with a division of comments into positive and negative ones.
   Comments with a sentimental score equal to or greater than zero are considered positive, while comments with a sentimental score lower than zero are considered negative.
   In this way it was possible to develop a frequency research that allowed to assign a correct and fair relevance and meaning to the detected results.

2. Co-occurrence starting from predefined expression

This method moves away from evaluating the frequencies of individual expressions and focuses on evaluating co-occurrences, i.e., detecting the frequencies of pairs of expressions.

The aim of this method is to directly assess associations between expressions with positive or negative connotations and the characteristics of a car.

By identifying the car features most often listed alongside positive terms, it is possible to identify the attributes most appreciated by consumers.
Specifically, the code that has been developed requires, as a starting matrix,the matrix consisting of the co-occurrences of all expressions within the comments.
After defining a starting term, the algorithm detects all pairs of expressions containing that term and orders them by decreasing frequency.

Model	1st	2nd	3rd	4th	5th	6th
Prius	Get	Age	Wheel	Look	Drive	Better
Focus	Transmission	Get	Problem	Issue	Start	Clutch
Highlander	Seat	Love	Build	Good	Purchase	Quality
Tucson	Bad	Old	Get	Seat	Drive	Better

Table 2, Table of Prius, Focus, Highlander and Tucson models. The first column reports the models considered and the six subsequent columns report the six expressions with the highest frequency for each model detected with the tf weighting method and ordered in decreasing order.

Table 2 shows the expressions with the highest frequency in the comments of the four models considered.
This table reports the six highest frequency expressions detected in the comments of each model using tf weighting and without dividing between positive and negative comments.
The purpose of this table is to show the problems associated with the tf method initially described.
In fact, for each model we see expressions such as “get” or adjectives used that

do not provide the information we are looking for in this analysis.

Models	1st	2nd	3rd	4th	5th	6th
Prius	Great	Love	Seat	Handle	Comfort	Quality
Focus	Problem	Transmission	Fix	sync	Good	Issue
Highlander	Seat	Love	MPG	Price	Better	Quality
Tucson	Dislike	Bad	Problem	Seat	Transmission	Better

Table 3, Table of Prius, Focus, Highlander and Tucson models. The first column reports the models considered and the six subsequent columns report the six expressions with the highest frequency for each model detected with the tf-idf weighting method and ordered in decreasing order.

Table 3 shows the results obtained by replacing the tf method with the tf-idf method.

By comparing the table just obtained with Table 2, one can understand the difference in results obtained with the two methods.

Through the idf method we obtain different and apparently cleaner and more useful results than those obtained with the tf method.

We lose those terms that are characterized by a very high frequency in each comment, such as, for example, the term “Get” which leaves the first six places in the table.

The analysis with the idf method was performed on nouns, verbs and adjectives, jointly;However, to further analyze the positive and negative characteristics associated with each model, it was decided to limit the frequency analysis with the idf method to the names only. This is to focus the research attention on the attributes of the models, in order to identify only the most commented characteristics, then dividing the comments into positive and negative ones.

The division of comments between negative and positive was set by using the sentiment analysis functions.

The “valence” of the comments was used as a criterion for discriminating between positive and negative comments, that is, the overall sentimental score of the comment, resulting from the sum of all the sentimental scores associated with each single expression in the text.

To calculate the valence, an algorithm was developed that, starting from the dataset of lemmatized texts and from English-language dictionaries of emotions, returned a matrix with the individual comments as rows and all the following emotions as columns: "anger", "anticipation", "disgust", "fear", "joy", "sadness", "surprise", "trust", "negative","positive".

Subsequently, through the R function “get_nrc_sentiment” of the “syueth” package, a pre-established weight was assigned to each emotion and, based on the detected emotions, the sentimental score of each comment was calculated.

Comments with a sentimental score equal to or lower than 0 were then considered “negative” and all comments greater than 0 were considered “positive”.

Once positive and negative comments were separated, the frequency evaluation using the idf method was performed on each group of comments with respect to the terms in the form of names only and the first five terms were evaluated in decreasing order of frequency in the case of Ford Focus.

The output of this process is reported in theTable 4.

Focus	1st	2nd	3rd	4th	5th
Positive	Interior	Quality	Look	Handle	Transmission
Negative	Noise	Transmission	Sync	Feature	Clutch

Table 4, Table of the five most used expressions in positive and negative comments on the Ford brand Focus model, detected using the tf-idf weighting method and ordered by frequency in decreasing order.

What we have obtained in Table 4 is the list of the attributes of the Ford Focus model most cited in the comments, divided into positive and negative.

The table therefore represents an ordered list of the characteristics that consumers most associate with the benefits and disadvantages.of the Ford Focus model.

From these results, it can be seen that the 2014 Ford Focus model was characterized by good overall quality and interior quality but by serious problems associated with transmission and noise.

To validate these hypotheses, the analysis was deepened by using the evaluation of co-occurrences, with the aim of better understanding the

consumer's perception of the model.

To set up the co-occurrence analysis, three most used positive and negative adjectives within the comments were identified and the characteristics most associated with them were evaluated.
The adjectives thus identified were: “good”, “great” and “nice” for positive adjectives and “bad”, “terrible” and “awful” for negative adjectives. The co-occurrences of the three adjectives were thus evaluated.both positive and negative with other expressions contained in the text and the first two, three or four terms with the highest joint frequencies were extracted.

In fact, we evaluated which characteristics of the car were most associated with positive and negative adjectives in the comments released by consumers.

They have also evaluated the co-occurrences of the terms used in the texts with the expressions “problem” and “issue”, to evaluate which parts or characteristics of each model were defective according to the consumer.

The results of the co-occurrence assessment were summarized in the
Table 5.

Model	1st (and 2nd) term plus used	Car. Pos.	Car. Neg.	Problems
Prius	Mpg // drive Great, average // fun	AccelerationHandle Drive	NA	NA
Focus	Transmission // noise Problem, issue, fix //wind, grinding	QualityInterior Look	Acceleration Transmission Noise Sync	TransmissionNoise
Highlander	Seat Comfort, leather, support	Look Drive Interior	Noise Wheel	Vibration
Tucson	Seat // Drive // Price Comfort // Fun // Great	Control Price	Brake Quality	Audio Dealership

Table 5, Table containing in the first column the names of the Prius, Focus, Highlander and Tucson models. In the second column the first terms by idf frequency found in the comments and below these the respective expressions most associated with them. The third and fourth columns respectively report the expressions with high co-occurrence with respect to the terms "good", "great" and "nice" and the terms "bad", "terrible" and "awful". The last column reports the expressions most associated with the terms "problem" and "issue". Fields marked with Na indicate that no associations or co-occurrences were found with a frequency greater than ten within all comments.

The symbol"//" is used to separate the most used terms and the expressions associated with them. In the Prius case, for example, "Mpg" is associated with the expressions "Great" and "Average", while "Drive" is associated with the expression "Fun".

Table 5 provides a schematic representation of the most commonly used expressions in reviews and the terms associated with them, the main positive and negative characteristics and the main defects encountered by consumers.

Remaining on the Focus case, it can be noted how the first expressions for

frequency reported by consumers are Transmission and Noise, to which negative and problematic terms are associated.
For example, the term “noise” is associated with the expressions “wind” and “grinding” which indicate the type of noise that consumers complain about hearing.
Moving on to the next columns, you can see the positive and negative characteristics most associated with the model.
According to consumer reports, the Ford Focus model has good aesthetics and a fair overall quality.
However, there are negative ratings regarding Transmission and Noise to which are added the expressions Acceleration and Sync.

It can be concluded that in 2014 the Ford Focus model has certainly suffered some repercussions in terms of evaluations, this is due to the problems detected and perceived by consumers with respect to Noise and Transmission.

By interpreting Table 5, it is possible to identify the advantages and disadvantages of the Ford Focus expressed through the words of consumers.
It was therefore possible to summarize the contents of the comments, obtaining an evaluation of the car's attributes through the words of the consumers.
This analysis made it possible to identify the main defects and advantages of the Ford brand Focus model, allowing us to obtain information also on the “opponent” models.
The information that has been collected represents an invaluable source of feedback regarding the characteristics of the car, which can allow a company to get a representation of your model's positioning in the consumer's mind.

4.2 Is it possible to evaluate how much the consumer's perceptions on the characteristics of a car impact the overall rating starting from comments?

The purpose of this paragraph is to understand which are the characteristics, reported in the comments, that most influence the final score assigned to the model by consumers.

The final score refers to the rating given by each consumer together with the review in the form of a comment.

If we have previously evaluated the main characteristics associated with a model, now we evaluate how much the perception of these characteristics actually impacts the final evaluation.

We want to understand if it is possible, through textual and sentimental analysis applied to comments, to understand the importance and relevance that consumers associate with some of the attributes of a model.

To develop this analysis, the most discussed and commented features in the reviews of the Ford Focus model in 2012, 2013, 2014 and 2015 were identified, such as the quality of the interior or the level of comfort.

Subsequently, the subtexts of the comments containing these characteristics were extracted and their sentimental scores were calculated.

Based on these sentimental scores and the evaluation given by the consumer, a multiple regression analysis was developed.

Operationally, the analysis starts from the dataset of cleaned texts with respect to the Ford Focus model in the years 2012, 2013, 2014, 2015.

Subsequently, the expressions with the highest frequency within the model comments were identified as the characteristics to be looked for in the analysis, namely: “Quality”, “Look”, “Interior”, “Acceleration”, “Transmission”, “Noise” and “Sync”.

A lemmatization operation was then performed on the cleaned texts, which generated a new dataset in which each line contains a word detected in the comments and the respective comment to which it belongs.

In this decomposition phase,they then list all the expressions used in the dataset

and the reviews that contained them.

The next step was to identify among the comments those that jointly contained all the characteristics underlying the analysis.

However, the comments containing these characteristics jointly were 0, therefore, the number of searched attributes was reduced.

The final combination that was selected was made up of the following characteristics: “Look”, “Drive”, “Transmission” and “Comfort”; these characteristics were found together within approximately eighty comments.

Once the new starting dataset, consisting of eighty comments, was identified, we proceeded to carry out the sentimental analysis.

Each comment was divided into four subtexts, so that each subtext included within it one of the characteristics considered.

We then calculated the sentimental score associated with each of the four subtexts of each comment.

Finally, a dataset was created containing the sentimental scores of the four subtexts, the identification code of the comment to which they belong and the score released by consumers in that comment.

doc_id	Look_sent	Drive_sent	Transmission_sent	Comfort_sent	Rating
1	+2	+1	-3	0	3

Table 6, Table of the first row of the dataset used in the multiple regression analysis. The first column contains the identification code relating to the comment considered, the next four columns contain the sentimental score of the subtexts containing the expressions: “look”, ”drive”, “transmission” and “Comfort” respectively. The last column contains the score released by the consumer.

Starting from the newly created dataset, the first row of which is shown in Table 6, it was possible to develop a multiple regression, having the Look_sent, Drive_sent, Transmission_sent and Comfort_Sent columns as independent variables and the Rating column as dependent variable.

The results of the regression thus formed are reported in Table 7.

Characteristic	Coefficient	P-value
Look	0.874	0.097
Drive	0.620	0.236
Transmission	1.196	0.088
Comfort	0.739	0.125

Table 7, Output of the multiple regression with the Rating as the dependent variable and the Look_sent, Drive_sent, Transmission_sent and Comfort_Sent as the independent variables. The first column shows the four characteristics considered. The second column shows the coefficients associated with each independent variable. The last column shows the p-values associated with each characteristic.

Analyzing the output, one can immediately notice how the p-values are overall high,generally greater than or equal to 0.1, this is probably due to the small size of the dataset used for the analysis.

In fact, the coefficients associated with “Drive” and “Comfort” have p-values that are too high to be considered significant and therefore to be considered reliable within this analysis.

However, evaluating the results obtained one can see how “Look”and “Transmission” are the only variables to be characterized by relatively low p-values, in fact the associated reliability level is higher than 90%.

It was decided to accept this level of reliability and therefore to consider the two characteristics as significant for the analysis.

The coefficients associated with “Transmission” and “Look” are 1.196 and 0.874 respectively, this result demonstrates how a unit increase in the sentimental score associated with the individual expressions can increase the average score.

For “Transmission,” for example, a unit increase in the sentimental score associated with this characteristic would lead to an increase of 1,196 in the final rating, if all other dependent variables remained constant.

Therefore, through this analysis it was possible to detect the impact that the “Transmission” and “Look” characteristics have on the final evaluation.

This information combined with what was seen in the previous paragraph showshow the generally negative rating of “transmission” may have directly influenced the final score released by consumers.

It can be concluded that, using the functions of sentiment analysis, it is possible

to organize an analysis capable of detecting the importance and weight associated with certain model features by consumers.

In fact, it was possible to quantitatively evaluate the impact that an improvement or worsening of consumer perceptions with respect to “Transmission” and “Look” would have on the final score in the case of the Ford Focus.
It has therefore been possible to identify the characteristics of those whose perception mostly determines the assignment of the final score and its actual impact on the latter has been quantified.

This type of analysis would allow a company to understand the aspects that are most important to consumers and that have the greatest impact on product evaluations.
So, by exploiting this type of analysis, it would be possible for marketing to identify the key aspects on which to focus advertising campaigns or product improvement activities and to maximize the effects of any investments.

Furthermore, having seen the regression output on the perceptions of the model's characteristics, the marketing manager could define a new objective of the marketing activity as a unit increase in the sentimental score associated with a certain attribute.
Thus allowing us to numerically evaluate any progress or failure in achieving this goal.
What has been developed is a tool that is also able to provide feedback to the development team regarding the quality of the materials and technologies used in the manufacturing of the model.
All this information vital to the company and the success of the product has been extracted from the comments, using knowledge of text analysis and sentiment analysis.

However, during the development of this analysis a major problem was encountered. This problem concerns the quantitative limitation of the data, more specifically the low number of annual comments reported within the dataset, which by construction contained no more than 200 annual comments.

This small data size, coupled with the search for comments with four common expressions within them, significantly reduced both the size of the dataset and the reliability of the analysis itself. Certainly larger datasets would guarantee greater reliability and would allow for the development of broader analyses also in terms of the characteristics considered.

4.3 Using text mining and sentiment analysis functions, is it possible to understand the causes of the trend of the average scores of a model over the years?

In this paragraph we want to find the characteristics that dictated the trend of the historical series of average and sentimental scores over the years.

Specifically, we want to find the most frequently reported positive and negative characteristics in the comments of each year, so as to understand which qualities of the car may have defined its success or vice versa which defects may have caused its failure in terms of scores.

In the final section of the paragraph we will report some of the commercials developed for the Ford Focus model over the years, evaluating the characteristics contained in the message.

The purpose of this last section is to observe whether the company has developed a marketing strategy by exploiting the information contained in the reviews.

4.3.1 Average score detection

This analysis took into consideration all reviews of the Ford Focus model, in the period from 2005 to 2018, and calculated the average annual value of the scores released by consumers.

The Ford brand Focus model was chosen because this model is characterized by an average value of the scores detected in the interval considered equal to 3.467 to which is associated a variance equal to 1.836 one of the highest among all the models. Once the model under analysis was defined, the historical series of the average annual scores released by consumers was graphically represented.

Figure 8, Average score chart of Ford Focus model in the years 2005-2018

The aim of the graph in Figure 8 is to identify the causes underlying the high value of the variance; more specifically, we wanted to find the years in which the average score showed very evident peaks or drops.

From the representation of the historical series, in Figure 8, it is possible to note a long initial period, composed of the years 2005-2011, during which the brand follows a predominantly straight trend, characterized by an average value of just over 4.

However, at this time from the high valuations for the brand, a decisive collapse of the average value follows, which drops by more than one point in just one year (2011-2012). This period of collapse in valuations culminates in 2014, the year in which the model records the lowest average valuation value, equal to 2.3.

From this year onwards, a recovery of the model can be observed, ending in 2018 when the average score of the model approaches the scores recorded in the years 2005-2011.

Starting from these first revelations, it is natural to ask what caused this dramatic drop in scores for the Ford Focus model.

However, since it was necessary to work on the comments to find the causes of the formation of this "hollow" in the scores, it was considered appropriate to evaluate whether the curve of the average annual sentimental scores associated with the comments behaved in a similar way to that of the average annual scores.

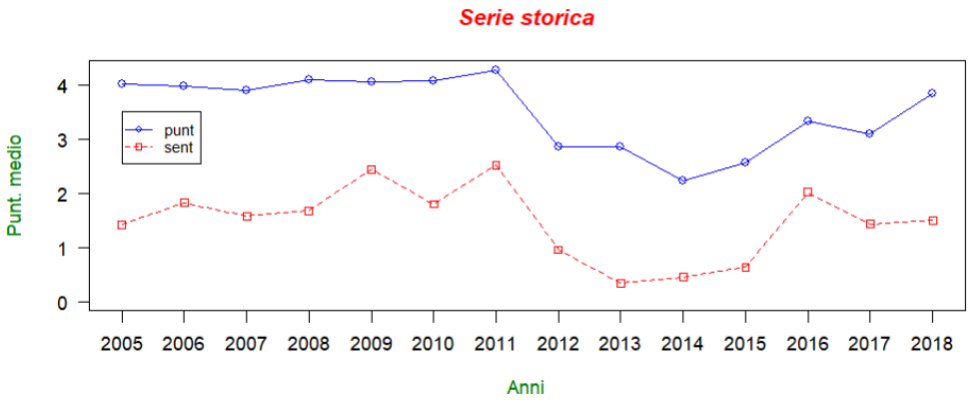

Figure 9, Graph of the historical series of scores of the Ford Focus model in the years 2005-2018. The blue line represents the curve of the average scores released by consumers in each year, while the red line represents the curve of the average annual sentimental scores

Figure 10, Graph of the historical series of scaled scores of the Ford Focus model in the years 2005-2018. The blue line represents the curve of the average scaled scores released by consumers in each year, while the red line represents the curve of the average annual scaled sentimental scores.

The graphs in Figures 9 and 10 aim to evaluate the presence of a similar behavior between the line of average scores and the average sentimental scores within the range of years considered.

In the graph in Figure 9, the sentimental scores are reported in their original scale,whereas, in Figure 10, the sentimental scores assigned by consumers have been scaled in order to make the differences between these two historical series more visually evident.

From these graphs, as expected, it can be noted that the sentimental scores show a similar trend to that of the scores assigned by consumers.

It is therefore noticeable that the collapse of the Ford Focus model is detected by both the comments and the ratings.

4.3.2 Detection and analysis of the most used frequencies and commercials

Having assessed the reliability of the use of comments in the study of the trend of the average score curve, we moved on to operationally set up an analysis of the positive and negative expressions most used by consumers in each year.

This process has the task of studying the evolution and alternation of characteristics considered positive and negative by consumers over the years, with the aim of finding the material causes that led to the lowering of the average value of the scores.

Using the concept of subtexts, introduced in the analysis of the previous question, the comments were divided into subtexts in order to divide each comment into sentences limited to a few words in length.

The score was then evaluated sentimental associated with each sentence, on the basis of which the sentences were divided into positive and negative.

Finally, the most frequently used expressions in the set of positive and negative subtexts were extracted using the idf weighting method.

The choice to calculate sentimental scores associated with subtexts rather than entire comments is intended to optimize and refine the search for positive and negative characteristics.

Specifically, we wanted to avoid the risk of incorrectly classifying an expression based on the score assigned to the comment.

For example, using sentiment scores associated with comments risks classifying a negative characteristic such as “noise” as positive just because the overall score of the comment is positive.

By dividing the text into blocks, we can limit this problem and make the method of identifying positive and negative characteristics more reliable.

The algorithm for identifying the most frequently reported positive and negative features in the comments has been improved. The output of the frequency analysis with tf-idf weighting has been reported in Table 11.

Year	Positive	Negative
2005	Drive, Fun, Interior	Transmission
2006	Comfort, Price, Look	Interior, Door
2007	Handle, Comfort	Tire
2008	Handle, Interior, Drive	Transmission, Tire
2009	MPG, Price, Sync	Interior, Small
2010	MPG, Sync	Interior, Steering
2011	MPG, Drive, Handle	Interior, Transmission
2012	Interior, Speed	Transmission, Package, Back_seat
2013	Interior, MPG, Handle	Transmission, Back_seat, Package
2014	Look, Comfort, Mpg	Transmission, Sync, Feature
2015	Interior, Feature, Quality, Sync	Speed, Small
2016	Interior, Package, Quality, Handle	Acceleration, Transmission, Shakes
2017	Comfort, Highway, Fun	Engine, Performance
2018	Comfort, Handle, Feel	Seat, Acceleration

Table 11Table of the most used positive and negative characteristics in each year, detected through the IDF method and with a division between positive and negative comments based on the calculation of the sentimental scores of the subtexts.
The first column contains the reference years. The second and third columns contain the highest frequency positive and negative characteristics, respectively.

In Table 11 it was possible to report the characteristics that have characterized the Ford Focus model over the years.

Considering the year before the drastic drop in scores, that is 2011, we find “Mpg” “Drive” and “Handle”, characteristics that were always present in the years between 2005-2011.It can be noted that these three aspects of the car were the most reported among the positive aspects of the Ford Focus model in the years characterized by the highest scores, this suggests the importance that consumers associate with these aspects.

It is possible to say that “Mpg”, “Drive” and “Handle” are the attributes that customers identify as the model's strengths and therefore aspects that have certainly led to the assignment of the high average scores typical of that period.

In 2011, each of these “historical” characteristics finds a place among the attributes evaluated as positive, while “Transmission” and “Interior” are evaluated as negative.

In 2012 the characteristics perceived as advantages and disadvantages change, in this version of the Ford Focus model it obtains an average score of 3 and among the positive characteristics of the car the historical terms “Mpg”

“Drive” and “Handle” are no longer found but rather the terms “Anterior” and “Speed”.
The 2012 version of the Ford Focus also presents new problems; in fact, the transmission will be joined by defects or dissatisfactions related to the rear seats and the "packages", various forms of software implementation accessories such as, for example, automatic parking and assisted driving

Analyzing Table 11, it can be stated that the cause that determined the collapse of the scores of the Ford Focus model in the year 2012 was a revolution of the model also in its main attributes.
The car has been improved in the quality of interior and engine performance, losing the perception of a car with good efficiency in terms of kilometers per liter of petrol.
A technological renewal was also introduced through the introduction of driver assistance service packages, which was not well received by consumers who reported these packages among the negative features.

Having completed this first evaluation of the table of the most used annual expressions,curiosity arose to detect and study the marketing operations that were adopted to ensure the recovery of the Ford Focus model.
The aim of this analysis is to see how the Ford company faced the collapse scores and evaluate the marketing activity carried out in relation to the defects that consumers have detected in the car.
Specifically, we want to see if it is possible to run this type of dataset using data extracted from comments through text mining and sentiment analysis and evaluating the commercials developed in the years considered.
From a practical point of view they were commercials from three different years that were taken into consideration: 2008, 2012 and 2017; searching for them within the multimedia content platform “Youtube”.

The texts of the commercials are reported in Table 12

Year	Advertising message
2008	A man lost in the savannah for years is found and reintroduced into society. He is dressed, his hair is done and he is made elegant, but despite his appearance the man does not stop roaring like a lion in the savannah. The advert ends with a voiceover, referring to the Ford Focus, saying: “they say it has changed completely,or almost”.
2012	Fascinating and fuel-efficient aerodynamic design, parking automatic, self-closing air vents for reduced fuel consumption and increased mpg
2017	We all drive, some just for the fun of it

Table 12,Table of contents reported within the advertising messages of the years 2008, 2012, 2017. The first column shows the reference years. The second column shows the contents of the spots.

In 2008, as can be seen in Table 12, the first important revolution of the Ford Focus model in the years considered occurred.

This version of the model features greater attention to look and design of the interior, as can be seen from Table 11.

This year's commercial highlights the cosmetic changes made to the Ford Focus model, comparing the car to a savage who is made more elegant and socially better off.

The machine is made aesthetically more appreciable and suited to the city environment, without losing the “wild” performances that distinguish it.

With this message the company has efficiently managed to renew and modernize the aesthetics of the Ford Focus model, reassuring its customers about possible concerns about performance drops and therefore about alterations to the mechanics of the model.

In 2012, a new revolution was introduced which brought a greater challenge, namely the introduction of functions and software packages and driving assistance, which automate the model.

This kind of innovation is introduced with a different approach than the one used in the 2008 commercial.

Rather than reassuring the consumer, the message presented aims to surprise the customer by listing the innovative features that the model has acquired.
This marketing strategy, together with possible technical problems related to new technologies, generally called sync, did not increase the average scores, but rather made them drop by more than one point.
The technological innovations introduced, together with defects associated with the new Transmission, damaged the scores and the image of the product even in the following years.
By consulting Table 11, it is possible to see how in the years 2015 and 2016 the Sync technology, the various packages and the innovative technological functions have passed into the positive attributes of the car, denoting their improvement and therefore the resolution of any initial defects.
Finally, in 2017 the company brings back in the commercial one of the most important aspects that was associated with the model in the years 2005-2011, that is, “Fun to drive”, recalling the pleasantness of the driving experience.
It can be concluded that the Ford brand has certainly been able to exploit the information contained in the feedback from its customers, both to identify the defects detected by customers and to find the characteristics to highlight in marketing activities.

4.3.3 Conclusion

It can be concluded that thanks to this analysis it was possible to identify the positive and negative characteristics that characterized the evaluations associated with the model in the years 2005-2018.

It was possible to evaluate which attributes caused the collapse of the scores in the year 2012 and it was possible to trace the marketing activities used by the Ford brand to recover from the shock.
By using text mining and sentiment analysis functions, it is possible for a company to get various information about your product.
Specifically, it makes it possible to:

1. Identify the main benefits and disadvantages that consumers find in the product.
2. Trace the characteristics of the car that led to a collapse

of the scores.

3. Detects consumer reaction to any product changes and innovations.

Furthermore, by comparing this information with the commercials used by other companies over the years, a brand would be able to study the marketing tactics of its competitors and learn from the mistakes of other companies.

Specifically, it was possible to analyze the marketing strategy that was used to deal with technological innovation, a situation that can be difficult for marketing to deal with.

5. Conclusions

The aim of this paper was to demonstrate and expose the potential of text mining and sentiment analysis, in terms of extracting useful information for market analysis.

This objective was concretely represented by the formulation of answers to the main questions of each paragraph of chapter four, with the final aim of obtaining a complete overview of the universe of the automotive world in the eyes of consumers, through their own experiences and evaluations.

Starting from comments unguided and evaluations expressed on a scale from 0 to 5 it was possible to set up the following analyses.

It was possible to trace the main characteristics of different models, highlighting their respective strengths and weaknesses, as well as the main problems, developing a summary of the market considered.

It was possible to evaluate the impact that the sentimental score associated with each characteristic had on the final evaluation of the model, identifying the most relevant and most influential characteristics.

Finally, the history of a model was analyzed from the consumer's point of view.and its needs, evaluating the characteristics that have defined its success or decline.

All these analyses were carried out using the Rstudio software, through which

codes in R language with Sentiment analysis and text mining functions were developed, working on reviews and comments.

6. Bibliography

W Medhat, A Hassan, H Korashy - Ain Shams engineering journal,
2014 M Wankhade, ACS Rao, C Kulkarni - Artificial intelligence
review, 2022
A Hotho, A Nürberger, G Paaß - Journal for Language Technology and
computational, 2005 U Kuckartz - Qualitative Text Analysis, 2013
NJ Horton, K Kleinman - Using R and RStudio for Data Management, Statistical
Analysis, and Graphic, 2015.