Dipartimento di Impresa e Management

Corso di Laurea Magistrale in Marketing Analytics & Metrics

Cattedra di Marketing Metrics

IMPLEMENTING SENTIMENT ANALYSIS TO ASSESS THE PERCEPTION OF POLARIZING PRODUCT: THE "TESLA CYBERTRUCK" CASE

RELATORE

Chiar.mo Prof. Michele Costabile

CANDIDATO

Federico Ferro

Matricola N. 704811

CORRELATORE

Chiar.mo Prof. Piermario Tedeschi

ANNO ACCADEMICO 2019/2020

Dedicata alla mia famiglia

Summary

Sharing opinions is an activity that defines social life of humans. People express thoughts to communicate their feelings about events, other people and objects that surround them. These can assume the shape of different emotions, according to how the input impacted and was elaborated by the subject. Thus, an opinion can be seen as a form of elaboration of taste made by a subject towards an object to which he found himself in contact with. Once an opinion is formed, sharing is a subsequent step that allows people to be in contact and thus giving sense to human as a social living being. Thus, expressing an opinion is sometimes a way to express oneself, to tray relating to people in the social fabric. The evolution that sharing have had in modern era is unprecedented, and the main point of this progress is given by the exponential increase of the audience that can elaborate a certain message expressed by one subject. In other terms, nowadays one sender has millions and millions of receivers: if this is multiplied by each subject expressing a thought, the number generated would be just huge. The tool that mostly influenced the change in people's communication and society in general is Internet, particularly in the form of social media. These are actual platforms that were built for sharing a multitude of information. Everything is this media is focused on allowing each user to share contents, from posts to comments. In some case and for some entities, an opinion can represent a source of value, as it is inserted into a process where its role is giving directions towards the achievement of a certain output.

Index

Figures List

Tables list

Introduction

Sharing opinions is an activity that defines social life of humans. People express thoughts to communicate their feelings about events, other people and objects that surround them. These can assume the shape of different emotions, according to how the input impacted and was elaborated by the subject. Thus, an opinion can be seen as a form of elaboration of taste made by a subject towards an object to which he found himself in contact with. Once an opinion is formed, sharing is a subsequent step that allows people to be in contact and thus giving sense to human as a social living being. Thus, expressing an opinion is sometimes a way to express oneself, to tray relating to people in the social fabric. The evolution that sharing have had in modern era is unprecedented, and the main point of this progress is given by the exponential increase of the audience

that can elaborate a certain message expressed by one subject. In other terms, nowadays one sender has millions and millions of receivers: if this is multiplied by each subject expressing a thought, the number generated would be just huge. The tool that mostly influenced the change in people's communication and society in general is Internet, particularly in the form of social media. These are actual platforms that were built for sharing a multitude of information. Everything is this media is focused on allowing each user to share contents, from posts to comments. In some case and for some entities, an opinion can represent a source of value, as it is inserted into a process where its role is giving directions towards the achievement of a certain output. This is the case in which an opinion turns into a feedback. In an economical sense, the subject turns into a potential customer and the feedback is directed to the evaluation of a product or service. This information is then received by the company that provided these elements and can used to understand the reactions towards them. Naturally, this system is prior to the advent of the social network era, but the changes brought by them are what made the real difference. In particular, two factors determined the revolution: the way the information is amplified and the way it is shared unfiltered. The former allows to think in terms of quantity, the latter in terms of quality. The amplification is what makes the process big and so why the businesses are more and more prone to analyse bigger and bigger data: for some aspects, in terms of opinions, the user needs to be thought as a cross-cultural and cross-national entity nowadays, with all its pros, cons and distinctions. In the era of social network domains, the communication is potentially worldwide, especially for big and renewed companies. Moreover, there is no "business filtering" when expressing feedbacks: in social networks, in most of the cases, users are not pushed by companies to express their opinions, neither are participating to an experiment: what they share is somewhat as close as possible to what they actually think about the object of discussion. These are what make social media information such valuable, in quantitative and qualitative terms, for businesses. That is also why the attention towards their analysis and so the implementation of specific computational developments is more and more increasing.

In particular, some programming tools were already in place that analysed big amount of information in form of comments and opinions, in particular by understanding the meaning of a word or sentence in a scalable way. However, these new kinds of sources have pushed the programs to rethink some aspects, especially in terms of language expressions, in order to adapt the system to the reception of the above described value. In particular, a form of computational analysis called text mining, that will be further described, have been adapted to the new and valuable source of textual information namely social media, leading to the creation of a new field of study whose name is social media analysis. Therefore, when the aim is to capture the feelings that lay behind a text, the focus of the mining shift towards the study of a specific emotion expressed by the user: this is when sentiment analysis comes into play. These aspects are then applied to a specific business or product, thus giving a structure and business value to the entire process of analysis. Ultimately, this is what the following study is about. The first chapter focuses on identifying the main elements and factors that characterize text mining, narrowing the field of research to the sentiment analysis and opinion mining in social media contexts like Twitter. The second chapter presents the product to which the analysis is addressed, namely the Tesla Cybertruck, an electric pickup presented by Tesla Inc. on November 21st, 2019, that polarized

people, critics and fans for its peculiar shape and features, together with its controversial unveiling. The third chapter applies some of sentiment analysis techniques to the Cybertruck, with the findings that were divided into four round and take into account different aspects and levels of study.

Elements of Text Mining

The study and analysis of written text allows to extract important knowledge that can be used to derive meaningful findings for businesses and researches. To achieve this goal, different approaches can be implemented, but they should all share some common steps that are fundamental in order to select relevant contents and discard non useful information. The chapter firstly highlights the importance that these types of information, known as unstructured data, are gaining throughout the years, considering also the vast development of different platforms where users, companies and brands in general can share knowledges through text. Moreover, some crucial aspects of the process of the extraction of information are presented, focusing on the main steps of analysis from data collection to final visualization. In addition, examples of algorithms and approaches, that are taken from machine learning fields and dictionary-based ones, are shown in a wider focus on Natural Language Processing, by this defining the ways through which human language is actually understood and then processed by machines and computers. One of the main scopes of the main analysis of text mining, namely opinion mining or sentiment analysis, that deals with capturing the actual emotion and thoughts about different topics, is further discussed. The focus then shifts to the discussion over particular domain for text mining studies, namely social network. These in fact represents one of the abovementioned platforms where user generated contents can be usefully and meaningfully levered, if not the main one, as its growth in popularity and subscriptions rose exponentially in the last decade. In them, a huge amount of written information is posted and shared every day by people and potential customers, thus explaining its importance for this field of study. Although there are different social media sources, one of the main relevant for text mining in undoubtedly Twitter, since its core functionality is strongly linked to the diffusion posts and comments in the written, that in this domain are also knowns as “tweets”.

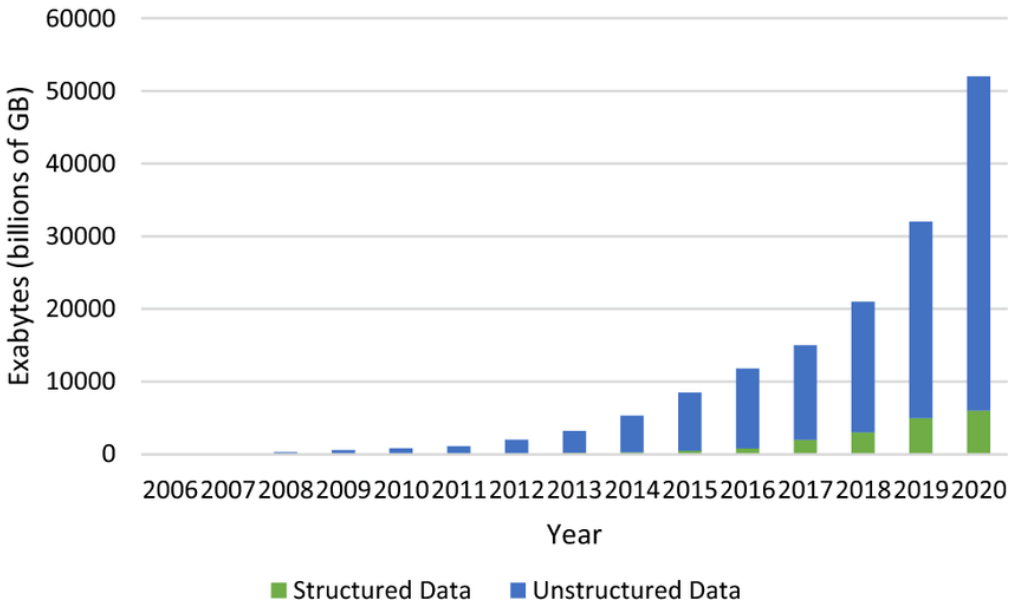
Different studies have used text mining techniques for numerous scopes in different domains, sometimes adopting an already existing method and others developing new ways of knowledge extraction form text. Some of them are presented in the final section of this chapter.

Unstructured Data and Text Mining

The information that come in the form of text is not only considerably relevant for its amount, but also for the undoubted quality and accuracy that is derived by using a written form. In fact, text allows to give thoughts, ideas and emotion an accurate shape, thus enhancing the level of understanding of the elements at issue. Writing allows the communication among human beings, and nowadays the way it is done drastically mutated because of the advent of networks and internet communities where information interchanged proliferate. Undoubtedly, the social media’s era constitutes a huge step towards a global human connection, and this link is built upon the opportunity for users to comment and share their opinions online with a billion of people over

a billion of topics. A relevant portion of this content is formed by consumers that comment over a particular product, service or campaign launched by a certain company. This latter can therefore deploy meaningful pieces of information to analyze in the form of text to measure the performances of its proposition and assets. A way to do this is by using text mining techniques. We can define text mining as the *"process of extracting interesting and non - trivial patterns from huge amount of text documents"* (Talib, et al., 2016). The interest is given by the fact that sometimes text allows to derive conclusions that cannot be drawn from any other means of communication. Hence, non-triviality is closely related to this, as relevant pieces of information are found through this kind of analysis. Technically, a document is not the only possible content that can be analysed through text mining. It is possible to broaden the concept, thus including different written forms. As a matter of fact, any text can constitute a form of textual unstructured data, where this is defined as a form of information that is not comprised in a pre-determined model or processable data frame, in contrast with the "structured data", namely data that are clearly comprised in a precise pattern or schema (Taylor, 2018). Therefore, the unstructured set also includes formats that go beyond textual contents and involve different elements, like multimedia contents. Moreover, it is important to mention that data online can be generated by every user, thus becoming an excellent non-filtered channel of knowledge for marketeers' analysis. Thinking about it, it is not difficult to find out that even the form of textual information that humans produce in their routinely life is enormous: emails, chats, website searches and the abovementioned social media posts are all different forms of unstructured textual data. Numerically speaking, textual nowadays constitutes 85% of business information (Hotho, et al., 2005) and unstructured data are growing at a rate of 55-65% per year (Taylor, 2018). The graph below shows that over 50000 Exabytes of unstructured data are available today, compared to about 5000 of structured.

Figure 1: The growth of structured versus unstructured data over the past decade.

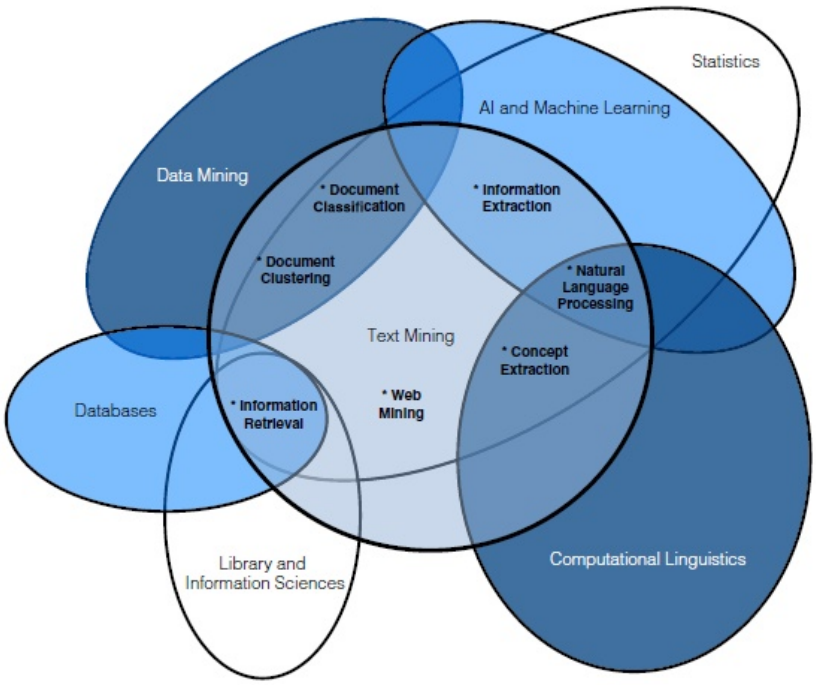


Retrieved from (Azad, et al., 2019)

Moreover, unstructured data can be defined as a form of Big Data. The reason why this definition is pertinent is that unstructured data score high values on the following three characteristics: Volume, due to their rapid growth in quantity of processable information; Variety, as the forms in which they can arise are various as seen above; Velocity, since creating these data is easy and very accessible to almost everyone in a rapid way. But what is more concerning from the business point of view, is that these factors combined can create added Value, thus constituting a real asset that can enhance company's intelligence (Taylor, 2017). What needs also to be taken into account is that the data produced are not only great in volume, but can host wide variables that help identifying the same unit: by taking a data frame as an example, the "big" factor is given by the number of rows, while the "wide" one by the number of columns that can help defining the row at issue. Thus, selecting the right features will mean deriving a frame that has more signal, that are valuable elements to retain, from noise, that on contrary are non-useful and avoidable ones (Yeomans, 2015). This clearly highlights the fact that managing to deal with this information properly could create and enormous advantages, explaining why this kind of techniques have become so crucial in business.

Text mining aims at processing text data whose principal characteristics, as seen above, are sparsity and high dimensionality (Aggarwal & Zhai, 2012). Therefore, "*Mining*" in this context literally means to gain new information from a certain database. It cannot be considered as a standing-alone discipline, as it in close contact with other disciplines like statistics, data mining, machine learning and computational linguistics, the latter describing a discipline that aims at *"understanding written and spoken language in a computational perspective"* (Schubert, 2019). Below is a Venn diagram that helps contextualizing text mining in the greater data and computer science area.

Figure 2: Venn diagram of the relation between text mining and other fields



Retrieved from (Talib, Hanif, Ayesha, & Fatima, 2016)

As the image shows, text mining is in the middle of different fields: all of them contribute to the objective of identifying useful patterns and deriving meaningful outcomes. Nevertheless, it has to be noticed that there is not an absolute and unique approach that needs adopted to achieve valuable results. This means that, despite text mining involve different fields, sometimes some of them do not need to be implemented. These can vary according to the objective of the study. What is crucial in this context is that a proper information is extracted., through systems and algorithms that analyse part of text with the aim of finding specific and already defined pieces of information. Thus, the objective can be defined as the "phase 0" of the entire process, since it delineates the structure and the method of work. For instance, a project conducted on reviews will require different approaches than a study conducted on formal documents or books. What is also to consider in this phase, is to understand if an automated text analysis is appropriate, which advantages can bring and what can discover more than other methods (Humphreys & Wang, 2018). For instance, text mining can bring additional knowledge when the goal is to identify patterns in the text structure or if the study is verted to an overtime comparison of different elements that are extracted by text. However, in order to reach the abovementioned goal, while text mining, some general steps needs to be followed, then different techniques and algorithms can be applied. Text mining is implemented through the following phases (Chang, et al., 2018):

1. Gather: the very first step is the collection of useful unstructured data to be processed. As seen above, various sources of this kind can be used, ranging from papers, book, documents, pdf files to social media posts. It can vary according to the different goals of the research. For instance, a study on social network would require a collection of different posts and comments of users, so using a social media sources would be more indicated for that.
    1.1. Information Retrieval: this phase is crucial in order to narrow the field of research. In here, the aim is to find answer to a specific question (Hotho, et al., 2005), namely "query", made by the user. The answer is retrieved from already-stored documents, information and catalogues. Hence, information retrieval system aims at showing the most relevant results for the query made levering also on its semantic and syntax, thus optimizing its utility for the user (Kahl, 2017). This is also the mechanism that underlays search engines, where the query has the form of keyword typed in the search bar. In particular, for a specific request, an algorithm automatically processes information available on the Internet and gives back the best match to the related query, even allowing to switch to disambiguation meanings if necessary. Even social media retrieval can work similarly to keyword search, as user can retrieve specific posts containing a desired word or hashtag queries.
2. Pre-process: once data are collected, they need to be pre-processed in a way that non useful information are discarded, allowing the algorithm and programming tools to process them with better qualitative results. In addition, this step allows to remove elements that do not have a valence for a text mining study. Hence, here is where raw data are converted into a more structured frame (Biswas, 2014). In order to do so, specific techniques are applied. The following are some of them:
    2.1. Stop words removal: this allows to remove words like articles, pronouns and prepositions, html links from the text data. Even words that occur very frequently or seldom can be deleted if necessary. The

reason why this is made is that the meanings that these elements embody are not relevant for further processing, so it is better to get rid of them. The user is free to choose which words to remove or not, but predefined "dictionaries" help to individuate some of the most frequent stop words. Therefore, especially when it comes to social media domains, more elements can be considered as stop words: URL, letter repetition and usernames are examples of them.

2.2. Stemming: The process allows to reduce words into their base form, or root. A stem is defined as a natural set of words that have equal or similar meaning (Hotho, et al., 2005). For instance, plural forms of nouns and past forms of verbs can be stripped into their root form through stemming, thus improving the tidiness of the data frame to build. So, "working", "worked" and "works" are all converted into "work".

2.3. Lemmatization: this technique returns word into their "lemma", or dictionary form. This is a more refined process as it is able to group synonyms into a single word and convert verb into the infinite tense; for instance, "best" changes to "good" and "went" changes to "go" (Yse, 2019). Although lemmatization provides better results in theory, it is not often used as every word needs to be tagged into their part of speech, namely their category in a phrase (Hotho, et al., 2005).

2.4. Tokenization: the step is needed to split a text into its unit forms, or tokens. In this way, a sentence like "I am working" is separated as "I" "am" "working". Tokenization is useful for determined further processing, as it allows to build a document dictionary, namely the set of terms contained in a certain document. To express the concept more technically, *"Let D be the set of documents and* $T = \{t_1, \ldots, t_m\}$ *be the dictionary, i.e. the set of all different terms occurring in D, then the absolute frequency of term* $t \in T$ *in document* $d \in D$ *is given by tf(d, t). We denote the term vectors* $\vec{t} = (tf(d, t_1), \ldots, tf(d, t_m))$*"* (Hotho, et al., 2005). However, splitting a sentence into its unit parts means that the original phrase semantic is not taken into account. Thus, its adoption depends on which approach is more useful for the study.

2.5. Part-of-speech tagging: once text is tokenized, another available option for the analysis is to assign the grammatical category to each word, namely its role in the construction of the sentence. This process is called as part-of-speech tagging or POS tagging (MonkeyLearn, 2019). In this way, more meaningful structure can be derived from the text, allowing to deepen the level of research if necessary.

3. Index: in this phase, data are transformed, as the text is encoded into a processable form for algorithm. Technically, (De Moura, 2009) defined text indexing as *"the act of processing a text in order to extract statistics considered important for representing the information available and/or to allow fast search on its content"*. For instance, the data frame obtained can be vectorized, namely converted into a numeric form called integer. In fact, *"most text mining approaches are based on the idea that a text document can be represented by a set of words, i.e. a text document is described based on the set of words contained in it (bag-of-words representation)* (Hotho, et al., 2005)." Bag-of-words is thus an example of data

transformation, that counts the frequency of a word in the document using a document-matrix (Shetty, 2018). It can also be useful to better understand the relevance and importance of specific terms for certain documents. More specifically, one of the approaches used to assess this is the vector space model one, that allows to derive the Term Frequency Inverse Document Frequency (TFIDF) method. This usually works better for machine learning tasks (Castañón, 2019) and is defined through the derivation of word's weight: *"Thus a weight w(d, t) for a term t in document d is computed by term frequency tf(d, t) times inverse document frequency idf(t), which describes the term specificity within the document collection*. […] *Term frequency and inverse document frequency are defined as idf (t) := log(N/n$_t$)"* (Hotho, et al., 2005). It therefore allows to index the data better, as terms are easily searchable. The space that word vectors occupy can also be optimized through a particular representation called word embedding. Basically, what word embedding does is to put similar words closer in the space. An example would clarify that:
"Consider the following similar sentences: Have a good day and Have a great day. They hardly have different meaning. If we construct an exhaustive vocabulary (let's call it V), it would have V = {Have, a, good, great, day}. Now, let us create a one-hot encoded vector for each of these words in V. Length of our one-hot encoded vector would be equal to the size of V (=5). We would have a vector of zeros except for the element at the index representing the corresponding word in the vocabulary. That particular element would be one. The encodings below would explain this better. Have = [1,0,0,0,0]`; a=[0,1,0,0,0]` ; good=[0,0,1,0,0]` ; great=[0,0,0,1,0]` ; day=[0,0,0,0,1]` (` represents transpose) If we try to visualize these encodings, we can think of a 5 dimensional space, where each word occupies one of the dimensions and has nothing to do with the rest (no projection along the other dimensions). This means 'good' and 'great' are as different as 'day' and 'have', which is not true. Our objective is to have words with similar context occupy close spatial positions. Mathematically, the cosine of the angle between such vectors should be close to 1, i.e. angle close to 0." (Karani, 2018).

4. Mining: This phase is the one in which text mining process livens. In fact, it is in this step that information is extracted or "mined" from the dataset through specific instructions. To do so, text mining levers algorithm and programming tools arising from Natural Language Processing or NLP intelligence, namely *"an interdisciplinary field whose goal is to analyze and understand human languages"* (Jothilakshmi & Gudivada†, 2016). As showed in the Venn diagram, Natural Language Processing is situated in an area that comprehends statistics, Artificial Intelligence and machine learning, computational linguistics and naturally text mining. This explains why this step stands at the core of the process. NLP will be better discussed further on.
5. Analysis: The dataset is now ready to be analysed, so at this stage a proper knowledge can be implemented, shared and visualized for different scopes. This is the phase where human intellect is necessary to derive meaningful results from the analysis conducted. The findings can thus be used, among the several scopes, for making comparisons, inspecting correlation, identifying pattern association or inferring predictions on

the elements considered (Humphreys & Wang, 2018). It is possible to define this stage as the catalyst of the entire process.

Figure 3: Text mining process

Retrieved from *fosteropenscience.eu, 2018*

Fields of application and Business relevance

Text mining is used for a plethora of scopes in different fields. The reason why they all share this technique is that, in many sectors, the need to analyse a vast amount of information to derive a desired outcome or to identify relevant entities is vivid. It can be seen as a tool that helps humans processing information, thus overcoming time and efficiency constraint and consequently making it possible to adapt results on a large scale for different tasks. For instance, text analysis can be crucial if properly used in different sectors, like the medical one: basing on patient's speech or electronic records, through text mining it is possible to derive useful patterns to make a medical diagnosis or to extract treatment outcomes. Moreover, it can be implemented for anti-spam purposes, to identify emails or posts, to assess trends in a specific geographical area, to cluster documents and research papers, thus helping human resources to select the best candidate for an open position, or even for warranty claims (Hotho, et al., 2005). From a business perspective, text mining allows to improve analysis in crucial departments and research areas. For instance, customer relationship management can rely on text analysis for processing and grouping a multitude of customer feedbacks and reviews over a product or service, in order to find a unified answer for similar needs. Moreover, it is useful when it comes to find information about its own company, as it allows to highlight useful patterns that can eventually constitute a strength or weakness of the firm. (Yse, 2019) (Biswas, 2014). Mining consumers and competitors could be an application too, especially when the aim is to gain competitive advantage in the market or when a specific campaign needs to be monitored or launched. Therefore, text mining can also assess why consumers choose a particular product and which are the features that like or dislike the most (Wonderflow, 2019). Another application, that will be further described, takes into account the use of social media as source of data collection. In fact, as seen before, the production of data on these media is incredibly massive and deals with users that can potentially turn into costumer. When the aim is to inspect consumers, thus text mining is implemented for consumer research, there are four areas that makes text analysis particularly valuable for researchers (Humphreys & Wang, 2018): one of them is certainly the attention, that tend to be captured by

words and make people more respondent to issues. The expression is in fact delivered by the use of words, and it reaches the recipient in a direct and precise way. In a research context, attention can be measured through different ways, like word frequency, and evaluated overtime in order to extract valuable knowledge from the data at issue. Another area is the processing one: when syntax is examined, it is possible to find out that there are different ways to encode a message, and according to this, different outcomes can be achieved. A phrase can be more effective than another that use the same words just because of how this is encoded, or structured. Researchers can thus highlight this factor to discover deeper findings about consumers, a part to discover the tone and emotion of the message delivered. Social dynamics are another point, as language can reveal information about the relationships about subjects that are involved in a textual conversation. The analysis of pronouns or names can allow to infer more details of who lays behind a message, even when this does not carry an informative meaning but has the only objective to express a certain feeling about something expressed in the phrase or word. Lastly, language can release information about groups and cultures: it is sufficient to say that the expressions used in languages evolve together with the progress made throughout time, and they can vary according to different size of geographical space. So, the inspection of text and language can be also targeted to the evaluation of diversities given by time and space about a specific object of study.

Natural Language Processing

As seen above, levering Natural Language Processing is the crucial phase to analyse textual data and to conduct a study on text mining in general. NLP is defined as *"a field of Artificial Intelligence that gives the machines the ability to read, understand and derive meaning from human languages"* (Yse, 2019). Basically, in this phase, algorithms, that can be seen as "classifiers", are implemented in a way that they elaborate the content of interest and gives back a specific output, according to the analysis method chosen. More specifically, "*Text mining systems rely on algorithmic and heuristic approaches to consider distributions, frequent sets, and various associations of concepts at an inter-document level in an effort to enable a user to discover the nature and relationships of concepts as reflected in the document collection as a whole"* (Biswas, 2014). An example of this can be found in sentiment analysis, where algorithm is capable of identifying a precise sentiment attached to a word by linking it to a pre-set dictionary. NLP is a form of machine learning as the process looks at patterns in data and, thanks to a previous instructions or examples if provided, can match an element with its correct meaning or definition (Expert System Team, 2017). To perform a task, NLP intelligence can be set for different paths, that differ accordingly to the degree of previous knowledge of the machine. In general, there are different task that NLP system can perform, and they vary according to their specificity and deepness. These are (Zhai & Massung, 2016):

1. Lexical analysis: at this stage of analysis, the objective is to derive the meaning of a single word, regardless of its position in a phrase. This level can be implemented by separating each unit through, as seen before, mechanisms like tokenization. So, the representation of text will be as string of characters or a sequence of words. What represent a challenge at this stage, is the correct interpretation of word level ambiguity: in fact, a single word can have different meaning, thus its sense is ambiguous

2. Syntactic analysis: the aim shifts here to the finding of interrelatedness among each word in the sentence, thus allowing to discover new information about the text taken into account and more specifically its representation through a syntactical structure. What needed to be considered here is that syntactic ambiguity can occur: in fact, the same sentence can have different interpretation that, despite easily understandable for humans, can create difficulties in a computational context.
3. Semantic analysis: based on the previous two levels, semantic analysis' purpose is to capture the meaning of the entire sentence. Advanced mechanisms can understand the topics under discussion by analysing and identifying the most relevant elements in the text, and some of them can also recognise the disambiguation of a certain word if needed (Expert System Team, 2017). The representation of text is made by highlighting the entities of a phrase and its relation. For instance, the system can recognize the word "car" as an object or the word "elephant" as an animal.
4. Purpose analysis: NLP techniques can also be able to determine the meaning of the context, namely its purpose in the communication. The system here can allow to represent the text as a speech act, so to derive the action meant to be communicated in a phrase.
5. Discourse analysis: in this stage, a group of concatenated sentences is analysed, each of them depending by one another. Hence, the analysis moves to a group of factors in which its order is a necessary element to consider.

Moreover, NLP embraces various ranges of techniques, that can be used to reach different goals and can vary according to the level of analysis that the research is willing to achieve. As an evidence of that, two different techniques take place in NLP: text classification, which is the process of assigning categories or tags to the set of unstructured data, and text extraction, that allows to extract relevant and already existing pieces of information from the unstructured data. These methods are not mutually exclusives and they can be adopted together when performing a text mining study. Examples for the former are (MonkeyLearn, 2019):

1. Sentiment analysis: formally, it is defined as *"an approach to natural language processing (NLP) that identifies the emotional tone behind a body of text."* (Rouse, 2019). In this way, it is possible to derive useful knowledge about not only the dataset, but also about the people that expressed that thought. Sentiment analysis will be discussed further on.
2. Topic analysis: it allows to understand which are the main topics of conversation within a text. By the use of specific techniques, arguments of interest are derived that allow to recognize the object of a specific text. Again, some ways to perform this analysis are described further on.
3. Intent detection: by using this technique, it is possible to understand the purpose or intention that lays behind what is written by people. Word recognition and its subsequent classification are steps to be performed when the aim is to implement this technique.

For what concerns text extraction, examples are (MonkeyLearn, 2019):

1. Keyword extraction: its aim is to individuate and extract the most relevant single word within a text, according to parameters like frequency. It is useful to give an overview of the dataset at issue.

2. Entity recognition: this represents a deeper level of extraction analysis, as it able to attach specific entities, like people or places, to words. For instance, if trained, the system would be recognizing "Italy" as a "country".
3. Word sense disambiguation: this technique is useful to discern the sense for ambiguous words, that can very their meaning according to different semantic contexts. Models need to be taught specifically for this task that is challenging from a computational point of view.

To perform these techniques and tasks, as seen above, sometimes NLP needs to be accompanied by classifiers. For NLP in text and opinion mining, classification methods fall into two main categories, both of them having different subset: Machine Learning Methods and Lexicon-based approach (Hemmatian & Sohrabi, 2019). These two aspect can also be divided by the fact that former adopts a "bottom-up approach", as they first inspect pattern in text and only in a second time a theoretical explanation is given, while the latter a "top-down approach", that deals with the inspection and analysis of occurrences of words based on a pre-defined dictionary or a specific set of rules (Humphreys & Wang, 2018).

Machine Learning Approaches

Specifically, machine learning should all perform good in three aspect: feature extraction, namely the ability to retain useful and processable variables or "features" in order to extract valuable knowledge from the set of data; regularization, that is finding a good weight for data in the model and thus collocating itself between a conservative approach, that makes highly generalizable inferences but discards useful information, and a complex one, that contrarily consider several variables and features even if they are considered as noise; cross validation, which is the accuracy of the model shown by its usefulness for different data (Yeomans, 2015).

The main subset for machine learning techniques of NLP are:

1. Supervised learning: in this case, the algorithm works in a way that it makes predictions based on instructions received in a first set, and then generates outcomes on a second set, from which information need to be extracted. We can define these two groups respectively as "training set" and "test set" (Sathees Kumar & Karthika, 2014). For instance, if a word is labelled as "x" or it has a certain score attached in the training set and the same word is present also in the test set, the algorithm will classify and give a score of "x" to that specific word in the test set. In particular, the algorithms that are used in supervised learning are called "classifiers", as their task is to classify the content in the test set based on the information available in the training set. More precisely, in this context, *"a classifier is a function* $f(\cdot)$ *that takes a feature vector as input and outputs a predicted label* $\hat{y} \in Y$*"* (Zhai & Massung, 2016). So, what the classifier does is to try guessing the relationship between an input and an output (Hemmatian & Sohrabi, 2019). There are plenty of supervised learning classifiers that can be implemented for text mining purposes. Regression-based classifiers are an example of these, that study the relationships between variables to predict a better outcome. In particular, there are two types of regression: linear regression, the classical model, used either to remove correlated variables, diminish the noise in the dataset and naturally to derive a significant output; logistic regression, mainly used in binary classification whose goal is to

weight variables in a function that is nonlinear. Instead, it is called logistic function (Rokad, 2019). Moreover, another example is given by the Naïve Bayes classifier, that levers on an application of Bayes' conditional probability theorem, namely $P(A|B) = \frac{p(A)p(B|A)}{p(B)}$, to attribute a certain class to a given document (Asiri, 2018); More specifically, what Naïve Bayes does is to create a probability distribution for the features taken into account over the class labels. Given a label y where $y \in Y$, and features x_i, in order to calculate the probability that x_i belongs to a certain label y, more formally $p(x_i|y)$, it is necessary to calculate $p(y|x_i)$ first as follows (Zhai & Massung, 2016):

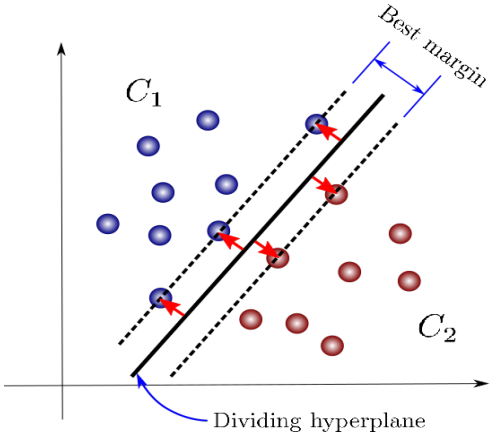
$$\hat{y} = \arg max_{y \in Y}\, p(y|x_1, \dots, x_n) =$$

$$\hat{y} = \arg max_{y \in Y} \frac{p(x_1, \dots, x_n|y)}{p(x_1, \dots, x_n)} =$$

$$\arg max_{y \in Y}\, p(y) \prod_{i=1}^{n} p(x_i|y)$$

Where the nominator was excluded because it did not affect the calculation and the independence of each features is assumed, thus the final score can be found by multiplying every single feature considered.
Moreover, Support Vector Machine or SVM, that builds a hyperplane from which classes are divided (Hemmatian & Sohrabi, 2019). Technically, SVM works in a way that it creates boundaries between features and tries to optimize the spaces in the areas considered in order to assess a better classification for new variables (Zhai & Massung, 2016). To perform SVM, categories needs to be converted into binary numerical data. So, for instance, in the example of email categorization into spam or non-spams, x would be the email taken into account, y=1 would identify spam while y=-1 a non-spam (Carrasco, 2019). There are two different ways of possible classification with SVM. The first one is linear, so by using the properties of linear combination it divides the observations in a hyperplane into two different categories; the second one is nonlinear, that is scaled on a multidimensional space. Nevertheless, if the number of features is too large, nonlinear classification could not improve the performance of the model (Hsu, et al., 2016). The following figure clarifies the definition of linear SVM classification for binary classification problem:

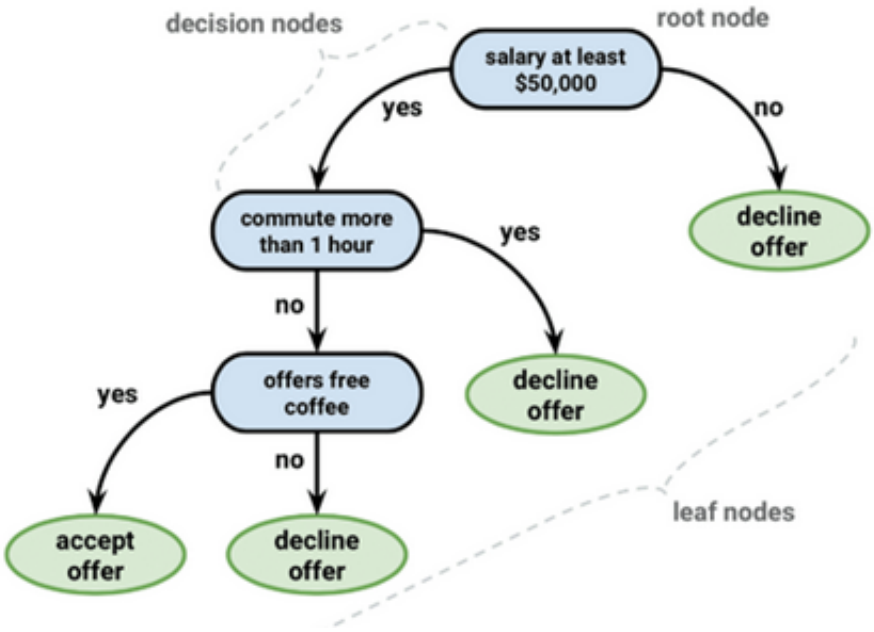
Figure 4: Graphical representation of SVM in a binary classification problem.



Retrieved from (Carrasco, 2019)

In this graph, the attributes are separated into two categories named C1 and C2 by a line that divides the hyperplane. The more optimized is the margin, the better the classification. Moreover, if the space considered was multidimensional, there would be another separator rather than the line above. Another supervised method per definition is the decision tree approach, which is a form of classification that works as a mutually exclusive if-then structure that, starting from a specified rule, uses a top-down approach to classify the set at issue. This one works as follows: *"For a training set M with labelled documents the word ti is selected, which can predict the class of the documents in the best way, e.g. by the information gain […]. Then M is partitioned into two subsets, the subset Mi+ with the documents containing ti, and the subset Mi− with the documents without ti. This procedure is recursively applied to Mi+ and Mi−. It stops if all documents in a subset belong to the same class Lc. It generates a tree of rules with an assignment to actual classes in the leaves"* (Hotho, et al., 2005). So, in decision trees, each variable is a decision node that be split into another decision node or a leaf node, that is the output of the model (Rokad, 2019). The figure below graphically represents an example of regression tree:

Figure 5: Decision tree structure



Retrieved from (Rokad, 2019)

Leaf nodes are the final green outputs. It is also important to notice the binary structure of the model, as the nodes are generated by a "yes" or "no" response and the output can either be "accept offer" or "decline offer".

Another used unsupervised system is called K-nearest neighbour or k-NN. This algorithm works in a way that it finds similar document of a query document. Once done this, k-NN assign to the query document the most common label of the similar documents (Zhai & Massung, 2016)

Having these characteristics, supervised methods are used for classification problems like image recognition (Salian, 2018) and they require more time to process at the training stage, but less for the outcome derivation.

1. Unsupervised learning: if the machine does not have reference set where an instruction tells if an outcome is right or wrong, it has to process the data by himself, trying to discover useful patterns among the dataset. For instance, the algorithm can group all the factors that share similar characteristics, as it happens in clustering methods, or infer a semantic relationship between two words, as it happens in Latent Semantic Indexing (Barba, 2019). These methods can lead to satisfactory results, and they require less time for the processing phase, but more on the outcomes' one. Moreover, dimensionality reduction system can be implemented in unsupervised approaches. The aim of this method is to inspect the dataset in order to select and retain only the most relevant features, thus simplifying the general structure without losing important pieces of information. In fact, there could be the case where some variables are not useful for the scope of the study, so sometimes it is better to just get rid of them. Additionally, this method can be implemented in order to select features that can be used in a regression or in a supervised approach. An example of a common dimensionality reduction algorithm is called Principal Component Analysis or PCA (Brownlee, 2019) (Edwards, 2018). In PCA, two phases are implemented, that are feature elimination and feature extraction. After the process of eliminating non useful variables for prediction is completed, PCA allows to extract features that, are all independent of one another, thus matching the assumption of independence of the linear model (Brems, 2017).

   Moreover, another dimensionality reduction model, one of the simplest and most used, is called "Latent Dirichlet Association", or LDA. This method is useful to extract relevant topics, which can be specified before the implementation, from selected documents, and work by following a logic given by conditional distribution. The assumption that lays under this model is each document to be analysed share the same topics, that are pre-identified, but with different distributions. More specifically, the aim is to find hidden variables, in this case topics, from observed variables, so documents and words in documents. This distribution is called posterior and is computed in the following way (Blei, 2012).

   The parameters are: $\beta_{1:K}$, that defines the topics distribution; θ_d, or topic proportions in the *dth* document and $\theta_{d,k}$, proportion of topic *k* in document *d*; z_d, namely the topic assignment for the *dth* document, and $z_{d,n}$, topic assignment for word *n* in document *d*; w_d, that are the observed words in the document d, with $w_{d,n}$ being the *nth* word in *dth* document.

   The posterior formula can thus be defined as:

$$\frac{p(\beta_{1:K};\, \theta_{1:D}; z_{1:D}; w_{1.D})}{p(w_d)},$$

   where

$$p(\beta_{1:K};\, \theta_{1:D}; z_{1:D}; w_{1.D}) = \prod_{i=1}^{K} p(\beta i) \prod_{i=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} (p z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1.K};\, z_{d,n}) \right)$$

As already touched on, clustering represents an alternative in unsupervised learning methods. Specifically, clustering is defined as *"the act of creating groups with different characteristics"* (Edwards, 2018). Since there is no label attached to the set, the algorithm finds hidden pattern on their own, trying to recognize similar characteristics in dataset and thus grouping the most similar units. The absence of restriction implies that there is no *a priori* number of clusters to be considered in the analysis. Thus, the optimal numbers of groups are found empirically. There are different types of clustering algorithms: one of the most used and popular is called k-Means, where k represents the chosen number of clusters. In this method, a random centre within the data is found. After that, each data point is assigned to the closest centre found. Subsequently, a new centre for the cluster is calculated, until centres do not or slightly change (Castañón, 2019). Another method is the hierarchical clustering. Similarly, it groups the closest and most similar points in a space into a cluster, but its main output is given by a dendrogram, namely a graph that shows the sequences performed to obtain the sets. The distance is usually calculated by the length of the straight line that is drawn from a cluster to another. This method of measuring distance is called Euclidean distance. (Bock, 2018).

2. Semi-supervised learning: this machine learning technique is based on a hybrid form of supervised-unsupervised, as it generally uses both labelled and unlabelled data as training set. The objective is thus already pre-defined, but the model has to learn the structures necessaries to organize the dataset in a way that the task can be achieved (Brownlee, 2019).

   Some other intermediate forms of learning have developed for further optimize algorithm results, like reinforcement learning, which works as a "trial-and-error" and "delayed reward" mechanism (Expert System Team, 2017).

Other algorithms can work either with a supervised or unsupervised approach. An example of these are Neural Networks, or NN. The name is given by the fact that the model is inspired from the neural networks or organisms in nature (Rokad, 2019). NN's objective is *"to capture non-linear patterns in data by adding layers of parameters to the model"* (Castañón, 2019). In NN, there are three layers, each of them collecting nodes called artificial neurons: the first one is called input layer, so this is basically the starting point of the process. Through edges, that are the connections between neurons, the information is transmitted to the hidden layer. The latter can be more than one and its task is to compute the weighted inputs. Finally, the last step is the output layer, where information is completely processed (Rokad, 2019). Neural Network structure is used to perform word embeddings calculations too (Karani, 2018). A Neural Network learns supervised if the desired output is known, otherwise it learns unsupervised (Fröhlich, 2004).

As seen, these classifiers do not perform equally in each domain and study. Thus, different evaluation of their effectiveness exists in order to better understand how a method or algorithm works in a determined dataset. In order to inspect this, different measures of evaluation can take place during the analysis. The most important measure is the accuracy, that is the number of correct predictions made by the classifier divided by the total number of predictions made. Accuracy can be defined as a default metric. If a two-class domain were considered, where the possible outcomes for classification are true positive, true negative, false positives and

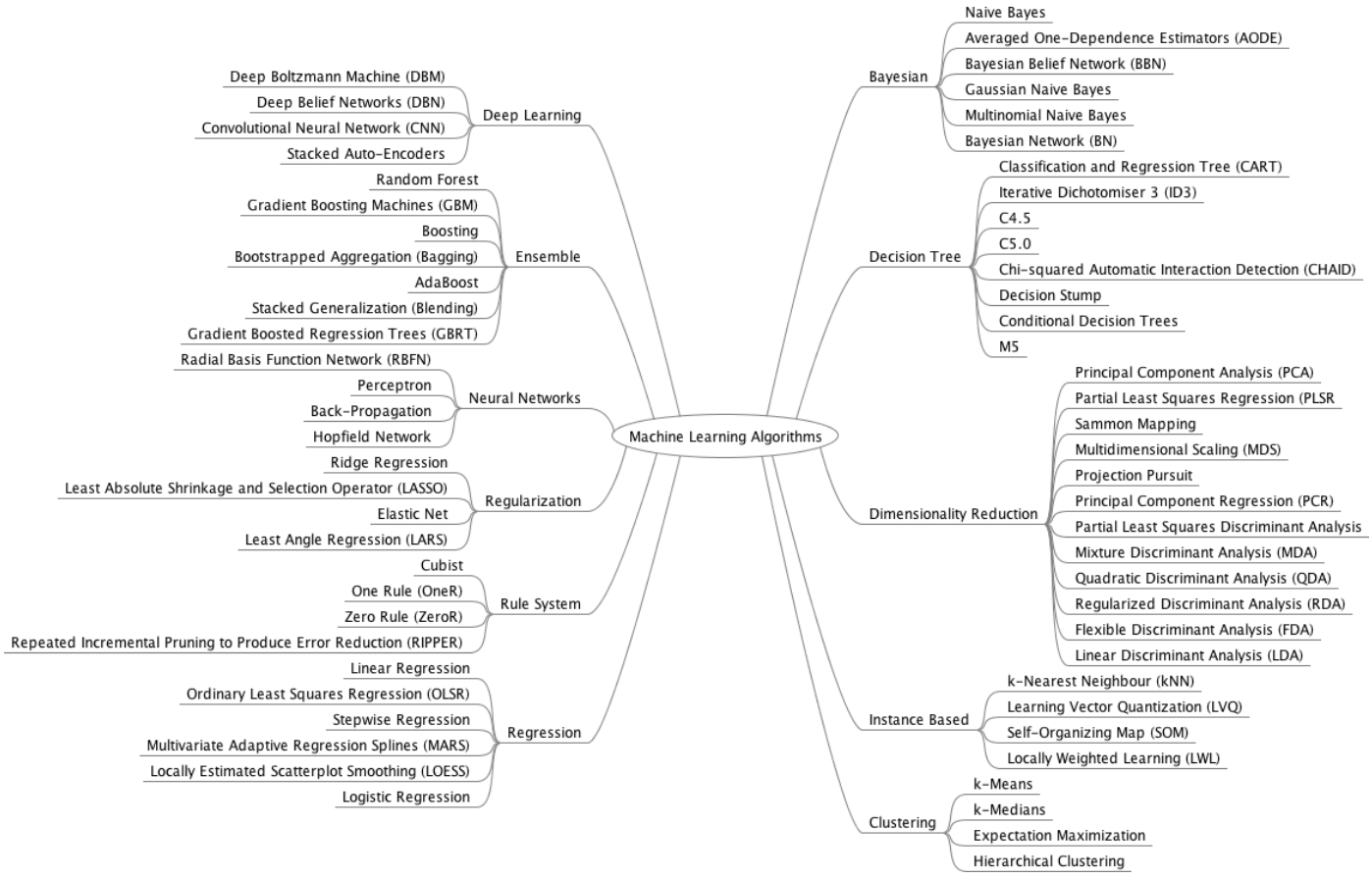
false negatives, accuracy can also be defined as the sum of true positives and true negatives on the sum of all the possible outcomes. Another metric considered for classifiers is called Precision. This is the measure the gives back the correct prediction out of the ones that were predicted to belong in a particular tag. In other terms, it can be considered as the true positives on the sum of true positives and false positives. Moreover, it is possible to consider the recall. This evaluation is made by the number correct predictions divided by the number of predictions that should have been considered as belonging to that given tag. Again, in other terms, the precision is the number of true positives on the sum of true positives and false negatives. Finally, the last evaluation considered for classifiers is F1. This is the harmonic means of precision and recall, where these two are considered as equally important. F1 is considered as a better performance of classifiers than accuracy (MonkeyLearn, 2019) (Fawcett, 2015) (Koehrsen, 2018). Below, these metrics are represented in formulas:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad Recall = \frac{True\ Positive}{True\ Positive + False\ Negatives} \quad F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

For the sake of completeness, the following figure graphically sums up different machine learning algorithms with their respective subdivisions.

Figure 6: Machine Learning Algorithms.



Retrieved from (Rokad, 2019)

Lexicon-based approach

Another way of studying text in NLP and opinion mining field is to use an approach based on the meaning of words, called lexicon-based approach. More specifically, this is a method that levers on the semantic orientation of words that contain information about polarity - so the ones that can express positive or negative feelings like adjective or adverbs - to assess a precise orientation of documents, using a pre-determined dictionary or a "first-hand" one, that contains a set of words labelled with a predefined polarity score (Hemmatian & Sohrabi, 2019). Generally, two ways of lexicon-based approach are used: dictionary-based approach, that searches for synonyms and antonyms of the opinion word at issue, and corpus-based approach, which considers the word as a part of a context, thus searching for other opinion words in the entire corpus. In addition, a hybrid approach of the two can be implemented (Hardeniya & Borikar, 2016). The advantages of such approach are its easiness of implementation and comprehension, the contribution that is given by sociologist and psychologists, that validated certain forms of dictionaries that thanks to that become "standardized", and the transparency of findings (Humphreys & Wang, 2018). Since it is an approach that is mainly implemented for sentiment analysis and opinion mining studies, more applications and characteristics will be further discussed in the specific section.

Opinion Mining

Having stated what stands above, it is pretty straightforward to understand that NLP stands at the base of Opinion Mining, as it is the real "engine" that makes the entire mining system work. The research and studies on opinion mining have grown exponentially throughout the last decade, following the evolution of social media and micro-blogging platforms and networks. A classic definition of it is the following: *"Opinion mining is extracting people's opinion from the web. It analyzes people's opinions, appraisals, attitudes, and emotions toward organizations, entities, person, issues, actions, topic and their attribute"* (Hemmatian & Sohrabi, 2019). Having said so, we can therefore highlight that opinion mining is a field of study that is deeply linked to the relationships among humans, that share what they think and are inspired and motivated by what others share. This social phenomenon was exasperated by the rise of different platforms, micro-blogging and social media platforms, that allowed to share an opinion to an incredibly vast audience. Thus, the Web evolved in a way that users are not passive content-seekers anymore, but they are actively involved in the system as content-creators (Liu, 2015). As the amount of information grows, the need of summarization becomes more and more relevant, making opinion mining crucial when it comes to social media analysis. Moreover, as seen previously, the urge of deriving patterns is vivid among subject towards which an opinion is formed, such as companies. More specifically, opinion mining analyses a text structure composed by five elements (Sun, et al., 2017):

$(e_{ij}, a_{ij}, s_{ijkl}, h_k, t_l)$,

where "e" is the ith entity, "a" is the jth aspect of the ith entity, "s" is the opinion or sentiment towards the aspect j of the entity i, which is made by an "h_k" opinion holder at a certain time "t_l". Nevertheless, not all the

elements are necessary in order to process the mining. To better understand this concept, an example is provided. If Marco says: "The design of this car is good", "Marco" is the opinion holder h_k, "design" is the aspect a_{ij} of the entity "car" e_{ij}, while "good" is the sentiment s_{ijkl}. Generally, entity, aspect and sentiment are the elements of major interest and are sufficient for the studies on opinion mining. As it showed, the study on the polarity of sentiment, called sentiment analysis, is one if not the main research area in this field. Specifically, sentiment analysis "*is a natural language processing or information extraction tasks that helps extract pro or anti opinions or feelings expressed by a writer in a document collection"* (Biswas, 2014). Usually, opinion mining and sentiment analysis are considered as synonyms, with the latter that is mainly addressed to the study of feelings and public opinion examination, but again, the borders are blurred (Hemmatian & Sohrabi, 2019) (Liu, 2015). In particular, sentiment is generally measured by its polarity and/or by the mood through which it is expressed.

Different levels of depth can be adopted for the study of sentiment analysis (Liu, 2015):

1. Document level: this represents the broadest level of analysis, sometimes considered as too coarse. Its aim is to assess the general polarity of the entire document, thus not taking into account the reasons behind certain results and keeping the outcomes to a surface level.
2. Sentence level: this approach goes deeper as it analyses the sentiment to the level of sentence, so it better comprehends the structure and the subjectivity of the phrase, thus assessing if an opinion or emotion is expressed or an objective fact is reported in it. Sometimes, the average sentiment score of each phrase is calculated to assess the general value for the entire set, thus achieving to conduct the analysis even to the document level.
3. Aspect level: the deepest level is the one that goes beyond the mere language syntax and aims at identifying the opinion and its target, namely the aspect of reference. In this way, a fine-grained feature approach can be implemented in order to find out a sentiment over singular elements. For this phase, it is really useful for the machine to know the semantic of the entire phrase, in order to understand the relationship among words.

The level choice is based on the different goals researchers wants to achieve with the analysis. Moreover, inside every level, various methodologies can be adopted. These recall in part the machine learning mechanisms described in NLP. In fact, the method can be machine learning-based, where training set is used to "train" a certain classifier, that is the algorithm whose task is to predict the outcome of the element in the testing set (Xu, et al., 2011). Each algorithm have different characteristics and level of accuracy: some of them are lazy learners, meaning that they store the training data and do not start until a test data is introduced, while others are eager learners, meaning that they start building a classification model before the introduction of test data. Example of the former is k-nearest neighbour, while examples for the latter are Naïve Bayes and Decision Tree (Asiri, 2018).

Moreover, the method can also be implemented as a form of lexicon-based approach, that assesses the polarity of a word of phrase via a lexicon, namely *"a dictionary of sentiment words and phrases with their polarities and strengths"* (Sun, et al., 2017). Through this mechanism, the text is generally represented in a bag of words

form, so every unit or string is made by a single word. After that, the sentiment score from the pre-defined dictionary is attached to the correspondent word in the data frame to analyse (Jurek, et al., 2015). For instance, a certain dictionary attaches a positive value to the word "good". If this word is found in the test set, it will receive the same positive score. Different systems of evaluations exist according to the dictionary used, and sometimes same words can have diverse polarities in different dictionaries. The lexicon-based method is faster and largely implemented, however its downside is that it does not take into account a potential target for the sentiment expressed. The most common case of lexicon limitation is given by negations: for instance, the word "good" taken alone can have a positive value, but if it is used after a negation like "not", its value should be reversed into a negative one. Even intensifiers like "very" could not be taken into account, even if they could adjust the sentiment score of the adjective that succeeds them. There are different types of already developed lexicons, each of them containing different numbers of labelled words. The most used in sentiment analysis studies are (Monsters, 2017):

1. SentiWordNet: this lexicon assigns a sentiment score to the synset – namely a set of synonims - of the lexical database of English words called WordNet (Princeton University, 2010). The scores attached can be three: positive, negative or neutral. These are derived by result combination of eight classifiers
2. SentiWords: in this lexicon, 155,000 English words are present, each of them having attached a polarity score that goes from -1 and 1.
3. AFINN: built between 2009 and 2011, the polarity score for the word here goes from -5, most negative score, to +5, most positive one
4. WordStat: this dictionary uses a different approach to sentiment measurement. In fact, a negative score is attached to either negative words that in the precedent three words do not contain negations, or to positive words that are preceded by negations within three words. For positive score, the mechanism is the same but from the positive perspective.
5. ANEW: the measurement in ANEW is made in terms of three factors: pleasure, arousal and dominance. It is both used for sentiment studies and attention ones (Bradley & Lang, 2017)
6. Whissell Dictionary of Affect in Language: the peculiarity of the lexicon at issue is that it does not classify the words based on their meaning, but rather to how they make people feel pleasant, active and to their ability to arouse imagery.
7. Linguistic Inquiry and Word Count: this dictionary has the characteristic to support different languages and to extract around 60 different categories for the word embedded
8. Multi-Perspective Question Answering MPQA Subjectivity Lexicon: MPQA contains about 8,222 words and phrases, labelled as subjective expressions, with their Part-Of-Speech tagging attached (Musto, et al., 2014)

These are only few examples of the lexicons created for the analysis of sentiment, and it contributes to understand its relevance in this field of study. More examples will be further presented. As seen, the limitations described above can be overtaken through the use of some lexicons and algorithms.

Lastly, a linguistic approach, where the outcome of text is derived from the semantic structure of it, can be used, even if it is not commonly used as it is time-consuming and not very implementable in real applications (Zhang, et al., 2018). The best choice varies according to the objective of the research and even to the data available for analysis. For instance, a certain model or dictionary can be better indicated for a study on social media analysis because it was specifically suited and built for this particular domain.

As it easy to intuit, sentiment analysis gives a great hand to company intelligence and analysts. To give a numerical measure of the importance that it has for business, it is suffice to say that the revenues arising from the whole software market for sentiment analysis and opinion mining are $500 million for now and will reach $3,8 billion by 2025 (Govoni, s.d.). In particular, some crucial advantages brought by sentiment analysis are (Monkeylearn, 2020):

1. Providing real-time results: sentiment analysis allows to identify points of strength and criticalities almost instantly, thus getting ready to respond adequately to the results given by the model.
2. Analysing data on a large scale: this is the benefit that all the fields of data mining share, even in the sentiment analysis subfield, as the data produced are considerably vast.
3. Giving a uniformed criterion to apply on data: fortunately, humans do not have the same point of views over things, as they are influenced by their memories, past experience and reference groups, and give opinions over different topics on different time. Taking this into account, the same sentiment analysis model can be used in different field of application, thus increasing the efficiency of researches.

The importance of sentiment analysis is therefore crucial for some elements of a business, like (Altexsoft, 2018) (Marketing Inside, 2017):

1. Brand/Product Monitoring: companies are interested about their audience reputation, market credibility in the market and image perception. They can also be willing to know who is talking about them, in order to propose the best outcome according to the eventual situations that occur.
2. Competitive analysis: the goal could also be the one to individuate some critical points in competitors, perhaps compared to their own. For instance, it can be a way to outperform competitors levering on their weak point and equalizing strong points.
3. Customer service: as users write opinion on blogs or social media, companies can focus on how they feel about them or if they perceive that some criticalities need to be fixed. As a consequence, companies can prioritize tasks and solve customers' concerns.
4. Product Analysis: the interest can also verge on evaluating which features of a product customer value the most. This can be also a great starting point to target marketing campaigns or work on a successful innovation.
5. Identifying trends: as people are the subjects that produce trends, what they write can be a direct source for detecting emerging trends that eventually can become the core interest of customers and society in general. This can also result to a great competitive advantage if its consistency in monitoring is kept overtime.

6. Employee engagement: the satisfaction of the workforce is crucial for companies' performance. Thus, knowing what motivates or makes employees unhappy through sentiment analysis can be a key factor in sustaining company's environment and eventually improving results.
7. Viral tracking: the analysis can also be a useful tool when the aim is to detect and track buzz marketing campaigns, in addition to online conversation over the element at issue.
8. Social listening: sentiment analysis can mainly be targeted to inspect social media performances to assess business' presence over these communication channels.

As said before, one of the richest and perhaps most active sources of user-generated content are undoubtedly social networks, thus detecting relevant characteristics of the functioning and implementation of sentiment analysis on these particularly busy domains can be of great interest.

Social Media Analysis

Every day, billions of people open their social media account and start scrolling down the homepage. As soon as they find a catchy content, that can be an interesting article, a friend's post or an advertising of a pair of shoes, reactions follow: some can just have a look on what other people said about that content and some others can directly contribute to a conversation or commenting section by writing down their opinions or feelings. If these actions were multiplied for each social media user in the world, the resulting number would be huge, enough to understand the relevance of the phenomenon.

There are consistent reasons that make social media relevant for business and marketing: just thinking that a quarter of purchasing decisions are influenced by content posted on social media should be enough to understand the power of these channels, adding the fact that 79% of people choose social network to share life moments with friends and relatives (Worldz, 2018). In this context, an effective online word of mouth chain is right behind the corner and can potentially happen every time a person engages with the content of a post, confirming recommendation as a powerful tool for deriving useful marketing results. These and other evidences paved the way to a lot of digital marketing tools, that led companies to the enhancement of their performances and sales. Thus, levering opinion mining in social media is certainly a useful idea to find out how people behave, what they think and share and to guide them towards a better relationship with the company or the business at issue. This operation is generally called "social media analysis". Nowadays, there are plenty of social media in which these actions are made every day: however, some of them are more indicated for a text mining task.

Among these, the most relevant is the micro-blogging social network Twitter. In this platform, users interact between one another through messages called "tweets". There are plenty of reason why Twitter is more indicated for social media analysis purposes (Lin, 2019): Twitter has 330 million active users monthly, as in 2019 Q1, with 500 million "tweets" sent each day, meaning an active social network where information is mainly delivered via text. Within them, 40% purchase something after having seen it on Twitter. Moreover, 67% of B2B businesses use Twitter as a digital tool, so the platform can count on a robust presence of companies. This also means that a direct contact and relationship between business and customer is common

and even bi-directional: in particular, when companies respond to users' tweets, 77% of these users develop a positive feeling toward that company's brand, product or service (Elrhoul, 2015). Last but not least, Twitter is more valued for its informational content and it is more used than Facebook, one of its main competitor, for breaking news (Hernández, 2016). When dealing with social media analysis on Twitter for sentiment analysis scopes, it is possible to define it as Twitter Sentiment Analysis, or TSA. Generally, TSA is mainly used for four different macro-areas of study: product reviews, movie reviews, political orientation and prediction on stock market (Pandarachalil, et al., 2015), but almost every objective can be achieved if the useful data are retrieved and implemented.

Nevertheless, sentiment analysis on Twitter or in every other social media has its pros and cons: as seen above, the large amount of information is certainly an advantage when it comes to the gathering phase, and its non-filtered characteristic makes it really valuable for "listening" to consumers; however, the downside is that not all the data are actually useful for the analysis, as a lot of tweets contain little or no relevant information, or otherwise not processable (Marta, 2020). Moreover, the tweets that can be collected will obviously represent just a little sample of users' opinion, although the research can be narrowed according to the needs.

In addition, social media contents cannot be considered as a standard document when it comes to analysis, as they embed unique characteristics that make them different from other domains. In fact, these contents are first and foremost short: it is sufficient to say that text limit for a Twitter post is 280 characters. Moreover, there is a completely different way to share emotions and thoughts, as users often express them in forms of emojis or in an informal and sometimes slang language, which are more challenging to process from a computational perspective. Hence, non-filtered form of expressions cannot be considered as ready-to-be-processed information. For these and other characteristics, directly using text mining techniques over social media posts can lead to significant bias (Zhang, et al., 2018). Moreover, even the post structure cannot directly be linked to a document one, and the levels of analysis, although structurally the same, slightly differ because of the diverse context, especially when it comes to a fine-grained approach. That is why social media analysis usually require peculiar approaches where generalization is more difficult to achieve as the online domain has its own characteristics, and that is also why different techniques were used when dealing with its study, as discussed further on.

Literature Review

It is possible to highlight different studies that levered text mining techniques to better comprehend the content to analyse. As seen, many different approaches can be implemented. (Zhai & Massung, 2016) showed that different level of deepness can be implemented using NLP, and most of the time the level of computational difficulty is directly proportional to the need of human interpretation. In addition, the drawback of a more precise knowledge of text gained through deeper level of NLP analysis is the fragility of the system, that can be more exposed to errors and imprecision. That is why a good payoff between information and robustness needs to be found while text mining. Another finding made in (Zhai & Massung, 2016) is that combining structural features that analyses syntax with lexical features showed better performances than adopting a single

feature. Thus, performing two layers analysis can guide towards more complete and correct results. Two methods were also applied in (Kennedy & Inkpen, 2006) for the determination of sentiment in movie reviews. More specifically, the study investigated over the role that valence shifters, namely words that can reverse or amplify the meaning of the accompanied word, has in the determination of the correct sentiment score. Valence shifters taken into account were negations, intensifiers and diminishers. The first method consisted in a corpus-based lexicon approach: positive and negative words were counted through the General Inquirer. This system provides a method for listing terms and give the exact sense of them through a little definition. In addition, valence shifters were included in the calculation: negations reverted the score of the positive word, while intensifiers and diminishers gave the word +1 or -1 in score, respectively. The second way was conducted through the use of machine learning mechanism, in particular the Support Vector Machine one. What was found is that, if bigrams with valence shifters were taken into account rather than unigrams, the accuracy slightly improves. In (Humphreys & Wang, 2018), a broad study on the relationship between consumer and language, providing a complete overview of the methodologies, techniques, issues and points of strengths in consumer research analysis through written text. Moreover, (Blei, 2012) explain a methodology that allows to extract relevant topics from a large set of documents. This set of systems is presented as "probabilistic topic modelling" and one of its main application can be found in a topic model called "Latent Dirichlet Allocation" or LDA. More specifically, as seen, what LDA does is to perform a conditional distribution, so to assess topic probabilities, hidden variables, knowing the words in the document, or observed variables. Nevertheless, one of the relevant problems of LDA is that it is not possible to derive the optimal number of topics *k* to take into account in a set of variables, and sometimes words may overlap in them. To tackle this problem, (Cao□, et al., 2009) developed a model that consider the correlation between topics to assess an ideal number of topics to select. Basically, when the density between topics reaches the minimum, it means that their correlation is also low, thus improving the independence score of each topic. In this way, a more stable structure can be built, and overlapping are avoided as the optimal number of *k* gives back semantic clusters, another way to define topics, that are similar within them and different between them. With the aim of optimizing the results given by Dirichlet multinomial distribution, (Yao, et al., 2009) developed an algorithm called SparseLDA whose main characteristics is its reduction of sampling time and less computational time and memory usage for processing topic modelling. Different methods were compared in order to understand which of them would produce better results with minimal computation, including classification-based inferenced like Naïve Bayes Classifier and Maximum Entropy Classifier, with the latter that does not assume independences among features, and sampling-based inference, with methods derived from Gibbs sampling. What resulted to be the best classifier, according to this study, Maximum Entropy or MAxEnt. Still from the NLP field, (Timoshenko & Hauser, 2019) deployed convolutional neural network or CNN and clustering for filtering non informative content and redundancies across a corpus made of user-generated content or UGC. The aim of the research was to assess if UGC-base consumer needs were comparable to interview-based ones, and if machine-learning techniques were able ti improve the detection and identification of consumer needs. To do so, word-embedding in UGC corpus was implemented to convert words into numerical vectors. Subsequently, a small sample of

sentence were labelled as informative or non-informative, and the machine learning classifier CNN was trained to complete the task. In this way, the informative content in the corpus was identified. After that, sentence embeddings were derived from the average of word embeddings and the former were clustered through Ward's algorithm, that minimize the variance within clusters. Finally, customer needs were formulated by extracting sample sentences from the clusters. (Hong & Park, 2019) implemented un unstructured approach based on clustering in a keyword analysis on airline reviews data. The process was structured in a way to deploy the synergies between text mining and survey. On the text mining side, the process worked in a way that the most relevant keywords were extracted out of data reviews of online customers, which were composed by two parts: the first one was text data with customers' experience information, while the second was a survey with questions about satisfaction and recommendation after having use the airline services. In order to extract words from the former, the method used was to assign weights basing on the frequency of appearance of words in document set, after having pre-processed them. The inspection of the most frequent words was performed by two visualization tools like frequency histograms and word cloud. At the end of this first part of the process, 45 keywords were retained. For these, hierarchical clustering was performed, and they were classified in two clusters. The result was that, in the first cluster, there were the providers of services, like "staff", while in the second the details of the services like "meal" were in place. To complete the analysis, survey items were used as variable, and correlation were studied between them and keywords. In this way, for instance, it was possible to understand which keyword affect the item "satisfaction" the most. The result was that "seat", "staff" and "cabin" were the keywords that had the higher and significant effect on satisfaction. In (Popescu & Etzioni, 2005), the aim was to extract relevant feature from online reviews. To do so, OPINE was introduced, that is un unsupervised system of information extraction that mines reviews in order to find not only features, but also their evaluations and quality across products. The steps were the following: firstly, identification of features takes place. Secondly, the opinions about the features were identified. Thirdly, the polarity of opinions was determined. Lastly, opinions were ranked based on their strength.

Nevertheless, plenty of previous studies have focused on social media analysis to conduct a research on sentiment over particular entities. What emerges from the literature is that different approaches, techniques and programming tools can be implemented to conduct sentiment analysis. This is mainly explained by the significant differences that arise in a different domain, where not only words but also punctuation, emoticons and slang can potentially play a role. The following are some of the approaches used in literature when dealing with text mining and, more specifically, social media and Twitter Sentiment Analysis.

For instance, (Pandarachalil, et al., 2015) focused on a lexicon-based unsupervised method for analysing tweets' sentiment using three different lexical resources, namely SentiWordNet, SenticNet and SentislangNet, with the latter that represents a specific lexicon built for recognizing slang words, that as seen are largely used in social networks. (Zhang, et al., 2018) gives advices for tackling the abovementioned challenges in social media analysis: in particular, these pitfalls can be overtaken by different measures. For instance, according to their study it is not indicated to consider every post collected as a single document, but rather group the ones that have similar targets and then create separate documents; alternatively, the sentiment of every posts can be

used to assess the calculation of the overall sentiment. Therefore, they also state that a way to overcome slangs and typos is to build personalized dictionaries. Finally, another advice is to assess not only the sentiment over a general target, but also over the features of that particular target, an approach used is the aspect-oriented sentiment analysis. (Choi & Wiebe, s.d.) created a lexicon based on positive and negative sense, identified as + and – sense, to overcome the potential ambiguities and difference in senses that can arise in simple word-based lexicon. A hybrid method between a graph-based method and standard classifier resulted to perform better to analyse new glosses. (Saif, et al., 2016) overcame the lexicon-based approach problem of static polarity score of words, proposing an approach called SentiCircle, a dynamic representation that evaluates sentiment and polarity of words taking into account its semantic context and so its co-occurrence. Thus, a word can score with a higher strength in a domain and lower in another, and vice versa. The study was conducted both at the entity-level, so detecting sentiment over a particular topic, and at tweet-level, assessing the sentiment of the individual tweet. (Buckley & Paltoglou, 2011) developed SentiStrength, an algorithm that was specifically made for assessing sentiment polarity and strength in short and informal social media contents: the algorithm was experimented in six different social web domains, including Twitter. Similarly, (Anna Jurek, 2015) adjusted lexicon-based approach outcomes to its intensity and through evidence-based and sentiment normalisation functions. (Kowshalya & Valarmathi, 2018) conducted a TSA study for analysing users' perception about Social Internet of Things. The sentiment analysis was performed by using algorithms called Improved Polarity Classifier (IPC) and SentiWordNet Classifier (SWNC), hybrid classifier, IPC+SWNC, and a semi-supervised algorithm called Fragment Vector, with the latter that performed highest results in accuracy. (Rathan, et al., 2018) performed an aspect-based sentiment analysis for mobile phone brand and determined features by using a training dataset labelled with lexicon-based approach and applying spelling and abbreviation correction, emoticon detection and emoji detection, the latter through the transformation of Unicode language. After these steps, SVM model is applied to the test set composed by tweets that contained smartphone's name and features of interest for the research. (Öztürk & Ayvaz, 2018) collected tweets in Turkish and English language to conduct a sentiment analysis study over Syrian refugees' crisis: for Turkish ones, a manual lexicon dictionary was created, for the English ones, a sentiment lexicon called *"RSentiment"* available in R programming language was used. The most frequent words were put into relevant clusters and the sentiment trend was studied overtime, following tweets' dates. By following a similar path, (Kabir, et al., 2018) pulled tweets through R and used a dictionary of positive and negative words available on the website "GitHub" and studied the polarity trend of tweets throughout a week. They therefore focused on the characteristics of *"wordcloud"* visualization tool. R was levered also in (Nigam & Yadav, 2018) to conduct a sentiment analysis in twitter domains. Tweets were collected and cleaned through R packages, in order to derive a processable text. To implement the analysis, two dictionaries for positive and negative words was created, with the range of scores that goes from -1 to +1. So, the words in the text were attached to the words in the dictionaries. Moreover, this lexicon was composed by both English language and Hinglish, namely language made by words that are a hybrid of Hindi and English words. After that, a histogram with the results was showed. A similar method was implemented in (Bail, s.d.). Tweets about Trump were

extracted through R and were analysed together with other labelled datasets. Some additional tasks were performed: word counting, term frequency-inverse document frequency and sentiment analysis. In particular the scores of the latter ware compared across different domains analysed. Nevertheless, an unsupervised approach called Latent Dirichlet Association, or LDA, for topic modelling of tweets was used in (Zhao, 2013), together with the visualisation of frequent term and their association, namely the network of terms based on their co-occurrence in the same tweet. LDA was also used in (Namugera, et al., 2019) in a wider study on the relationships between tweets generation and traditional media houses in Uganda: in particular, it served as a driver for discovering hidden topics in that media houses. After that, sentiment analysis was conducted on each of them over the time under consideration. Even in (Y.K.Lau, et al., 2014), an LDA-based topic modelling was implemented to develop a social analytics methodology for aspect-oriented sentiment analysis, and was inserted in an instantiation called Ontology-based Product Review Miner or OBPRM. Moreover, the methodologies applied in (Cho, et al., 2017) were both LDA and social network analysis, for a study that analysed topics trend in marketing journal from a time rang that goes from 1995 to 2014. By leveraging two R packages called "tm" and "lda", it was possible to clean and pre-process the text data coming from 25 marketing journals. After that, bigram features were considered to extract relevant topics in journals. The word in topics were then scaled in a way that the most representative was weighted more and put in front. In addition, social media analysis was implemented to inspect the association between paper and authors and to detect the co-authorship structure across the different marketing journals. The findings of the study revealed that, as business priorities and interests evolved, marketing followed the path. Thus, for instance, topics moved from the focus on product quality to the brand differentiation one, arriving to identify the emerging trends of sustainability and digitalization. The authors-paper analysis instead showed the preference of authors to collaborate with new partner and the importance of social capital in creating an authors' network. Moreover, in (Hartmann, et al., 2018), 10 different methods, five machine learning and five lexicon-based, were compared in the level of performance across social media datasets covering different platforms, size and languages. The study states that, in marketing research, mainly two approaches are used: one is the Language Inquiry of Word Count or LIWC and the other one is Support Vector Machine or SVM. Another point highlighted was that, especially in marketing and economics researches, the implementational costs, interpretability of results and inspection of economic difference play a crucial role that has to be taken into account when choosing the approach to text mining. The abovementioned models were compared with others, namely Naïve Bayes, Artificial Neural Network, k-nearest neighbour and Random Forest, together with other four lexicon-based approach. What emerged from the study was the confirmation that there is no model that performs the best across every domain or dataset, thus each approach gives different performances results according to the different elements taken into account. In addition, Random Forest method resulted to be agile, flexible and robust especially for three-class sentiment classification, and Naïve Bayes were one of the best models for recalling human intuition, and outperformed other approaches when classifying two-class sentiment datasets. To what concerns lexicon approaches, these did not perform less in accuracy, compared to the simplest machine learning methods like Naïve Bayes, contrasting findings of other researches. Another

research conducted in social media domain, in particular Twitter, is the one of (Sailunaz & Alhajj, 2019). In addition to detecting the sentiment of twitter posts, a dataset containing information about text, users and emotion was created. To perform, sentiment analysis, Naïve Bayes classifiers was used. Moreover, user's information was collected to detect the influence score of each user, and together with the sentiment and emotion ones, a system recommendation based on past tweets activity was implemented. In particularly, to determine the closeness of users, the Acquaintance Affinity Identification score was calculated.

Moving on, different studies in literature levered on lexicon techniques to implement studies on text mining, sentiment analysis or social media analysis and to introduce or adapt dictionaries to be used in specific or various domains. For instance, (Whissel, 2009) refined a dictionary called Dictionary of Affect in Language, that was created to specifically attribute Pleasantness and Activation of emotional word. To do so, a group of volunteers were asked to assess a score from 1 to 3 to individual word for its Pleasantness and Activation, that are part of the affective space, and Imagery, so the easiness to create a mental picture of that individual word. The higher the scores, the higher the abovementioned factors. The main findings were that, first of all, although weak, the three scales were significantly correlated; moreover, even function words, that are words which contribute to the syntax of the entire phrase rather than having an absolute meaning in their self, were resulted to be distinct in their emotional undertone. The dictionary was applied in two different samples of natural, showing its ease of adaptation to different contexts. Moreover, (Cho, et al., 2014) built a and advanced lexicon that was the result of the merge of different available sentiment lexicon dictionaries. To do so, scores of the words in every dictionaries were normalized in order to obtain values within -1.0 and 1.0; therefore, a "remove" and "switch" approach was used, where words that were not considered useful, according to a positive/negative occurrence ratio and positive/negative review ratio, were removed, while the score of the others was adjusted to three different domains, namely smartphone, movie and book. (Hutto & Gilbert, 2014) developed a lexicon tailored to microblogs content called Valence Aware Dictionary for Sentiment Reasoning or VADER, also adding abbreviations, emoticons, acronyms and slang. Instead of using a machine learning approach, a gold standard, human validated one was adopted. In fact, pre-screened and trained "raters" were asked to give a score of -4 to +4 to the polarity of phrases and words, in order to build an efficient, general and domain-independent lexicon. Social media played a crucial role also in (Ordenes, et al., 2019), where text mining studies on famous brands' posts over a time range of two years were conducted in Facebook and Twitter domains to understand how contents posted by brand can affect user's message sharing. What the study found out was that the use of cross-messages in different platforms and the use of rhetorical style enhanced consumer message sharing. Similar results were also found in the image-based study, were the presence of visual features or image acts showed to improve message sharing of consumers. These findings would thus contribute to enhance the effectiveness of social media marketing campaigns. (Katz, et al., 2015) developed ConSent, a context-based approach that identify key terms among classified positive and negative documents and context terms, that reinforce the intensity of key terms. Although this approach is effective in noisy datasets, it was not indicated for social media even if it is applicable in this type of domains. To refine lexicon results, Sentiment Orientation Calculator, also known as SO-CAL, was presented by (Taboada, et al.,

2011), that takes into account intensifiers and negations. (Kauffmann, et al., 2019) developed a sentiment analysis for mobile phone reviews in Amazon on different levels: in fact, NLP techniques and tools were used to firstly detect the sentiment of each reviews, using AFINN lexicon and normalizing the star score. Therefore, the analysis was narrowed to sentence level for every review, assigning a score for each phrase. Finally, after having processed the text through Part-Of-Speech tagging, the most frequent nouns were detected to extract the most relevant features in the sentences, in order to understand which particular factor was mainly cited and levered reviewers' satisfaction. Finally, visualization tools were implemented to better share the results obtained. Again, (Khan, et al., 2016) created SentiMI, a sentiment dictionary that was built upon SentiWordNet 3.0 and improved its performances mainly for what concerns its specificity, namely the correct classification of true negatives words. To do so, it levered mutual information on positive and negative labels after having pre-processed the datasets and implemented the Part-Of-Speech POS tagger. (Warriner, et al., 2013) took the Affective Norms for English Words, also known as ANEW, as a base to build a more complete dictionary. ANEW distinguishes three emotions: valence, or the pleasantness of a stimulus; arousal, namely its intensity of emotion; dominance, which is the degree of control of a stimulus. While ANEW contains 1,034 words, the new set was extended to 14,000 English lemmas, adding information about gender, age and education differences in emotions. Another lexicon was created and implemented in (Mohammad & Turney, 2013). The peculiarity of this study was in the fact that, to create the dictionary for sentiment analysis, researchers levered Amazon's crowdsourcing project called Mechanical Turk. Technically, crowdsourcing is defined as "*The act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call. This can take the form of peer-production (when the job is performed collaboratively), but is also often undertaken by sole individuals. The crucial prerequisite is the use of the open call format and the large network of potential laborers"* (Howe & Robinson, 2006). In Mechanical Turk, volunteers get paid for solving "HITs", namely "Human Intelligence Tasks", thus contributing to the research. Through the platform, a term-emotion association was derived in a dataset called EmoLex. The emotions on which EmoLex focuses are joy, sadness, anger, fear, trust, disgust, surprise and anticipation. These were derived by the "Plutchik's wheel of emotions", where similar emotions are placed next to one another, while contrasting ones are placed to the diametrical opposite of the circle. Moreover, the dataset included both unigrams and bigrams for most frequent English nouns, pronouns, adjectives and verbs. In this study, terms were retrieved from the "Macquerie Thesaurus" and for each of them a HIT was created. The HIT consisted in a questionnaire about a single word were volunteers were asked to assess if association between the word at issue and the abovementioned emotions was sensed. In second place, to validate the annotation, an automated script was used. So, the research showed that wisdom of the crowd can be able to create a term-emotion association to build a dictionary quickly at a minimum expense. Moving forward, in the research of (Ordenes, et al., 2017), the focus shifts on the different ways and range of linguistic features to express emotion, that can better assess the value scores for a sentiment analysis study. In fact, through the introduction of speech act theory or SAT, which allows to underline the higher-order features like reinforcements or attenuative terms that contribute to adjust the tone of the phrases.

Moreover, the study investigates over certain phrases where language is used in a way that sentiment or feelings are not expressed or expressed implicitly. In this way, the differential effects of these different ways of expressions are detected. The study was conducted over different online domains, where a star rating scale from 1 to 5 was in place, although applicable to others like social media. Some of the most important findings were that, when positive explicit expressions were used, the probability that a user would rate the experience one point higher was doubled. The same effect reversed was not found for explicit negative expression. In addition, it was confirmed that implicit expressions can convey sentiment. Two different approaches, dictionary for verbal cues and grounded for nonverbal cues, were applied in (Marinova, et al., 2018) to isolate and then inspect the effectiveness of frontline employees' problem solving on costumers satisfaction in face-to-face interactions. The research found that a positive interaction between the two factors exists, increasing in magnitude across the interaction time. However, this effect tends to diminish for high levels of relational and displayed affect, while it is stronger for low level of the same characteristics. So, over displaying of affection can lead to higher level of dissatisfaction and disappointment in customers complaints. Nevertheless, another finding made by the study was the one that the level of frontline employee's problem solving remains significant even if the solution of service's problem cannot be found. That is, customer tend to value the efforts made by the employee, regardless of the final result, thus separating the process of problem solving from the final outcomes and results. The study was conducted in an airline setting, both in field and experimental, so the two different types of cues were extracted from this context. Another mixed approach was used in (Ordenes, et al., 2014): in this case, text mining technique was implemented together with other crucial characteristics for the evaluation of customers experience feedback. In detail, a more holistic approach was implemented that comprehended three value creation elements, namely activities, resources and context. In this way, the findings arising from text mining were enriched. The model was defined as an "open learning" model as it demonstrated flexibility through diverse trainings and domains. Finally, a previous study (Pierre, 2019) used a Phyton package called "textblob" for conducting a sentiment analysis of the Tesla Cybertruck. However, the analysis was limited to just a mere assessment of sentence polarity, either positive or negative, and the count of their frequencies, without going deeper into analysis that will be further discussed in this study, where different and more detailed tools will be implemented. Again, these are just examples of the various methods that can be used to perform a sentiment analysis on social media domain and in text mining. As seen, diverse studies were conducted for numerous scopes, meaning that a constantly increasing attention is put on computational analysis, often going beyond marketing and business in general, that actually used and are still using these techniques to optimize customers, product and brand researches.

Object of study: product characteristics

As seen, a sentiment analysis needs to be directed on an entity, namely the object over which the emotion or sentiment is expressed (Sun, et al., 2017). Moreover, it was also highlighted that one of the main areas of research in Twitter Sentiment Analysis or TSA is product reviews (Pandarachalil, et al., 2015). So, the object of this study is introduced in the present section and comes from the automotive sector. Along with the reasons why Twitter is a valuable source of information for brands and consumers' purchasing process (Lin, 2019) (Elrhoul, 2015), there are also additional evidences highlighting the importance that social media cover particularly in cars and vehicles purchase decisions. As a matter of fact, researches evidenced that potential car buyers are approaching more and more to social media, especially during the selection process. A survey on 500 car buyers conducted by Crowdtap, the online community that works as a marketing platform website together with known brands to get consumers' opinion and ideas through the use of pools discussions and even product tests with the opportunity to earn money or free goods (Wealth Inflator, 2018), showed that 87% of people search potential cars purchases on social media during the selection phase, 68% of car buyers purchased a car that was found in social media and 80% of them are more willing to search advices on social media than on car salesperson. What is more, is that 95% of people would post about a car they purchased or liked on social media (Businesswire, 2015).

Another study of Digital Air Strike, a social media marketing and digital engagement company, was conducted on 2000 car buyers and 2000 service customers. What was found is that 75% of car buyers and 68% of service customers stated that Internet and social media research, together with reviews sites, were the most helpful medium for selecting car dealership (V12, 2016). A confirmation of that comes in the 2019 study by Cox Automobiles Company, an automobile retailer and wholesaler: what emerged is that car buyers spend 61% of active shopping time online: *"consumers are spending more of their shopping time online while making faster decisions and spending fewer days in market."* (Cox Automotive, 2019).

Moreover, another research conducted on Twitter by the social media audiences' insights platform called Canvs, showed that more than 327,000 auto-related tweets are sent every day, and 75% of them are consumer driven, so not arising from brands. In addition, 60% of Twitter auto-conversations are made in the "pre-purchase" phase, while 30% are occurring in the "during purchase" one (Jr., 2015). These numbers are just another evidence of the fact that levering social media and reinforce online presence for brands can drive to meaningful sales results even for the automotive sector, showing an increased trust of consumers in the purchase of long usage products like cars.

The product at issue is a vehicle that has made the news for its distinctive design and its controversial unveiling. This is the electric pickup truck of the American company Tesla, called Cybertruck.

The company: Tesla, Inc.

Before introducing the characteristics of this electric pickup, it is important to introduce the company first, including its relevant personalities and their choices in online and social media context. Tesla was founded in 2003 in San Carlos, California by American entrepreneurs and engineers Martin Eberhard and Marc Tarpenning under the name "Tesla Motors", in honour to the Serbian American inventor Nikola Tesla, known for its contribution to electrical engineering whose works led to the invention of the "alternating current". The initial goal of the company was to develop an entirely electric sports car, after the promising results arisen from a test market conducted by General Motors. In 2004, the company received $30 million in investment by the CEO and founder of SpaceX and X.com Elon Musk, who then became the chairman of Tesla's Board of Directors on the same year. In 2006, Eberhard and Tarpenning unveiled the first prototype of the electric car, called Roadster, the production of which started in 2008. What made Roadster unique was that it met customers' need, namely a long-lasting battery and a powerful motor that could reach highway speeds. In fact, Roadster reaches 394 km on a single charge and goes from 0 to 96 km/h in less than four second, with a maximum speed of 200 km/h. Moreover, the position of the battery in the car was at the front of the vehicle. The Roadster's initial cost was a little more than $100,000, that contributed to position the product and the brand to the high luxury niche segment. The production of Roadster stopped in 2012, with the efforts of Tesla that moved to another electric vehicle, a sedan car called Model S, that was immediately acclaimed by critics especially for its design and performances. One of the main characteristics of the car was that the battery was located on the floor, thus optimizing the space and giving a more equilibrate balance and centre of gravity. In the same year, charger stations were built across United Stated and Europe, in which Tesla owners could charge their cars at no extra cost. After that, in 2015 a crossover called Model X was released, which had seven seats and reached a battery range of 475 km, and in 2017 another sedan called Model 3 began its production. The latter was less expensive than the other Tesla models, having an initial cost of $35,000. (Schreiber & Gregersen, 2018) (Reed, 2020). Two more products were then unveiled in 2019: in March, the mid-size SUV called model Y, while in November the pickup truck called Cybertruck (Hawkins & O'Kane, 2019).

Two other crucial moments in Tesla history need to be mentioned: firstly, the change of name from "Tesla Motors" to "Tesla, Inc." in 2017, signalling the willingness for the company to enlarge its business to the production of solar energy products useful for powering houses and businesses; secondly, the relocation of the company in Palo Alto, California, known for being the headquarter of famous innovating companies (Squatriglia, 2009). The move was crucial since it allowed Tesla to put the basis for the creation of a real competitor in the automotive market, targeting the electric car niche that was almost unexplored by incumbents. Moreover, the influence and philosophy of Palo Alto's companies, focused on the constant seek of innovation, often reached by seeing businesses from a different point of view and embracing risks, were and are still vivid in Tesla, and contributed to represent one of the main drivers of its success. Two of the most

relevant innovations proposed by the company are the creation of an entire set of vehicles completely electric with zero emissions and the developing of self-driving mechanism also known as autopilot.

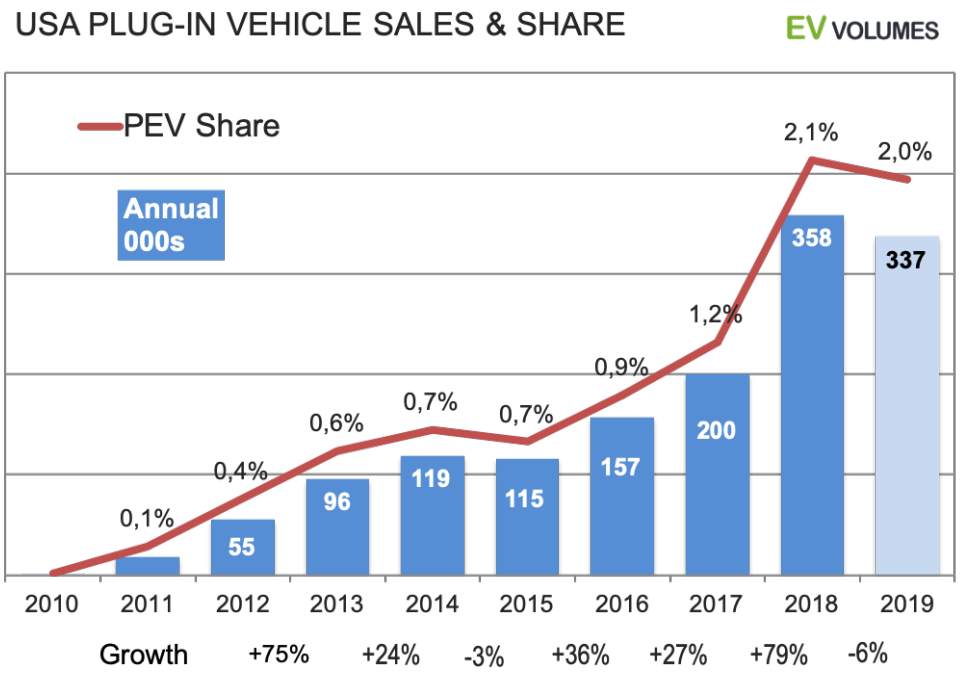
Concerning on the first one, for the sake of completeness and clearness it is important to make a distinction among different types of electric vehicles that include an electrical component in their engines. The family of Electric Vehicles or EVs cars comprehends (Electric Car Home, s.d.) (EVgo, s.d.):

1. BAV, Battery Electric Vehicle: these vehicles all completely powered by a battery; thus, they do not need other diesel or petrol engine as a support. The battery is rechargeable through an external source. The chargers work with a plug-in system that links the external source of energy to the vehicle and are classified into three different levels: the first one can be typically done at home or work and is called level 1, that uses an household outlet of 120v and takes over 8 hours to charge for around 120km-129km; level 2, that requires a charging station that delivers power of 240v and takes about 4 hours for 120km-129km; level 3 or DC fast charging, that are located in dedicated stations and takes 30 minutes for a charge of about 145km.
2. PHEV, Plug-In Hybrid Electric Vehicle: these cars are provided with both a battery and petrol or diesel engine. The battery can be charged by either the abovementioned plug-in system or through the so called "regenerative braking", which works in a way that the motor uses some of the energy generated by car's braking for charging. In addition, these vehicles move through battery only to a certain moderate speed and, above that threshold, the petrol or diesel engine is activated.
3. HEV, Hybrid Electric Vehicles: the hybrid types include both normal electricity and gasoline, but the electrical component cannot be charged through plug-in systems, thus levering only on regenerative braking. Similar to the Plug-In Hybrid, the HEVs start using electrical motors, switching to diesel or petrol or diesel engine for higher speeds. However, the latter is predominant over the former, as the battery is small and typically helps improving normal engine's performance.

Tesla cars are all BEVs and can also count on DC fast charging through the already mentioned dedicated stations.

Nevertheless, entering in such a consolidated market like the automotive one, where incumbent have enormous capacities and know-how thus constituting a very high barrier to entry, was not an easy task for Tesla, plus its move would be also considered risky for a new-comer company with no prior knowledges of the sector. Other relevant obstacles to the development of the electric cars were high fixed costs, the need for a large production capacity and the lack of an ecosystem to sustain the development of electric cars (Stringham, et al., 2015). Despite that, Tesla's mission "*to accelerate the world's transition to sustainable energy*" (Tesla, s.d.) made company go over these initial obstacles; in addition, this objective was clearly sustained by studies on unmet needs of consumers: in fact, at that time there was not a valid model that *"could address the false perception that people have that an electric car had to be ugly and slow and boring like a golf cart"* (Musk, 2014). To better clarify and quantify the demand that was targeted by Tesla's electric vehicles, the following bar plot is shown:

Figure 7: Sales and share of plug-in vehicles in USA



Retrieved from *ev-volumes.com*, 2019

The plot highlights the fact that the niche pursued by Tesla was in growth, thus meaning that a higher percentage of the population were interested in buying an electric car. Moreover, it is dutiful to say that Tesla itself contributed to the rise of electric cars' market share. So, there is also a mutual relationship that explains the plot: the growth of the segment caused Tesla's development and vice versa.

Another incentive arising from the electric market in US, together with the increased demand, is the subside received by the government in the form of tax credit. More in detail, since 2009 the government of the United Stated of America found an agreement on the release of little extra incentives to new Tesla buyers: this means that government contributed to a little part of the expense through an incentive for the purchase of a new Tesla car. In addition, a tax credit for electric vehicle maker was planned, that worked in that way: each electric car maker was eligible for $7,500 in credits for each vehicle sold, up to 200,000 sales. Six months after having reached this threshold, the credit would be halved, and then halved again after others six months, until the credit reached zero. Tesla hit the 200,000 sales in 2018, thus starting a descending phase for tax credits that

led to its end in January 1st, 2020 (Marshall, 2019). So, what constituted a point in favour can now be transformed into an additional cost that needs to be put into account in Tesla's strategy.

In fact, the company of Palo Alto adopted a specific strategy to face the challenges that it encountered and still faces. As stated by the CEO Elon Musk himself, *"The strategy of Tesla is to enter at the high end of the market, where customers are prepared to pay a premium, and then drive down market as fast as possible to higher unit volume and lower prices with each successive model*." (Musk, 2006). So, the strategy pursued by Tesla is to generate economies of scale starting from a high-end luxury product with high price and low volumes and ending to a product with affordable prices but higher-scales volumes, plus pursuing the goal of changing the automotive business and the ecosystem around it. More specifically, the firm levered on different actions, partnerships and investments to consolidate its position in the market. Firstly, Tesla finalised an agreement with Lotus in 2004 where the latter would have helped the former with the design and technology of the car, together with its structure and safety. This allowed Tesla to focus on the chassis and other characteristics of the vehicle, thus guaranteeing a crucial time saving and optimization of resources. After that, in 2009 it received more investments and established partnerships from companies like Daimler, Panasonic and Toyota for sustaining a large-scale production of the Model S. It is sufficient to say that, in four years of production, 2,500 Roadsters were produced, a low quantity compared to the 35,000 Model S cars produced in two years. Moreover, to tackle the problem of the ecosystem, Tesla built and still continues to build station networks for fast charging vehicles, with zero marginal cost to the final user. Although costly, the strategy helps not only its business, but even the others to reach the goal of moving to a transportation standard with zero emissions. As an evidence of that, Tesla did not reclaim a patent of property on their projects, because this would not contribute to the development of such a florid network. Instead, Musk pursued an open source philosophy (Stringham, et al., 2015):

"Our true competition is not the small trickle of non-Tesla electric cars being produced, but rather the enormous flood of gasoline cars pouring out of the world's factories every day. We believe that Tesla, other companies making electric cars, and the world would all benefit from a common, rapidly evolving technology platform. Technology leadership is not defined by patents, which history has repeatedly shown to be small protection indeed against a determined competitor, but rather by the ability of a company to attract and motivate the world's most talented engineers. We believe that applying the open source philosophy to our patents will strengthen rather than diminish Tesla's position in this regard" (Musk, 2014).

These words clearly explain the reason behind not patenting and adopting an open source philosophy instead: to let projects be analysed and improved for free from others would encourage talents to join the cause. On a large scale, this means that the interest towards the business grows, and more people will be attracted by the company and its product and this would finally be an important incentive for investors to put their money on the business. In this way, the abovementioned Tesla's strategy can be implemented, in addition to reach the goal of a different and eventually less polluted ecosystem without renouncing to the pleasure arising from driving a car and actually improve the consumer's experience in the vehicle.

This all together can be considered as the ideal virtuous circle for Tesla, but in reality, it has to contend the market with the actual competitors, that did not underestimate this rise and instead started working on their electric models. In addition, the infrastructure for Tesla's project is not fully established and the production capacity still cannot be compared to the ones of other bigger firms, putting the Palo Alto company in a not completely stable position. It is possible to state that, nowadays, Tesla has great points of strengths, but important weaknesses as well. Each of them shapes what the Palo Alto Company represents in the industry today. To better individuate those, a SWOT table is presented (Business Strategy Hub, 2020) (Serna, 2018):

Table 1: Swot Analysis of Tesla, Inc.

Strengths	**Weaknesses**
• Talent employment company • 1st company to produce an electric luxury car • Actual leader in U.S electric vehicle sales • Innovation-driven company • Renowned and influent CEO • Diversification	• Manufacturing inefficiencies • Lawsuits and liability claims • Unbalanced supply and demand • High volume production still not in place • Supply of batteries • CEO's other commitments
Opportunities	**Threats**
• Expansion in untapped international market • Price decrease for cars • Market sustain and confidence • Growing environment concern and interest • Advancement in technologies for battery • In-house production of battery • Entrance in pickup sector	• Established and prepared competition • Costumer adaptation • Defects of products • Shortage of materials • CEO's behaviour

As seen, some of the abovementioned points were addressed before. Naturally, all of them produce a chain of dependency that can affect the company both in short and long run. For instance, a growing customer concern and adaptation towards the electric would unlock different strengths and reduce other concerning weaknesses, but mostly depend also by the possibility to sustain the productive process and ensure that the partnerships would keep pace with the developments of electric models. Moreover, the adoption of the open source system facilitated the acquisition and insertion of talents that would contribute to the realization of company's project. Another tendency that would enhance Tesla's position is the cost of batteries: in fact, if this one goes down, Tesla's margin would consequentially rise. Generally speaking, Tesla's total expenses have raised from $6 billion in 2015 to $22.5 billion in 2018, where Cost of Revenues was the biggest driver, accounting from 76% of revenues in 2015 to 81% in 2018. A decline in Cost of Sales as percentage of revenues is forecasted for 2020, declining to 79% and leading to operating profits (Forbes, 2020).

Moreover, Tesla had to face lawsuits arising from product malfunctions, together with complaints about defects in design and manufacturing (Business Strategy Hub, 2020). In addition, the weakness of batteries supply is directly linked with the opportunity for Tesla to produce them in-house, thus vertically integrating one of the main features of the cars. Since 2009, Tesla entered into an agreement of supply with the Japanese

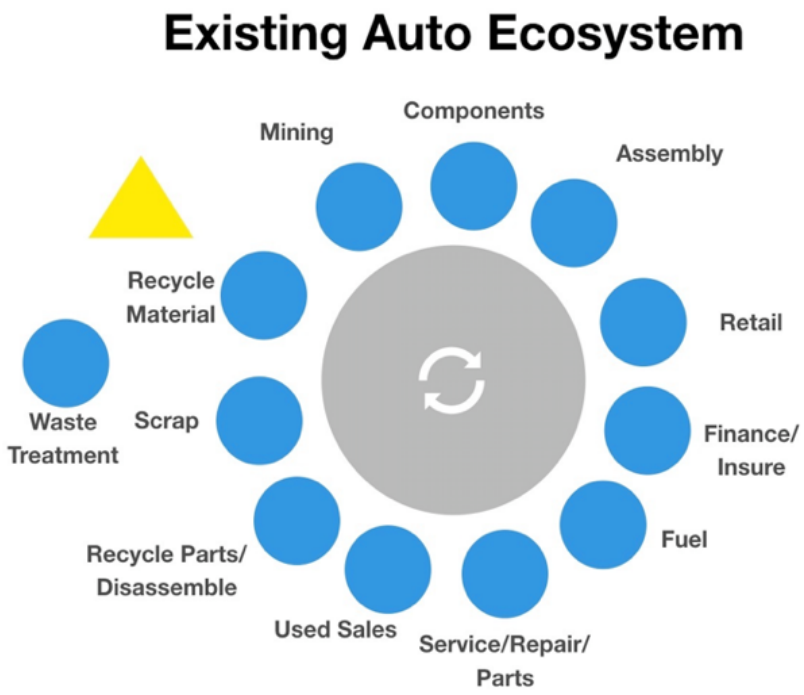
multinational electronical corporation Panasonic, that was reinforced with an additional investment of the latter in 2018 for making the electric industry market grow. Subsequently, in 2011 the two companies finalized another agreement, this time for the supply of automotive-grade lithium-ion battery cells, that would be used to produce 80,000 vehicles in four years (Tesla, 2011). However, since then, Tesla's high demands and Panasonic's limited deliver capacity led to some frictions. To tackle these problems, in 2014 Tesla announced another deal with Panasonic that would bring the manufacturing of batteries in-home (O'Kane, 2019). In fact, Tesla started the construction of a huge subassembly battery factory called Gigafactory. Its ultimate scope, as stated by the company itself, is to *"supply enough batteries to support Tesla's projected vehicle demand"* (Tesla, s.d.). The first one, named Gigafactory 1, is set outside Spars, Nevada, the construction of which started in 2014. Its size is 1.9 million square-foot, with 5.3 million feet of operating space divided in three floors (CleanTechnica, 2019). The second one, named Gigafactory 2, is set in Buffalo, New York and its construction started in 2017. It is mainly directed to the production of solar panels and solar cells and its size is of 1.2 million square foot and has already creating nearly 800 new jobs (Tesla, s.d.). Other two Gigafactory are in Shanghai and Berlin, Gigafactory 3 and 4 respectively, with the former that began the production in October and its focus is on the production of Model 3 and partially Model Y (Hull & Zhang, 2019), and the latter, which is still under construction, would also produce 500,000 Model 3 and Model Y per year (Lambert, 2019). These two will serve the main Tesla's markets after USA, namely China, Norway and the Netherlands (Hull & Zhang, 2019).

Nevertheless, the relationship between Panasonic and Tesla still remains unstable, as the latter is the biggest client of the former for the electric vehicle battery supply. This would mean that each investment made by the Japanese company has to keep diverse factors into account, like general demand among that, while Tesla's goal is to try concentrating the production in its own factories without losing important investments for their construction and production capacity. Moreover, as seen, one the biggest challenges for Tesla is the implantation of an ecosystem that can sustain the production and usage of its products on a large scale. This means that, to grow its business, Tesla needs not only to continue innovating its internal projects and systems but also to commit itself for a substantial change in the business, in both upstream and downstream fronts. In fact, in addition to the actual assembly, design, supply and cars composition, there are relevant challenges to be faced on a downstream and even post-purchasing level. Together with the abovementioned charging stations and grids, particular attention needs to be addressed to used vehicles and parts market and recycling of components. For the former, it is sufficient to say that the market for used cars in the United States is more than two times bigger than of the new ones (Wagner, 2020). For Tesla products, this is not necessary a disadvantage as its car components would last for several years. Even for that reason, the price of used Tesla still remains high after years, sometimes making not worth to buy a used model. Plus, this would mean to sacrifice a part of factory warranty, potential deterioration of battery capacity and the losing of unlimited charging in Tesla's Supercharger stations (Gorzelany, 2019). For what concerns the recycling process, the company is working to create a complete circle of recycling that will start and end at the Gigafactory. The words of Tesla's Chief Technology Officer Jeffrey B. Straubel clarify this point:

"*Tesla will absolutely recycle and we do recycle all of our spent cells, modules and battery packs. So the discussion about this waste ending up in landfills is not correct. We would not do that. These are valuable materials in addition to it's just the right thing to do.*
We have current partner companies on every major continent where we have cars operating that we work with to do this today. And in addition we're developing internally more processes and we're doing R&D on how we can improve this recycling process to get more of the active materials back. And ultimately what we want is a closed loop right at the Gigafactory, that reuses the same recycled materials." (Straubel, 2018).

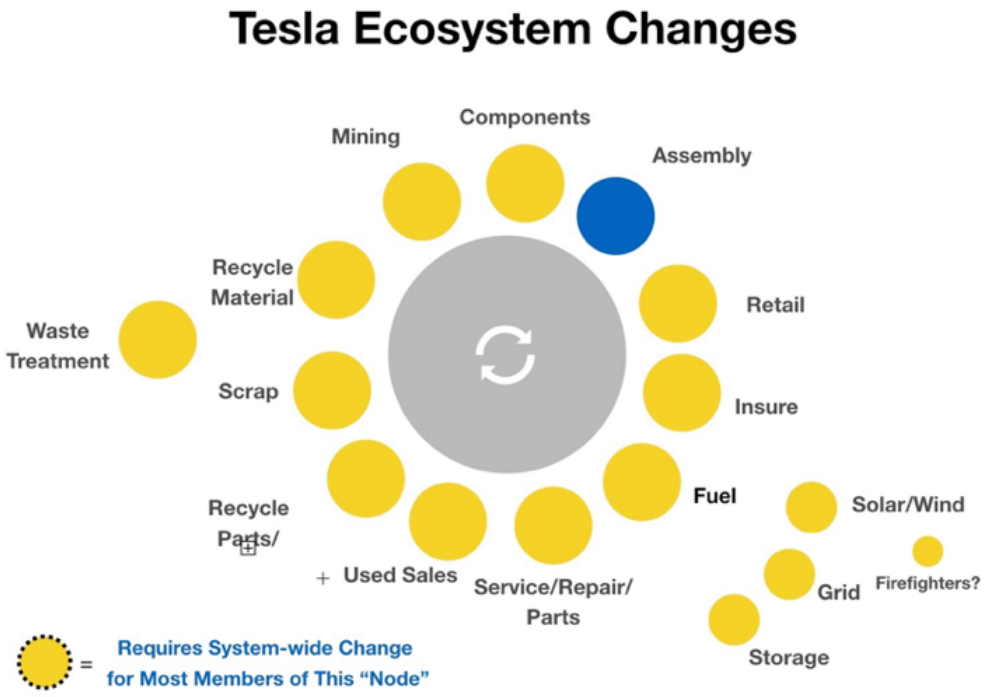
So, for now, Tesla is joining a partnership with companies around the world to provide the recycle of the batteries and materials, but the final goal is to move the process internally in Gigafactory, in order to lever on advantages arising from integration. Despite that, the study (Anderson, 2018), that compared the potential ecosystem of Tesla with the more established one of Toyota, shows that the Palo Alto company is requiring its stakeholders to make relevant changes, for instance in the processes of waste treatment, used sales, retail, fuel and components. In contrast, Toyota ecosystem, with its strategy to invest on hybrid and the use of a battery chemistry that is easier to recycle, is not asking the stakeholders meaningful changes, thus levering more on the actual state of technology for implementing its strategy to include electric components in vehicles. In fact, in contrast to Tesla, Toyota is focusing on the production of hybrid cars, namely a vehicle that, as seen, combines different propulsion systems. The following two images compare these ecosystems of Tesla and Toyota Prius, the hybrid model of the Japanese automotive manufacturer, that combines gasoline engine with an electric motor. The blue nodes indicate that not big changes were required, so the process put in the field by the stakeholders remains basically the same in its main components, while yellow ones indicate the opposite, meaning a revision of the system on a deeper level, requiring higher efforts not only to the company itself, but also to the system that gravitates around it.

Figure 8: Ecosystem of Toyota Prius



Retrieved from *Forbes.com, 2018*

Figure 9: Ecosystem of Tesla



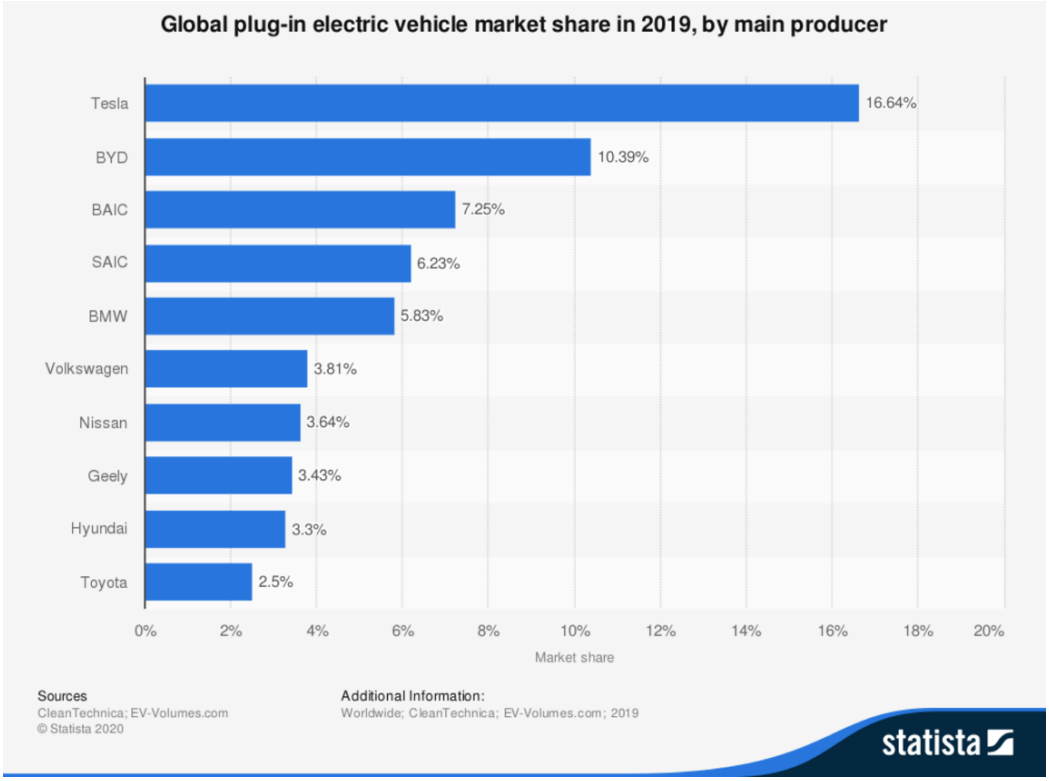
Retrieved from *Forbes.com, 2018*

Nevertheless, despite all these contrasting factors, the efforts of the company in electric vehicles paid back in that sector, in the sense that Tesla is leading the global plug-in electric market, as it is the main producer in market share with its selling that went between 367,000 and 368,000 units in 2019 (Wagner, 2020). In fact, in the first semester of 2019, Tesla sold around 83,875 electric vehicles, 10 times the volume of GM sales. Tesla Model 3 then became the most sold plug-in electric car in the US, overtaking every other EV in the same period and country by at least 750% (Matousek, 2019). 2019 was a record year for plug-in electric cars also in the European market. In December, sales went over 75,000 units and growth rate settled at 88%. Another relevant factor is the specific growth of BEVs sales, that increased by 91% year-over-year, specifically 68% of all plug-ins, a higher growth than the PHEVs which increased by 81%, that composed the remaining 32% of all plug-ins. The forecast for 2020 and beyond delineates an additional growth of the market. In this European scenario, Tesla Model 3 was again the best-selling EV, with 95,247 sales in 2019, 22,137 of them in the month of December (Kane, 2020).

This situation generated its consequences even in Tesla's stock price. Tesla' stocks were worth over $500 in January 2020 and the all-time highest Tesla stock price, $917.42, was registered on February 19, 2020 (MacroTrends, 2020). These numbers made Tesla the most valuable car company in America, despite General Motors sold twenty times more cars and Ford six times more. On January 2020, Tesla market capitalization was $93 billion, GM was $50 billion and Ford $37 billion (Lee, 2020).

The following chart shows the global market share of Plug-In electric vehicles for 2019, that saw Tesla at first place overall:

Figure 10: Plug-in electric vehicle market share, worldwide, in 2019



Retrieved from *statista.com*, 2020

Tesla, marketing and social media strategy

The specific marketing strategy adopted by Tesla reflects what it was defined as its general business strategy, namely, to start producing high-end luxury vehicles and then trying to lower prices with the objective to lever economies of scale and thus increasing sales in mass market. As seen above, the project is not straightforward to be applied and it presents different obstacles and impediments that Tesla has to overcome. A solid help in that sense comes from marketing and communication, that can be directed to convince stakeholders about the goodness of the general vision of the company and the quality of its products. In its registration statement in 2011, Tesla defined its main marketing goals as to (Tesla Motors, Inc., 2011):

1. Generate demand for vehicles and drive leads to sales team
2. Build long-term brand awareness managing corporate reputation
3. Manage existing customer base to create loyalty and referrals
4. Enable customer input in product development process

These four main key points are useful to detect how Tesla used marketing tools to become a recognized and generally appreciated brand worldwide. Throughout the years of its development and growth, Tesla tried to put in practice the innovative strategies and methods of young and agile Silicon Valley companies to a more traditional sector like the automotive one. Marketing does not constitute an exception in that sense. In fact, a high-end strategy can be seen as something unusual in the general automotive panoramic, whose main goal is to directly aim at selling the higher number of cars and thus focusing on mass marketing. On the contrary, Tesla wants its brand to be perceived as attractive, distinctive and environmentally friendly by forward thinking consumers. In this way, it attached itself to one of the most relevant trend of these years and eventually of the decade, that is the reduction of CO2 emissions and the implementation of the "green economy" (Shipley, 2020), whose hint of its relevance in transport is shown by the following data: "*In the European Union, the transport sector accounts for 30% of the total energy consumed and 22% of Greenhouse Gas (GHG) emissions. In Spain, it accounts for 40% of the total energy consumed and 27% of GHG emissions. In the USA, it accounts for 29 % of the total energy consumed and 26% of GHG emissions*." (Iberdrola, s.d.).

More specifically, what the company did especially with its first product is to target that segment of "early adopters" present in the premium sports car consumer niche segment. Later on, it distinguished three main segment market clusters (Mangram, 2012):

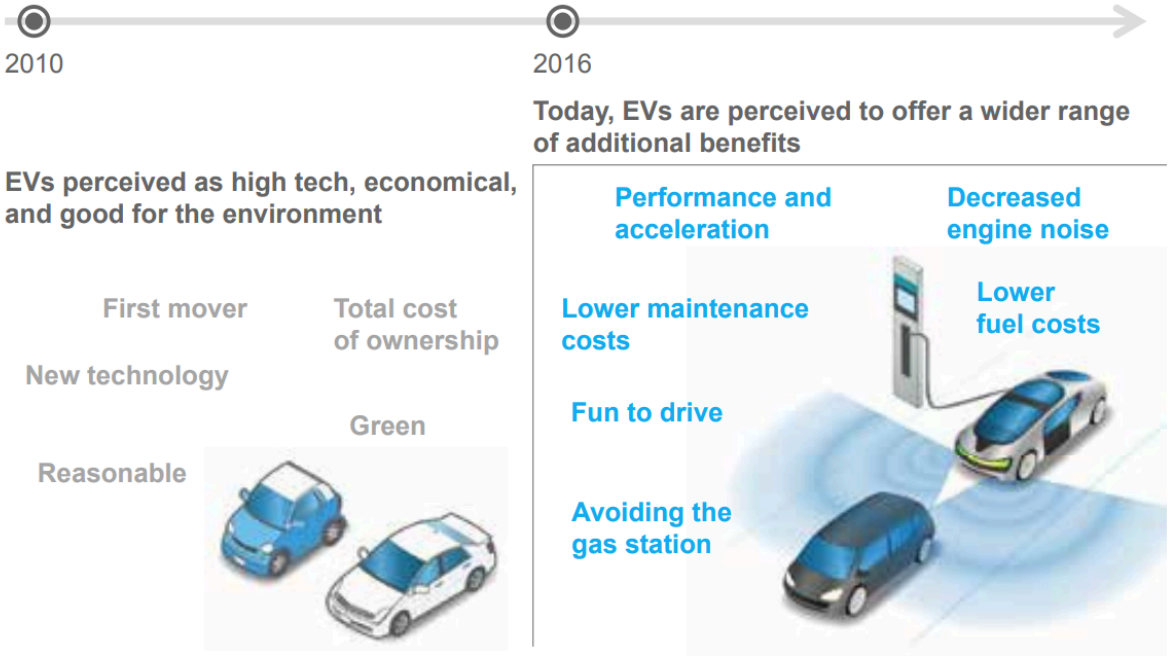
1. High-end premium sports car market: the very first objective for the initial phase of Tesla
2. Luxury vehicle sedan market: it can be seen as a further step that embraces a broader set of consumers and competitors
3. Mainstream vehicle consumer segment: the last phase of Tesla's expansion is to produce a great quantity of cars for entering the mass market.

Moreover, a McKinsey research on e-mobility (McKinsey&Company, 2017) delineated the main characteristics that contribute to define an EV consumer nowadays. The research was conducted through a global online survey of 3,500 EV consumers across US, Germany and Norway, and other 3,500 in China, and

made use of statistical approaches to derive a data-driven composition of current and potential EV consumer segment and the comparison of their perceptions towards e-mobility. In this study, four megatrends in automotive industry were firstly highlighted: Autonomous, that comprehends the set of features that makes the car a smart product, such as the use of central control unit, self-driving system and in general the latest software implementations; Connected, referring to the advantages that can arise from an EV ecosystem, for instance increasing the convenience of charging; Electrified, that deals with the goals of reaching the lowest level of emission possible thus protecting the environment; Shared, which is the trend that it is developing to tackle the total cost of ownership for cars through a sharing-economy system and to allow consumers to test and access multiple vehicles. It is possible to refer to these four areas as "ACES". More specifically, the study also highlights the fact that the "electrified" megatrend is composed by additional four areas of development, that are: the increasing demand for e-mobility, improvements in technologies and thus decreasing of production prices, urbanization increasement that would lead to both the need of cleaner air and the coverage of shorter distances, and finally the regulatory forces and governments rules that incentive green-economy and mobility and penalise high levels of car pollution.

In addition, McKinsey' research shows the guidelines to understand e-mobility market form a consumer point of view. It does so by highlighting more in detail fundamental levers, that embed the understanding of consumers' preferences for e-mobility and so the proposition of the right vehicles in different segments. What emerged from the studies is that, first of all, a perception gap exists between actual and potential consumers of EVs, with the former that are more satisfied in driving ranges and infrastructures terms than the latter. This can be partially explained by the fact that potential buyers are concerned about higher maintenance costs and pleasure in driving for EVs, despite these two factors are denied by different studies. However, McKinsey showed that a growing portion of potential buyers in US and Germany are referring to more perceived benefits when it comes to electric vehicles, including also acceleration and performance. The following image specifies these ones:

Figure 11: Perceived benefits of potential consumers for electric vehicles overtime

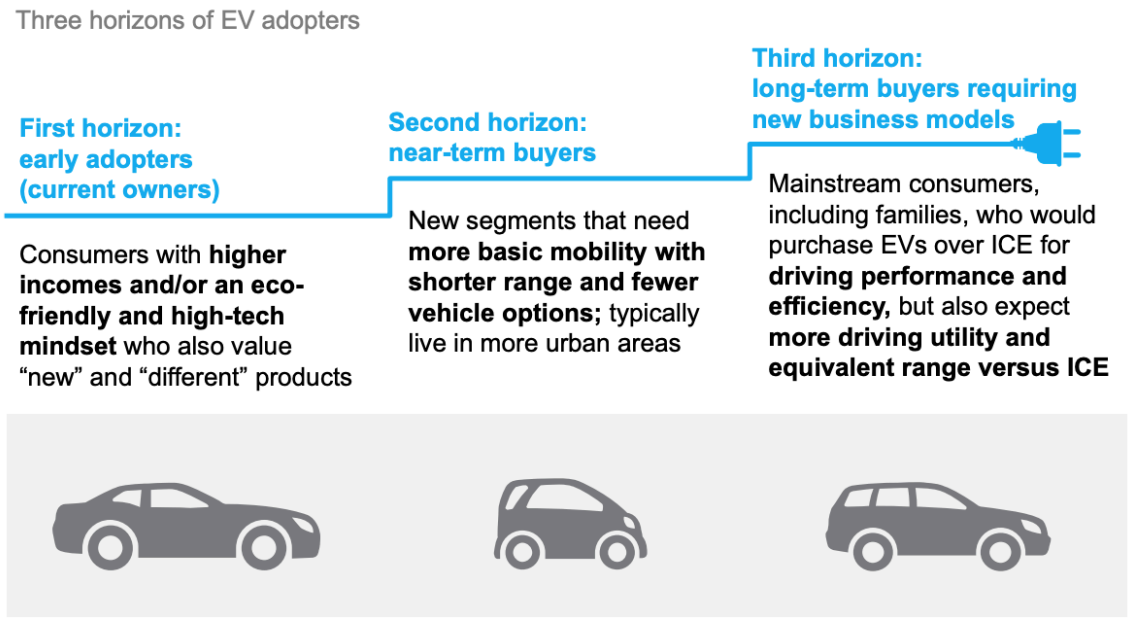


Retrieved from (McKinsey&Company, 2017)

What is more is that the trend can be also attributed in part to Tesla. Again, McKinsey conducted an experiment where respondents were exposed to non-branded EVs next to a brand logo, with random combinations. The result was that the willingness to buy an EV increased from 20% to 40% when the vehicle model was shown in combination with the Tesla badge.

Finally, another finding remarks the differences in needs that occur in the abovementioned segments. These are identified as “early adopters”, that are the tech-savvy and high-income consumer who already own an electric vehicle; the near-term buyers, that is the segment of potential consumers that are willing to purchase EV for short range; finally, mainstream segment, whose main drivers towards the purchase of an EV are its superior or comparable efficiency, performances and driving utility to a normal internal combustion engine car. These results are shown in the image below, and they partially match with the segmentation proposed for Tesla with its products:

Figure 12: Three EVs segment characteristics



Retrieved from (McKinsey&Company, 2017)

As seen above, the whole process, especially in its first stage, can count on an active involvement of costumers in promoting Tesla brand and product. In fact, the advantages of targeting tech-savvy and prepared consumers is that they can actually become ambassadors and advocates not only of the company in general, but also and mainly of its cause, since they are enthusiasts of how it is moving towards a goal that they share, projecting themselves as belonging to a system that want to revolutionize business and transports for the better. That is an effective way to turn costumers into fans (Marketing Examples, 2019). The involvement of consumers as leads and the creation of active referrals is crucial to develop a word-of-mouth strategy whose main consequence is to enhance company's reputation and support the perception of brand as prestigious and distinctive. To attract this specific type of clientele, Tesla adopted strategies that in a way are different to the competitors, taking into account that the investments are also different since the targets are not the same. As an evidence of that, a valid example is how Tesla handles its sales and service strategy. In contrast to competitors, that sells cars mainly through franchise dealership, Tesla levers more on direct selling, where the level of consumers involvement is higher, and the bureaucracy is remarkably lower. In fact, it mainly sells online, with easy-to-follow steps that guide to purchase through few links. Moreover, Tesla managed to make its stores and showrooms an enjoyable event: the former are located in notorious "trend-setting" markets, and the latter not only are directed to actually show the product at issue, but also to engage and inform actual and potential customers about the advantages of driving an electric car and owning a Tesla, also through test-drives, direct contact with Tesla's employees and catchy settings (Mangram, 2012). In this way, an "inbound" sales model is implemented, that attracts informed customers and guide them through a more conscious purchase process (Shipley, 2020), together with tackling the problem of misperception that is demonstrated to be still in place among potential buyers. The efforts made Tesla Model 3 and Tesla Model S gain 8 positions in the 2020 Consumers Reports' ranking, that valued the score of 33 cars on four factors, namely safety, test

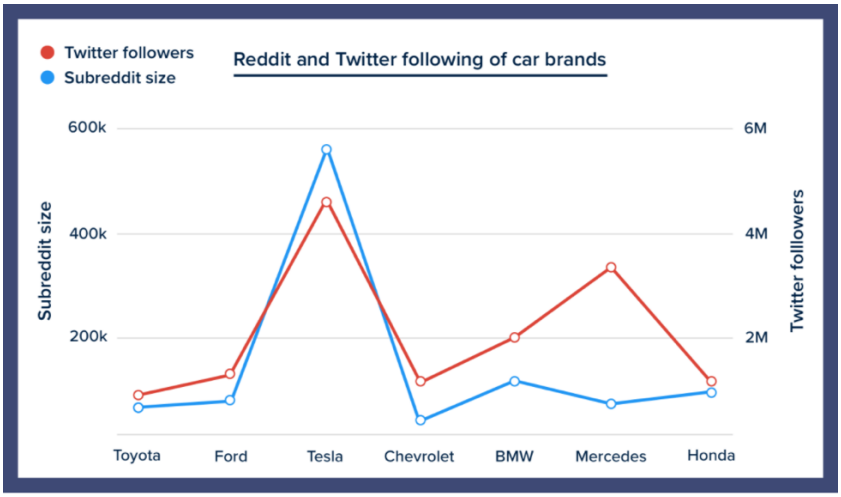
drive, owner satisfaction and reliability surveys, thus allowing the brand to be ranked 11th overall (Matousek, 2020).

So, Tesla relies a lot on consumers becoming advocates of its marketing and promotion strategies. In fact, in this way, Tesla is capable of levering on a powerful tool that is customer advocacy marketing. This can be defined as *"an advanced form of market-orientation that responds to the new drivers of consumer choice, involvement and knowledge. Customer advocacy aims to build deeper customer relationships by earning new levels of trust and commitment and by developing mutual transparency, dialogue and partnership with customers."* (Lawer, 2006). These active catchment areas generate a big amount of multimedia contents online in different platforms: Tesla fan pages, video trials, test drives, opinion pools, webpages and general social media comments proliferate on the Internet, generating a huge amount of advertisement that is not directly made and thus paid by the company, but it is entirely produced by product's owners and fans all around the world. As said, the easiness of sharing and the giant resonance that these tools provide can make product become famous in a very little time, with advantages that arising from a lot of perspectives. By looking at this scenario, it is possible to recognize some aspects of Tesla's marketing model, as they put their customers in a level where they not only consume the products but contribute to their realization and implementation. This strategy makes sense since, as seen above, consumers are more willing to be guided by previous users in the purchase process of a good or service. Moreover, other advantages can arise from this marketing form. In fact, an effective customer advocacy can contribute to increase brand value, allowing the firm to ask for a higher products' price than the competitors since the additional focus on service quality is well valued by consumers. Plus, a focus on that sense can enhance consumers' trust, that is crucial since a user that have trust on company is not only more likely to purchase or pre-order a product without precisely knowing it, but also makes its substitution with competitors' ones harder. These factors are also drivers to build loyalty, that is arguably one of the highest levels of engagement that customers can experience with a brand. Loyalty would create a bond that strongly links customers and brands, and the former become real participants in the product journey process, in the sense that they provide constructive feedbacks and share their experiences with potential buyers, in addition to developing faith and feelings towards the brand and having one of the highest willingness to pay for company's products (Duel, 2018). An additional focus is reserved to active loyal, namely that portion of loyal customers that are highly motivated to give their opinions about their experience with the product to a vast public, levering channels like social media and online reviews, which are becoming increasingly important to companies, especially for the ones that aim at focusing their attention on the development of advocate customers (Abu-Alhaija, et al., 2018). These then lead to the creation to of the social phenomenon of communities, that is defined as *"a place where consumers with common interests, needs and experiences come together to share ideas, communicate easily, create information and co-develop products"* (Duel, 2018). These facilitate the communication among people, thus in many cases creating a network of people that not only are interested in purchasing the brands' goods, but also follow its progresses and movements in a long-term time range. In many cases, a direct mutual relationship and sharing of ideas between brand and the community can be created, thus reinforcing and inspiring both parts.

The concepts of community and advocacy are particularly important for a company like Tesla, whose business, ideas and actions embrace not only economics, but also actuality themes that go in the directions of mega-trends like sustainability, that is defining the evolution of society in general. So, Tesla is able to embrace different subjects that go beyond the automotive sector, thus allowing the company to engage with diverse and bigger communities. In other words, as said, what Tesla would like to achieve with its business is not only a development of a car company, but more in general the evolution of an entire ecosystem, the transportation one, in the direction of some of the most important green-economy trends, thus enlarging and broadening its catchment area and increasing the level of participation and commitment of their fans, advocates and customers. As seen, the perception of Tesla as a premium brand contribute to attract interest and allows to develop positive feelings in the social imaginary.

Turning these concepts into numbers, to understand the impact of Tesla in online communities, it is sufficient to say that the communities of Tesla-centric YouTube channels count about 250 million subscribers combined for a total of 400 million views, and these are only number for active fans on YouTube. In addition, Tesla's YouTube channel is one of the most popular among the other car makers' ones. Moreover, social media popularity in other channel is considerably high, and a myriad of blog posts containing Tesla's stories are present on the Internet (Evannex, 2019). The following image highlights the fact that Tesla performs the best on two relevant Internet social media and network community domains, namely Twitter and Reddit:

Figure 13: Twitter followers and Subreddit size of car brands

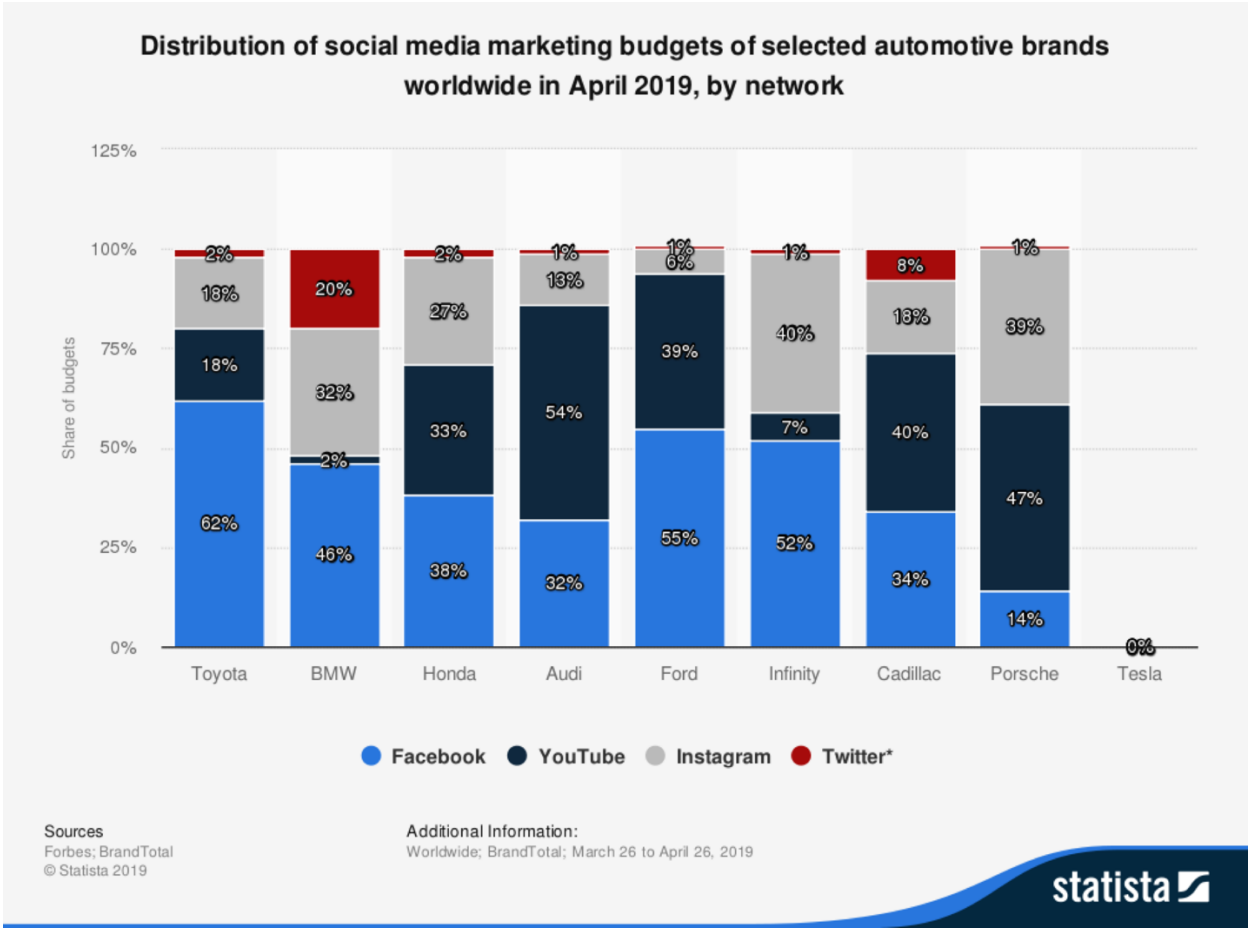


Retrieved from (Marketing Examples, 2019)

What is more is that this form of advertising, that is closely linked with the concept of the e-WOM, is really efficient in terms of costs for the companies, in addition to be particularly effective for company goals (Marketing Examples, 2019). In general, social media marketing can be conducted through paid advertisements in selected platforms, but in the case of Tesla, no money were invested on that. Excluding the offering of goods for owners that bring new customers to conversion through referral codes, the social media

marketing expenses for Tesla are zero (Evannex, 2019). This is not the case of some of their competitors, as it is shown from the bar plot below:

Figure 14: Social media marketing budgets of selected automotive brands worldwide in April 2019, by network



Retrieved from *statista.com*, 2019

That is a clear representation of the efficiency of Tesla's online presence, as it managed to become one of the most popular brands in social media networks without paying for that, but just counting on the almost spontaneous collaboration of its enthusiasts.

It follows that the biggest investments in marketing and sales are made to the realization of company-owned stores, galleries and service centres, so in on-field experiences that are fully controlled by Tesla and thus do not put their main focus into the auto dealers' market. In particular, over 200 Tesla stores are present worldwide, and 112 of them are located in the United States. Moreover, as seen, Tesla has its own website, where consumers can easily buy their model after having customized it in accordance to their tastes. That is another proof of the Palo Alto company's attention on customization and general purchase experience satisfaction (Business Strategy Hub, 2020).

The great online presence and resonance that Tesla actually has is largely given by the behaviours, comments and ideas of arguably its main figure, the CEO Elon Musk. As seen above, Musk is leading different companies and projects, whose common denominator is found in innovation. He was behind the creation and spreading of the online payment system PayPal, that sold in 2002 for $1.5 billion. Then, he started SpaceX, a spaceflight

company whose primary goal is to contribute to human colonization of planet Mars. Moreover, apart from Tesla, he is involved in another project that promotes the diffusion and utilization on a large scale of electric energy; this project is SolarCity, an energy company that provides low-cost solar services. Moreover, he is also working on the Hyperloop, a tube-based superfast train that would supposedly revolutionize the way people use transportations services, as it is claimed to take passengers from San Francisco to Los Angeles in 30 minutes (Wittmeyer, 2012).

Despite these projects are interesting and ambitious on their own, what contributes to the actual popularity of Musk is its unique way to interact with his audience, that goes beyond the mere promotions of his projects and thus creates a deeper contact and level of communication. Firstly, his presence in social media is constant and active, in the sense that he is always open to give answers on disparate users' questions and engage them in virtual conversations. The main social media through which he operates is Twitter, which facilitates a direct and instant interactions to followers since its formats allows to write a limited number of characters and to answer directly to every post through comments and "retweets". Going deeper into the way he communicates in this platform, it is possible to better understand why this is actually the most important channel for him and the companies he owns and works for. In fact, he not only uses Twitter to post information, but only to ask for direct feedbacks to their followers, thus encouraging a discussion and so enhancing the interactions and attention towards the content of the tweet. Responding and engaging with the audience allows to build trust, develop recognition and express approachability and authenticity. The latter point is another distinctive point of Elon Musk's communication on Twitter, as his answers and posts are perceived as honest or not filtered by any strategy behind. In addition, he has no problem with embracing critics and talking about events, mechanisms or project that did not work out as he was expecting. This again is valued by people especially in social media domains and contribute to build feelings and a deeper connection with the brand and the public figure. Eventually, these factors will reinforce the brand value (YS, 2018).

In a business perspective, what is ultimately tried to be achieved through the use of social media is the highest coverage possible without costs, meaning a maximization of efficiency for advertising and communication. As said, there is more than just promoting products and brand: Musk and Tesla also gave their help to solve humanitarian problems, like in the 2018 Campfire in California or in Puerto Rico to help rebuild power infrastructures after a hurricane (Hughes, 2019). Nevertheless, these informal and direct way of interacting can have its drawbacks: in 2018, a controversial interview of Elon Musk on a live web shop make Tesla share prices to drop and cause several problems in public relations plans of the company (Folschette, 2019). Moreover, there were many cases in which his opinion was not shared by the majority of people. So, to have a company's reputation that deeply depends to the actions and words of a single figure has its pros and cons. However, this path pays back at least in terms of numbers: as in April 2020, Elon Musk' s Twitter page counts over 33 million followers, while Tesla's one over 5 million, plus, as seen, advertising and social media costs are close to 0, a pretty efficient strategy considering also that the cost for a Silicon Valley marketing team is about $40 million (Popkin, 2018).

The product: Tesla Cybertruck

As previously mentioned, Tesla has produced five models so far, namely Roadster, Model S, Model X, Model 3 and Model Y, and other models are actually on development phase. Among these, there is the electric pick-up truck whose name is Cybertruck. Unveiled on November 21, 2019 in Los Angeles County, California, the Cybertruck will be available in three versions. The first one features a single motor rear-wheel drive or RWD, that goes from 0 to 96 km/h in less than 6.5 seconds and has an autonomy of more than 250 miles and is planned to be released in late 2021 at $39,900 before incentives. The second features a dual motor all-wheel drive or AWD, going from 0 to 96 km/h in less than 4.5 seconds with an autonomy of more than 300 miles; the third is provided with a tri-motor AWD, covering the distance in less than 2.9 seconds having an autonomy of more than 500 miles. In addition, Cybertruck will be set for 6 people and has a bed of about 2 metres, and a really strong body that is claimed to resist even to bullets. Together with these features, there are two other critical points that needs to be addressed when dealing with a truck: these are its hauling and towing capability. For what concerns the Cybertruck, its towing capacity is of about 3400kg for the single motor version, while for the AWD models, it goes up to about 6350kg. In addition, the payload is set at about 1587kg for each version (Leanse, 2019). The production of the last two models will start in late 2022 and will be priced at $49,900 and $69,900, respectively (Electrek, 2020) (Tesla, s.d.). Tesla already started the pre-orders for the Cybertruck, that consists in a refundable reservation of $100 regardless of the version chosen. The pre-ordering process is completely conducted through Tesla's website, where users can choose their favourite model and customize it based on their preferences. What is more is that, after the reservation, a referral link is released after having created an account on Tesla website, that gives the opportunity to win a Model Y SUV or a Roadster sports car if other people use that link to buy a Tesla vehicle or a solar panel (Matousek, 2019). Just after few days from the unveiling, pro-orders for Cybertruck overtook the amount of 200,000, with the majority of them that were for the dual and tri-motor version, with only a few portions of 17% was for the single motor one (Electrek, 2020).

Even for the Cybertruck, the role of Elon Musk was and is still crucial not only for its promotion, but also for its development. In fact, all the data previously mentioned were reported on his Twitter page, together with other tweets that anticipated the release of the pick-up. What it more is that Musk announced his willingness to scout a new location in Central America for another Gigafactory that would be built to start the production of the Cybertruck, generating a lot of reactions towards users, that also provided reasons why he should pick their cities or states for its construction (Silver, 2020). That represents an example on how Elon Musk deals with Twitter in different stages of the product journey.

As will be further discussed, there are many reasons that makes Tesla Cybertruck an unique product: first of all, its peculiar design is definitely different from other similar vehicles, thus contributing to produce more rumours and anticipation in that sense; secondly, its unveiling generated a buzz on the Internet since the claimed bulletproof windows shattered during a demonstration. Naturally, the images went viral and were spread out rapidly, thus again attracting visibility to the pick-up truck. These factors then need to be inserted

into the context of the American pick-up truck sector, that is one of the most historical, traditional and lucrative automotive market of the United States of America.

The market-edged nature of Cybertruck

At a first glance of the model at issue, it is straightforward to understand why it is considered a unique product. For this purpose, the following is an image of the Cybertruck:

Figure 15: Tesla Cybertruck

Retrieved from *businessinsider.com,* 2020

What firstly captures the attention and makes a strong visual impact is undoubtedly the peculiar design of the car, where angular and edgy shapes make it not only easily distinguishable, but also to be perceived as a futuristic and in a way more focused on essentiality and functionality, making it look resistant and captivating at the same time. The shapes of the Cybertruck contributed to make a breach in pop culture, since many people saw in the vehicle some references of past futuristic cars that appeared in cult movies: for instance, Elon Musk itself tweeted that the "*Cybertruck design influenced partly by The Spy Who Loved Me"* (Musk, 2019), a 1977 James Bond movie; in addition, others saw resemblances with model cars from cult movies like Blade Runner, Mad Max: Fury Road and Akira, and even from the videogame Cyberpunk 2077 (Young, 2019). Another relevant comparison was the one that saw the Tesla pick-up truck to the 1981-1983 DMC DeLorean, a vehicle made iconic by the movie "Back to The Future": what was observed is that they both were and are disruptive in the auto industry for their respective eras due to their characteristic features that made them easily

recognizable, among them the design and the stainless steel colour; some comparison with the two main figures, DeLorean an Musk were made too as they can both considered as flamboyant personalities that embrace risks, with the latter that however sold more cars compared to the former (DeBord, 2019). This is another evidence of the fact that the Cybertruck itself, with its extreme shapes and edgy design, generated a buzz in the Internet and thus curiosity among the brand Tesla in general. As an evidence of the contribution of the online network, suffice is to say that even Twitter users contributed to some ideas for the realization of the truck back in 2018, in perfect Elon Musk's style (K10, 2019).

In addition, more considerations need to be taken into account: the Cybertruck does not provide different colours, thus the stainless steel shown in the image is the only version of exterior colour. This leaves room for customizations and decreases environmental impacts arising from vehicle painting. Moreover, there are not any external rear-view mirrors, as they are replaced by a camera-based and digital setup, available on the screen, a 17-inches touchscreen installed in the front next to the driver. The lighting system on the front will not need any additional supports and is capable of ensuring good visibility on off-road trails. What contributes to the external perception of robustness are also the big wheels that sustain the vehicle. The windshield is also really spacious to the extent that it resembles a transparent roof, thus still giving points in terms of driver's visibility and aesthetics (Leanse, 2019).

Tesla Cybertruck is thus defined as a full-fledged polarizing car model, as some people were enthusiastic about its futuristic appearance and technological-advanced software and features while other were really sceptical about both the vehicle, which seemed uncomplete and so looked more like a prototype, and its impact on the American pickup market, whose characteristics will be further discussed. Nevertheless, in order to get a more complete overview of the scenario, it also has to be considered the business strategy of the Cybertruck and the segment Tesla is targeting with the pickup. It is undoubtedly true that this vehicle wants to represent a disruptive element in its market, as Elon Musk himself stated during the unveiling: *"Trucks have been the same for a very long time […]. Like a hundred years, trucks have been basically the same. We want to try something different."* (Musk, 2019). However, that truck proposition can be seen as a starting point for a broader strategy of Tesla to enter the market, starting with a product whose objective is not be targeted for mass-market, but rather for a niche of enthusiasts and eventually curious consumers who want to change its choices and persevere in the attention towards the environment without renouncing to the standard performances and functions of a truck. An analysis (Smith, 2019) for CNN highlights this point: what it is stated here is that for Tesla to enter in the pickup market, a lookalike product would not have been noteworthy plus not reflecting the business style of Tesla and its CEO. Rather, the proposition of a distinctive product would help to reach a niche a target of 5% to 10% of potential pickup buyers, which can be still a profitable percentage considering the size of the market. Once having established with the vehicle, the next step would be to build credibility, that could then lead to the proposition of another pickup model that will try to capture a higher user base and potential buyers. In addition, other analysts of the sector agree on the fact that potential Cybertruck buyers are mainly enthusiasts Tesla customers, pop culture lovers celebrities and green millennials, thus not constituting a thread to the pickup market as it is known nowadays (Wayland, 2019). This was in a

way confirmed by Musk during an interview, where he admitted that the Cybertruck could represent part of a bigger strategy in the sector: *"You know, I actually don't know if a lot of people will buy this pickup truck or not, but I don't care […]. If there's only a small number of people that like that truck, I guess we'll make a more conventional truck in the future."* (Musk, 2018). In other words, Tesla seems to re-propose its general strategy in the reduced scale of the pickup market: start as a niche, then try enlarging the market to target the mass with a more standard product, all while converting the network into the electrification and the adoption and installation of "in-car" software.

The Cybertruck unveiling

The episode that amplified the already relevant rumour over the new Tesla pickup truck, with all of its mediatic consequences, was undoubtedly the one that happened during the unveiling that took place on November 21st, 2019 at Tesla Design Center in Hawthorne, Los Angeles County, California. The presentation showed the abovementioned technical characteristic of the Cybertruck and featured different demonstrations. Firstly, as said, Elon Musk began by saying that, throughout the years, pickups have always been the same and no innovations were proposed. Then, after the entrance of the truck in the scene, the first demonstration saw Tesla's head of design Franz Von Holzhausen hitting a claimed regular door first and then the Cybertruck one with a sledgehammer, to show the better performance of Tesla's pickup skin compared to its competitor. In addition, as a demonstration to show the bed capacity, an ATV was brought into stage, drove on Cybertruck's bed and put into charging through an electrical outlet on the side of the vehicle. Other highlights were videos of the Cybertruck towing a Ford F-150 truck and a drag race between the truck and Porsche 911, with the latter that even received a head start. But what captured the attention and generated controversies and buzz reactions was the demonstration whose aim was to show the resistance of the claimed unbreakable Tesla Armor Glass, equipped in the Cybertruck. The first attempt was conducted in separate glass by throwing a metal ball from a considerable height to a regular glass first and then to Tesla Armor Glass, following the same format of the first demonstration. The result was that the former was considerably damaged while the latter was intact. However, when Von Holzhausen threw that same metal ball to the Cybertruck glass, it shattered, surprising everyone from the audience to Elon Musk. The attempt was repeated for the back window, but the outcome did not change. The words of Musk after that scene were the following:

"We threw wrenches; we threw everything, […]. We even literally threw the kitchen sink at the glass and it didn't break. For some weird reason, it broke now. I don't know why. (Musk, 2019).

Figure 16: Tesla Cybertruck glass shattered during unveiling

Retrieved from *commons.wikimedia.org,* 2019

The whole presentation lasted less than half an hour, but its consequences are still lasting. Attempts of explaining the episode succeeded. On Twitter, Musk blamed on the misshaped order of the demonstration, claiming that the sledgehammer caused an unseen crack that then led to the shattered glass: *"[...] Sledgehammer impact on door cracked the base of the glass, which is why the steel ball didn't bounce off. Should have done steel ball on window, then sledgehammer the door. Next time."* (Musk, 2019).

Despite that, what can definitely be considered as a bad episode for Tesla's reputation, it turned to not having that serious negative repercussions, but rather generated even some positive results, thanks to the management of communication by the company and Elon Musk. The worst repercussion was on Tesla's stock price, that after the release dropped by 6,1% to $333.04. Despite that, stock going down was also a consequence of the fact that is has been up for a long amount of time, and does not reflect a long-term decrease (Root, 2019). In fact, just days after the unveiling, Musk tweeted that the Cybertruck received more than 200,000 orders, with 41% of them that were for the triple-motor version (Liao, 2019).

This episode can make think about how a positive outcome could have been generated by a negative accident, as it seems to be contrary to logic. An explanation of that can be found again on the strong relationship and loyalty that customers have and feel towards Tesla, which is able of compensating bad events as the trust towards the project overtakes the difficulties arising in the path. Others (Bariso, 2020) specifically pointed out the ability that Elon Musk and the company had on having embraced the moment rather than having hided it,

making it work for them rather than against them; in fact, on January, Musk announced the selling of a t-shirt depicting the image of the Cybertruck's shattered glass. This showed an attempt to keep the episode real, building authenticity and showing that imperfections are possible even for high-end companies. Therefore, the episode was also described as a case of emotional intelligence, or EI, defined as *"the ability to engage in sophisticated information processing about one's own and others' emotions and the ability to use this information as a guide to thinking and behavior. That is, individuals high in EI pay attention to, use, understand, and manage emotions, and these skills serve adaptive functions that potentially benefit themselves and others"* (Mayer, et al., 2008).

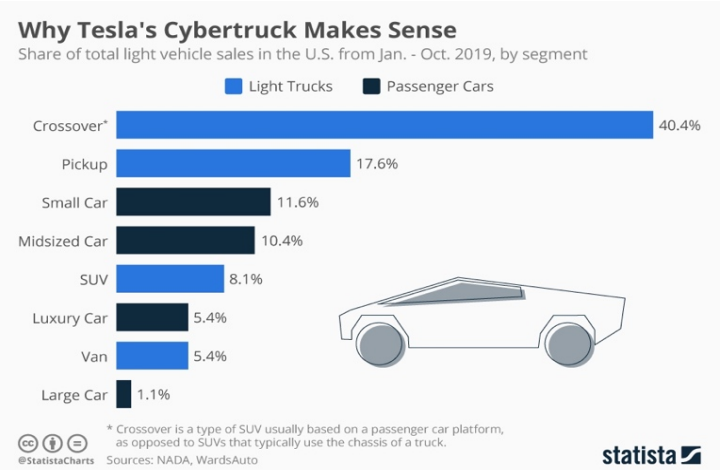
That skill is part of Elon Musk's personality at least when relating with his audience, and it represents another contribution towards the building of a solid engagement and relationship between the consumers and the brand, together with showing the good outcomes that arise from being authentic with the public.

The Electric Pickup sector

Before going into the specific electric sector, it is relevant to inspect the general characteristics of one of the most important market in the automotive sector in the United States, namely the pickup one. For definition, the pickup truck is "*a small vehicle with an open part at the back in which goods can be carried"* (Cambridge Dictionary, s.d.).

Historically, this vehicle type is very popular among northern Americans, especially for the ones that live in rural areas or often travel off-road. Its production in the US market boomed in 1963 thanks to the introduction of the "chicken tax", that banned the import of foreign pickups, and was reinforced in 1973 after the introduction of the Corporate Average Fuel Economy policy, that hit many classes of large passenger vehicles but was less strict for the pickups (Statista, 2020). These were some of the reasons that laid behind the popularity of this model, contributing to become an iconic symbol of authentic American lifestyle. Economically speaking, the pickup market is one of the biggest and most lucrative in USA. For what concerns its size, the plot below explains the relevance of the sector in the American automotive panorama:

Figure 17: Share of light vehicle sales in the U.S. from January to October 2019



Retrieved from *statista.com,*2020

Together with highlighting the second place for pickup in the American market, only overtaken by Crossover that comprehend different types of vehicles, the plot shows also that the entrance of Tesla in this market makes sense in terms of the objective of becoming mainstream and thus enlarging the production.
In terms of number of sales, pickup trucks went from about 2.5 million units of 2015 to 3.1 million in 2019, meaning a constant growth overtime that has not saturated (Wagner, 2020). What characterizes this market is also the undisputed dominance that some models of major car makers have established throughout the years and decades. In fact, nowadays, there are essentially three models that are dominating the market: these are the Ford F-Series, Ram Pickup and Chevrolet Silverado. To put that in measurement terms, from January to September 2019, the first sold 662,574 units, the second 461,115 and the third 412,258. This means that roughly half of the U.S. pickup sales in 2019 were made up just by these three models.
Moreover, for what concerns profitability, it is suffice to say that General Motors, that owns Chevrolet among others, reported to make a profit margin of $17,000 for each pickup sold (McCarthy, 2019), considering a general average price of $29,585.1 in 2019 (Statista, 2020). Therefore, these numbers suggest a market composed by really traditional buyers, who are risk-averse and tend not to substitute their pickup with another vehicle. Traditionally, the average pickup buyers are perceived as ones that carry family, tow trailers and load the bed with a lot of objects, meaning a small room for compromises despite their attention to the real value of purchase and trade-offs (Williams, 2019).
In this context, the development of the electric pickups is starting to take place. This can be considered a subset of the broader pickup sector and its evolution is still in an initial phase. In addition, what needs to be highlighted is that the switch towards the electrification would also change the way consumers perceive and use pickups. Again, misperception can be addressed to change the perspective in favour of the electrification. In fact, analysis estimated that just about 10% of the pickups sold in the US are used strictly for working purposes, with the rest that use them mainly for everyday family activities like going to sports event and are more concerned about the aesthetic of the car, with an additional focus on the ride quality and driving experience (Gertner, 2020). That tendency could better spouse the concept and the perception of some electric vehicle automakers, whose one of their main aim is to suggest a smart purchase without renouncing to the exterior pleasures arising from owning a car of such type. In other words, the initial niche the electric car manufacturers could target can be the one that actually show the real usage Americans do of the pickup.
Since these good premises, not surprisingly, the Cybertruck raised the attention of fans and enthusiasts, but also of other car manufacturers, both established companies and rising start-ups, that observed Tesla's moves and are working as well towards the proposition of their all-electric pickup models. In other words, Cybertruck will not be the one and only model that will operate in a market that showed to have potential for several reasons, not least the environmental and sustainability ones, and whose sales forecast should not exceed 70,000 units a year according to studies (Lienert, 2019). Some of the most relevant models that are expected to compete in this sector, except the Cybertruck, are (Valdes-Dapena, 2020) (Gorzelany, 2019):

1. Rivian R1T: The Michigan-based start-up is working on the model called R1T, that according to the company's website it will be provided with a quad-motor all-wheel-drive system, going from 0 to 96km/h in less than three second and towing a maximum of slightly less than 5000kg. Therefore, the vehicle should be able to cover a range of about 400 miles with a single charge. The starting price should be therefore set at about $70,000 and the production could start in late 2020. A massive advantage that this company has comes from two of its main important investors, that are Ford and Amazon. While the former would help with production needs and components, the latter would contribute to its know-how in the hardware and software for the installation of vocal assistant.
2. Ford F-150 Electric: An all-electric model of the best-selling vehicle in the entire U.S. market going even beyond the mere pickup sector, is expected to be released. On summer 2019, few demonstration videos showed a F-150 electric prototype towing a freight train filled with F-150s. Although there still are not a lot of information about this model, the production could start on late 2021 and the expectations for this vehicle are high. A straightforward advantage comes from the solid established position that this model has, thus potentially making a switch to an electric setting easier in terms of consumers' adaptation.
3. GMC Hummer EV: Even General Motor decided to move into the pickup truck market with its GMC Hummer EV. The information actually available about the vehicle are that it goes from 0 to 96km/h in about 3 seconds, it is provided with up to 1000 horsepower and the initial availability would be on Fall 2021. The revealing of this all-electric car, planned to be on May 20th, 2020, was postponed on a date to be determined (GMC, 2020).

Again, these are and will not be the only electric pickups that were announced or are planned to be produced in the following years: many companies have started to propose their electric truck for the American market and are working towards a complete switch to electrification not only for this segment, but also for the industry in general. However, the abovementioned vehicles are certainly some of the most awaited and most talked-about cars by potential consumers.

Application of Sentiment Analysis on product

Once having introduced both the characteristics of text mining and the product, the following section will aim at applying some of the features of sentiment analysis to Tesla Cybertruck, in order to assess the impact that this car had on the audience at an emotional and opinion level in the domain of one of the most important and most-used social media that is Twitter. Moreover, an overtime evaluation and evolution would be implemented to see how the focus of the audience shifted across different themes, that changed according to the eventual happenings and events that surround the product and its company. In fact, the analysis will also identify major elements, like relevant figures, tweets, strategic company decisions and general contingencies, that in a way were and are still associated to the Cybertruck. That step goes beyond a mere sentiment analysis, thus allowing to capture a useful and larger range of information and eventually to explain the reasons behind certain audience responses. As a matter of fact, relevant actions cause equally strong reactions, especially for a renewed and ambitious company like Tesla that, as seen, relies a lot on an aggressive and sometimes provocative communication strategy on social media. Moreover, understanding the opinion helps to reach specific above mentioned goals (Altexsoft, 2018) (Marketing Inside, 2017) that are really useful from a strategical point of view. Some of them, like the identification of trends, viral tracking and product monitoring, can represent a good target for a sentiment analysis conducted on a domain like Twitter. These objectives work well with an analysis whose aim is to detect debated topics and assess a value to an opinion.

Moreover, the analysis provides a combination of approaches to inspect the features on a deeper level and so to retain pieces of information that come from different sources. Besides, the definition of the word "analysis" itself, *"to study or examine something in detail, in order to discover more about it"* (Cambridge Dictionary, s.d.) encloses the general aim of an opinion mining study and, to do so, experimenting different approaches in different time periods can only contribute to enrich the contents and the findings.

The benefits for the latter are also amplified if considering that the object of the study is part of a market that is actually evolving and thus needs to have solid bases on which to start consolidating. So, understanding the audience's response about a changing trend is also an important aspect to take into account, as it goes beyond a single company's strategy and it is part of the cited elements that surround the Cybertruck. In addition, the social media dynamism of Tesla and its main figures like Elon Musk contribute to provide a great amount of information to be analysed, representing a factor that not only differentiate them from competitors, but also make them more easily searchable and in a way analysable, thus making a relevant impact on the available sources for this study. In other words, the more the information are made, posted and shared about an object, the more are the data available for the analysis, especially for text mining where large data sets conducts to large quantitative results and the possible application and more appropriate approach depend on the type of information available (MonkeyLearn, 2019). Having stated these premises, the following sections will go through the steps implemented for the sentiment analysis and the detection of some relevant topics, from data collection on Twitter to the visualization and inspection of findings, both considering the data as a general entity and splitting them according to specific reference periods.

Theoretical Background

As seen, the structure of the study is focused on the application of the necessaries steps that need to be implemented to conduct a sentiment analysis, from gathering to analysis and visualization of Twitter data with the aim of inspecting the sentiment at a tweet level. The gathering, that will be further discussed, was organized on a monthly base, in order to refine the understanding of data and eventually track the evolution of audiences' score and general news. Moreover, preliminary data visualizations tools would allow to actually give a first glance at the relevant focal points, trends and topics in the dataset. Nevertheless, to deepen the knowledge of data, the Latent Dirichlet Allocation unsupervised approach will be implemented, giving back some of the most relevant topics on tweets that need to be addressed and focused on. Thus, a specific sentiment analysis for them is conducted by creating a subset composed by the tweets containing the identified topic. The approach to the sentiment would be lexicon-based but organized in different levels. The first and most superficial one would detect the single words present in the tweets and gives back scores according to the dictionaries used that will be further presented. This method, together with being largely used in previous studies, is efficiently implementable in different domains and can give solid results. Therefore, its application on topics would allow to understand which of them are the real driver of the general sentiment results, giving a clearer idea of which factors are more appreciated by tweeters and users.

Although these methods already give meaningful results for a sentiment analysis, the study goes beyond the mere classification based on a single word-emotion association. To do so, another lexicon would be then implemented that catches more pieces of information inside a single tweet. In particular, the second and deeper level of lexicon-based sentiment analysis presented in this study would take into account form of languages that are unique to Twitter and social. In fact, together with considering the mutation that is given by scores of negations, amplifiers question marks and adversatives, the second lexicon will detect emoticon, emojis and slang words, giving back a more detailed and precise sentiment score for the tweets at issue, not for just a single tokenized word.

The combination of both monthly-based and general-based analysis would eventually broaden the relevant results, giving a more detailed overview on the tweets.

In addition, a glance to the sentiment analysis of relevant and abovementioned competitors will be implemented, in concordance to the actual possibility of getting a meaningful amount of data. In fact, as already mention and as it will be further discussed, the amount of information on Tesla, especially for models that are still not on commerce, is not comparable to the one generated for Cybertruck's competitors. Despite that, it is still possible to have an idea on how these models were and are perceived by users, thus another specific round of analysis would be targeted to this particular task.

The points touched in this section will be the object of discussion in the analysis. Moreover, it is relevant to point out that the entire study, so going from the gathering to the analysis and visualization, is coded on R language, a programming software whose characteristics can also fit for this kind of analysis.

Data Collection

As previously seen, the first step when performing a text mining study is data gathering. R allows to pull social media data, in particular Twitter ones, through the activation of a specific package that, after having received specific authorization, gives back a set of tweets that can be then analysed for academic purposes. To understand that in detail, the precise procedure is further presented.

When dealing with Twitter data, a Developer Application on the social network needs to be created. This step is crucial as it broadens the functionality of the app, among them the gathering of data. To create the Developer Application, the user has to communicate to Twitter the scope of the research, namely the reason why these data need to be retrieved. After having received each authorization, the social network releases specific keys and access tokens that identify users' account while conducting the researches. These codes are also the essential part as they allow to activate the API of Twitter, namely the Application Programming Interface. More specifically, the API is the code that, since it manages the access point for a server, allows applications to communicate with each other and thus exchange data. In other words, in this case the API is the tool that let the user access to a certain database. Therefore, the process starts by a request made through the app by a user, then it is received by the website's servers and data are returned in response through the API (Eising, 2017). In this specific case, Twitter API create a connection with R with their codes and Twitter, retrieving the specific information requested. In addition, the set of keys and tokens that are provided by Twitter App constitute what it is called "OAuth", that let the user to access some actions by using Twitter Developer Application. OAuth is frequently used when accessing contents of different websites or online services. In fact, this is the mechanism that allows not to create an account for each specific website, but rather to use the social networks' ones instead. An example of it is the common "sign in with Facebook" button that appears when accessing content of a new platform: in this case the service identifies the user on Facebook and then creates an account on its behalf. OAuth is moreover used when third-party apps wants to access some information of users' accounts, like the list of friends, to perform tasks after its authorization. Even in this case, a set of tokens limits the actions and the access the apps have on users' account (Hoffman, 2017). In the case of Twitter domain, it is like the third-party app is represented by the application created by the user, asking Twitter itself the access to some of its information.

To sum up, Twitter APIs have the dual function of both allowing the communication between the social media and other software and providing user's account the access to some publicly available pieces of information in Twitter, like tweets.

In a certain way, the digital era has paved the way to the development of a real "API economy" as this interface is often an object of exchange for both software and non-software companies, making it an appetible and, in most of the cases, profitable business. In fact, releasing APIs can be at the core of a certain company's proposition when its main goal is to distribute information for a specific task, or can constitute a substantial part of it, as other stakeholders can be interested in acquiring company's data for implementing their selling and thus increasing profits. In a certain way, the mechanism is reproduced in social media: despite their core contents are user-generated, the data generated on them can be shared through the Application Programming

Interface to companies that include them for perfecting and implementing their product or service proposition. In addition, social media can then attract more developers to its platform, thus potentially enhancing the final user experience. The API topic is very current as it is associated to both future protected systems like blockchain information exchange and downsizes arising from the excessive amount of information released to third-parties that sometimes led to data scandals (Yao, 2018).

Twitter Data Scraping

After having completed the abovementioned steps, the process moved on R, with the aim of pulling tweets for the analysis. So, what was done in this phase is called data scraping, that is "*a technique employed to extract large amounts of data from websites whereby the data is extracted and saved to a local file in your computer or to a database in table (spreadsheet) format*" (Webharvy, s.d.). One of the systems that performs this task on R is given by the package called "twitteR" (Gentry, 2015). What this package exactly does is to provide an access to the Twitter APIs already generated through the creation of the Developer Application and, after having set some specific instruction for scraping, it returns a certain number of tweets, from which text is extracted. To understand more in detail this functioning, the following is a portion of the code including an operation conducted with a function in "twitteR" (Gentry, 2014):

```
# Getting Tweets

install.packages("twitteR")
library(twitteR)

Consumer_key <- "xxxxxxxxxxx"
Consumer_secret <- "xxxxxxxxxxxxxxx"
Access_token <- "xxxxxxxxxxxxxxxxx"
Access_token_secret <- "xxxxxxxxxxxxxxxxx"

setup_twitter_oauth(Consumer_key,
                    Consumer_secret,
                    Access_token,
                    Access_token_secret)


Cybertruck_English_Mentions <- searchTwitter("cybertruck -filter:retweets",
                                             n = 5000,
                                             lang = "en",
                                             since = NULL,
                                             until = NULL,
                                             retryOnRateLimit = 120)

Cybertruck_English_Mentions_dataframe <- twListToDF(Cybertruck_English_Mentions)
```

The code clarifies the process: the package at issue is installed and then loaded into the software. The APIs' codes and tokens, retrievable from Twitter account, are then recalled, here they were hidden for privacy reasons. These fours codes then composed the OAuth, that is necessary to grant the access to Twitter data. The

actual function of scraping is then computed. This phase needs an additional focus on the instructions that can be set. In fact, "searchTwitter" function allows to specify several factors (Gentry, s.d.):

1. The query to issue to Twitter: This is a word that should constitute the object of the research, as it allows to identify the tweets and retrieve the ones that include that specific query, in a similar way to the functioning of a Google search. In this case, since the objective was to pull tweets that were about the Cybertruck, the word "cybertruck" was specified as a query and, in addition, retweets were filtered to give back a more precise result, especially for what concerns the text. The process was repeated by specifying the hashtag "#cybertruck", trying to capture that tweets that indicate the Tesla's model not just as a part-of-a-speech word, but as a single hashtag entity as well. This step broadens the outcomes in some cases, while in others it led to redundancies that were eliminated.
2. Number of tweets: this specific set the maximum number of tweets requested to the APIs. A number of 5000 can be considered as a good sample to compute a Twitter sentiment analysis, considering also the fact that the APIs does not always restitute that exact number indicated, as it can give back less of them.
3. Language: if specified, the function restitutes only tweets that the system recognized as belonging to that indicated language. In this case, since the research is more focused on American market, the tools for sentiment analysis are sometimes not applicable cross-linguistically and simply because it is the most widespread language in western society, English was specified.
4. Since/Until: this could give a time range of recalled tweets, but the actual APIs in possession for this study cannot recall this part of the function. Thus, this was not specified and put as "NULL". This argument represents one of the limitations that will be further discussed: the tweets that are restituted by the function are created in a time range of 8-10 days before the query request. For example, if the query was made on May 11th, 2020, the function pulls tweets that go from that date to about May 1st, 2020. This means that time range is limited and reprocessing the same function in a different period gives back different results. This factor represents both a point of strength and weakness for the analysis, as it allows to retrieve tweets from different times, but that range cannot be specified, and it is limited. This clarification is important to understand how the study was structured, since samples were retrieved and saved monthly.
5. Retry on Rate Limit: this number simply indicates the maximum number of trials that can be computed to extract the tweets. A high retry count slows the running time but eventually enhances the probability to complete the task.
6. Others: "twitteR" allows to recall more arguments, like geocode or IDs, but these were not considered as relevant for the study or their access was limited by the APIs in possession.

The function gives its results, but these are not still in the form of a spreadsheet or data frame. That is why the function "twListToDF" is used, as it converts the object created into a data frame that can be exported and saved as an .xlsx Excel file, among others.

In this case, the final result gave a data frame composed by 5000 rows and 16 columns. Despite that, not all of the variables are useful for the analysis: in particular, the one that is of real and actual interest is the "text" one, since it contains the words that will be then analysed. The columns "created" is also useful as it specifies

the date and time in which that tweet was created. The others are pretty much non useful for the purpose of this study.

The scrape was conducted in a temporal range that went from January to May 2020, in order to track eventual evolutions of trends. In addition, the process was also repeated for the indicated competitors of Cybertruck from March to May 2020, but with the relevant downsizes of a considerable reduction of numbers of tweets retrieved: in fact, despite the maximum number specified was the same even for the other models, way less texts were available. This factor limits the implementation of the research, as the small numbers for the other products are too low to be processed and analysed properly. However, some findings can be still retrieved and, in general, the difference in numbers already gave a first relevant demonstration of the higher popularity Tesla Cybertruck has compared to its competitors. To go deeper in detail, the following table sums up the results obtained:

Table 2: Number of tweets scraped for each product, monthly

Product	January	February	March	April	May	Total
Tesla Cybertruck	7114	2089	4773	6604	3183	23763
GMC Hummer EV	N.D.	N.D.	227	287	49	563
Rivian R1T	N.D.	N.D.	30	113	35	178
Ford Electric F-150	N.D.	N.D.	66	56	16	138

Considering this first outcome, it is straightforward to understand that the study would be mainly focused in the detection of Tesla Cybertruck's sentiment characteristics and word topics. Nevertheless, few comparisons with competitors' products in the electric pickup sector will be possible and implementable anyway.

The gathering is only the very first step of the entire process, and despite its formatting into a data frame, these data can be still considered as raw for the purpose of this study. That is why it is necessary to implement some pre-processing tools through other specific R codes and packages.

Text Mining Processing

Once the data are retrieved and stored, preliminary cleaning of text represents, as seen, one of the most important task to implement, especially in social media domain where the level of noise is considerably high due to the presence of emoticons, slangs and emojis that make the individuation of the single word more difficult. However, these elements can contain useful pieces of information, so in a social media domain they cannot be considered as data to be discarded: that is also why some specific lexicon were developed that are able to detect these special characters and assign a value to them.

For the purpose of this study, the cleaning and pre-processing phase was organised mainly into two sets: the first follows the classic functions with the aim of obtaining a token with only one-word-per-row structure, so by tokenizing the content of the tweet; the second method follows a different approach as its main goal is to

maintain the structure of the tweet, punctuation and symbols as they contribute to create meaningful emoticons or slangs, in order to deploy the specific abovementioned lexicon and refine the findings of sentiment analysis. Nevertheless, these two approaches share some common points in cleaning: naturally, they both need a previous conversion into a corpus, that can be defined as "*a collection of text sources*" (Devopedia, 2019) and creates the set of documents that are object of the analysis; then, steps like the conversion of text into lowercase, removal of https links and duplicates allow to get rid of non-useful elements for both methods. The package that allows to perform this cleaning phase is called "tm" (Feinerer & Hornik, 2019). This source allows to perform different activities related to text mining as it embraces a relevant portion of text document management supporting different text formats. Some of the steps implemented in this phase were:

1. Conversion to lowercase: R in sensible to the capital letters. Thus, for example, "work" and "Work" are recognised as different words. Thus, a lowercase conversion is useful to avoid redundancies.
2. Usernames, punctuation, symbols, URLs, special characters removal.
3. Elimination of English stop words, that bare little or no additional sense to the words that need to be analysed.
4. Specific words removal: some words, despite not being recognized as stop words, still do not bear relevant information to the analysis of sentiment. Thus, its removal simplifies computation and further processing.
5. Stemming: the words here are converted into their root to simplify and lighten the data frame as it groups words with same meaning but written differently.
6. White Space stripping: all these eliminations produce strings with null values or excessive spaces between words. This argument thus addresses the problem by bringing a correct word-space sequence back.
7. Duplication removal: this step is important in order not to consider the same tweets and words two times, thus avoiding redundancies that would alter the analysis, especially when it comes to frequency distribution.

The alternative approach follows the same structure except for the fact that it does not take into account punctuation and specific words removal, as they bear value for that type of analysis. The next steps are what really differentiate the approaches. In fact, in the former a term-document matrix is created, that is a grid that represent every document, tweets in this case, and the word that are in them, using the dichotomous variable of 0 if the word is not present in that document or 1 on the contrary and thus generating a really sparse matrix as the majority of the elements would be of value 0 (BSC, s.d.). In the latter approach, the tokenization and matrix are not necessary as the analysis is focused on the tweet on its entirety. Naturally, these approaches led to different findings that will be presented in the further section.

Findings

As already mentioned, the study conduces to diverse findings that vary basing on the method and time range chosen. In detail, the findings are structured in this way:

1. Findings of round 1: A broaden inspection of the dataset mainly based on general visualization tools. This round was repeated for each monthly dataset of Cybertruck and it is useful to have a first glance of the word present in tweets and thus individuate potential point of discussion. Moreover, still starting from the monthly datasets, a subsequent phase deepens the comprehension of the words, as it examines the relationships that can occur between them through the decomposition of tweets into the most frequent bigrams. Again, this can be a good way of inspecting the dataset as it goes beyond a single word inspection and thus captures more context. A month's lexicon-based sentiment analysis is then conducted.
2. Findings of round 2: This is the round in which an unsupervised learning method taken from NLP field is implemented: Latent Dirichlet Allocation or LDA. This one is used to clearly find the most relevant topics in the tweets and further deepens the level of the analysis. This phase is conducted for the entire dataset of Cybertruck tweets, without monthly division. After having completed it, a set of topics were given back: their content is then evaluated to manually extract topics that are judged to be more relevant. Finally, a singular sentiment analysis is conducted over the arguments chosen. This round allows to analyse the single features and drivers of the sentiment, thus going beyond document-level to inspect the score of precise characteristics that surround Cybertruck.
3. Findings of round 3: The final round that was entirely conducted on just Cybertruck dataset is the one that implemented the special dictionary that detects particular elements of Twitter's language, like emoticons, emoji and slangs, together with taking into account negators, amplifiers, question weights and adversatives. A numerical value range that went from -2.5 to 2.5 was then assigned to each tweet and graphs and other visualization tools are used to show the final score of the Cybertruck dataset. This phase is conducted by taking the data frame of Cybertruck in its entirety.
4. Findings round 4: The very last round is the one that compares Cybertruck to its abovementioned competitors by levering a special R package that allows to specifically conduct a competitive sentiment analysis of Twitter data. Again, the data frames were used with no monthly division.

Findings of round 1

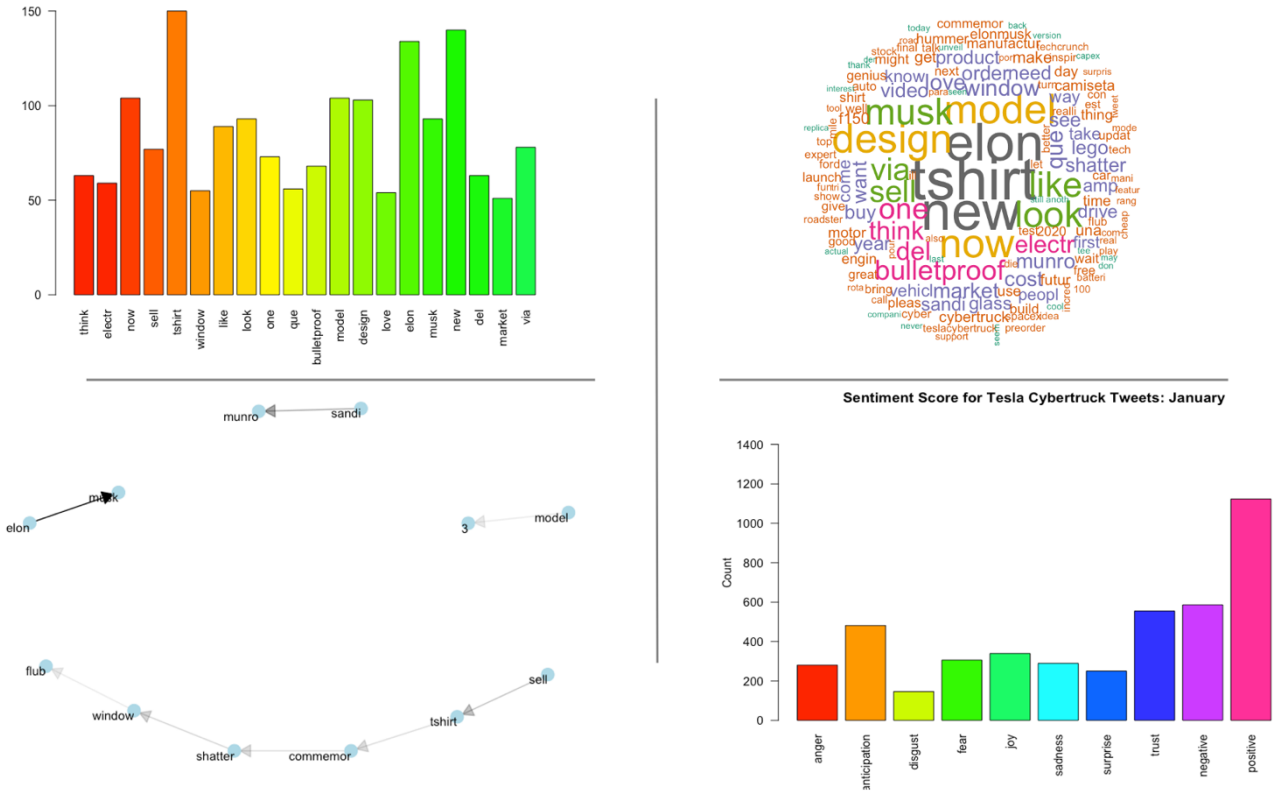
As already anticipated, the broadest level of the study will inspect general frequency distribution and visualization tools of the most relevant words of the tweets in a monthly cadence. This can be useful to first detect the topics under discussions in tweets, and meaningful information can already be retrieved.

By finishing the cleaning phase in R, the following step was to create the term document matrix to inspect the frequency of words in the documents-tweets, a data frame with the words and their frequency was then generated and a plot was created as well. In particular, there were retained the words that were cited more than

50 times or 100 if the bar plot resulted too crowded. In addition, another kind of image was generated from that dataset: this is the word cloud, a type of visualization that shows text data organized in a shape of a cloud, whose size varies based on words' recurrence on the dataset: the higher the frequency of the word, the bigger is it scape occupied in the cloud. In other words, *"a word cloud is a collection, or cluster, of words depicted in different sizes. The bigger and bolder the word appears, the more often it's mentioned within a given text and the more important it is"* (Boost Labs, 2014). The command for word cloud allowed to show words with a minimum frequency of 5 and a maximum number of showable words equal to 150. The other set of findings focused examined how the word are linked to each other: this was made possible after having tokenized the text and inspected the frequent bigrams, namely words that occur together the most. Specifically, only the bigrams with a frequency higher than 20 were retained. In some cases, the bigrams were even linked to each other, building a path that resembles a sentence structure. The final monthly analysis conducted was a sentiment analysis based on the lexicon called NRC Word-Emotion Association Lexicon, or EmoLex (Mohammad & Turney, 2010) (Mohammad & Turney, 2013), that is a dictionary composed by 14,182 unigrams divided in eight basic emotions, namely anger, fear, anticipation, trust, surprise, sadness, joy and disgust, and two sentiments, namely negative and positive. The classification of EmoLex was manually conducted through crowdsourcing in Mechanical Turk platform (Mohammad, s.d.). The dictionary is retrievable on R by using the package "syuzhet" (Jockers, 2017).

The following plots were the results of this first monthly analysis and were all retrieved from the R Studio code:

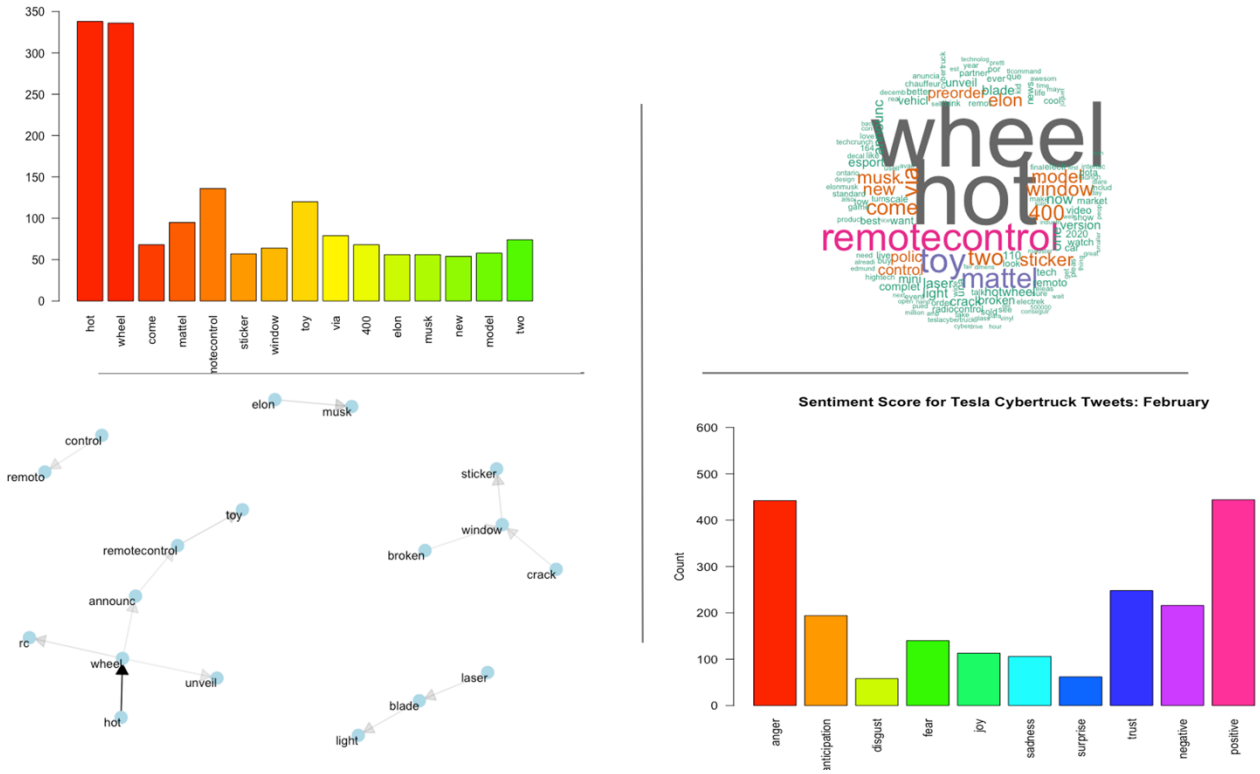
Figure 18: Cybertruck Word Frequency, Word cloud, Bigram Relationship and Sentiment Analysis, January 2020



Retrieved from *R Studio*

January word plots highlights some relevant topics: firstly, it is possible to still find referrals to the shatter of Cybertruck's window happened during the pickup's unveiling on November 21st, 2019. In particular, the words "window" and "bulletproof" are clear referrals, but also "tshirt" is one of them. In fact, as seen, after the episode of the broken glass Musk started the selling of t-shirt depicting the broken window. Moreover, an expected mention to the CEO Elon Musk himself are in place: this is a particular recurrent referral in each part of the study, as Musk is inextricably linked to its company and their product, Cybertruck included. In addition, in the word cloud it is possible to find even more details: a proof of that is the word "shatter". Other curious information retrieved in the cloud arose, for instance, from the word "lego", in the upper mid right section of the circle: in fact, after making some researches, it was found that a group of Tesla enthusiasts submitted a concept for a Lego Cybertruck to Lego Ideas, the site that collects new and original ideas for the creation and commercialization of new Lego pieces. This sort of request was sustained by about 10,000 people, pushing Lego to start the project review process, beginning on May 2020 (Paukert, 2020). From what concerns the bigram relationship, these confirm the information retrieved in the figures before, adding another frequent bigram: "sandi munro". This refers to the CEO of Munro & Associates Sandy Munro, whose company's core business is *"helping companies reduce "time to market", R&D, engineering and manufacturing costs all while increasing the quality of our customers products, processes and systems".* (Munro & Associates, Inc., s.d.). Munro provided a review of the Cybertruck after having accurately analysed it from an engineering perspective, even receiving a positive feedback from Elon Musk (Munro & Associates, Inc., 2020). Lastly, the sentiment analysis conducted through NRC showed a stronger presence of positive words compared to the negative ones. For what concerns the emotion analysis, trust and anticipation were the dominant ones, probably meaning that, since the attention for the Cybertruck was kept high, users are still impatient to know more about the product and have in a way forgiven the mistakes made during the unveiling.

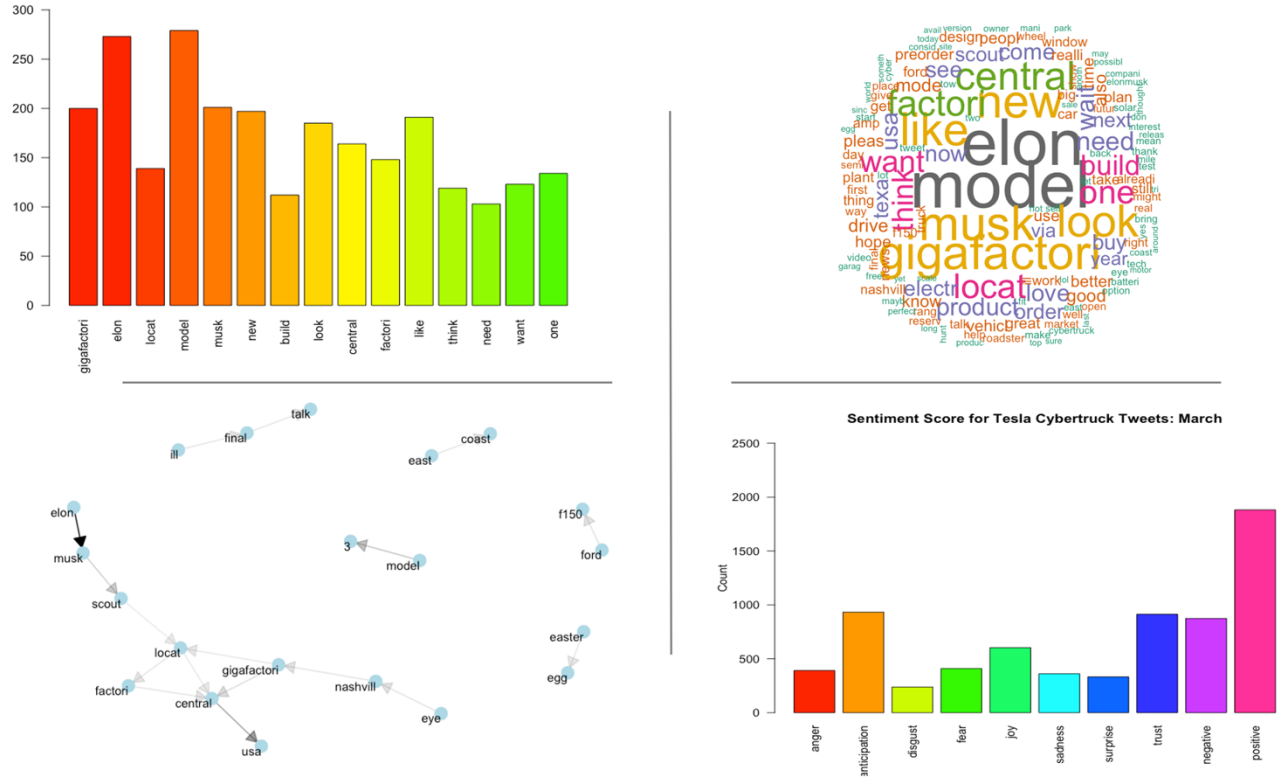
Figure 19: Cybertruck Word Frequency, Word cloud, Bigram Relationship and Sentiment Analysis, February 2020



Retrieved from *R Studio*

The analysis for February clearly highlights a word trend and most discussed topic: this is about the collaboration that Tesla made with Mattel's Hot Wheels for the realization and commercialization of a remote-controlled Tesla Cybertruck in both 1:64 scale, at the price of $20, and 1:10 scale, that would cost about $400. The ship is set to start on December 2020 in limited edition (O'Kane, 2020). The only adding that can be retrieved from the bigram relationship graph is the one that concerns the laser blade light feature on Cybertruck, announced, as usual, by Elon Musk in his Twitter account (Lambert, 2020). For what concerns the sentiment analysis, an unexpected outcome is shown in the high frequency for the "anger" emotion: however, this can be clearly justified by the fact that it is a misclassification of the dictionary, that recognizes the word "hot" as referring to the "anger" category, while, as abovementioned, it is just associated to the Mattel's Hot Wheels brand. This case clearly explains the reason why it is necessary to go beyond a mere word classification by inspecting how most relevant words are used and in which context. In this case, the combination of computational analysis and human observation was useful to detect this factor.

Figure 20: Cybertruck Word Frequency, Word cloud, Bigram Relationship and Sentiment Analysis, March 2020



Retrieved from *R Studio*

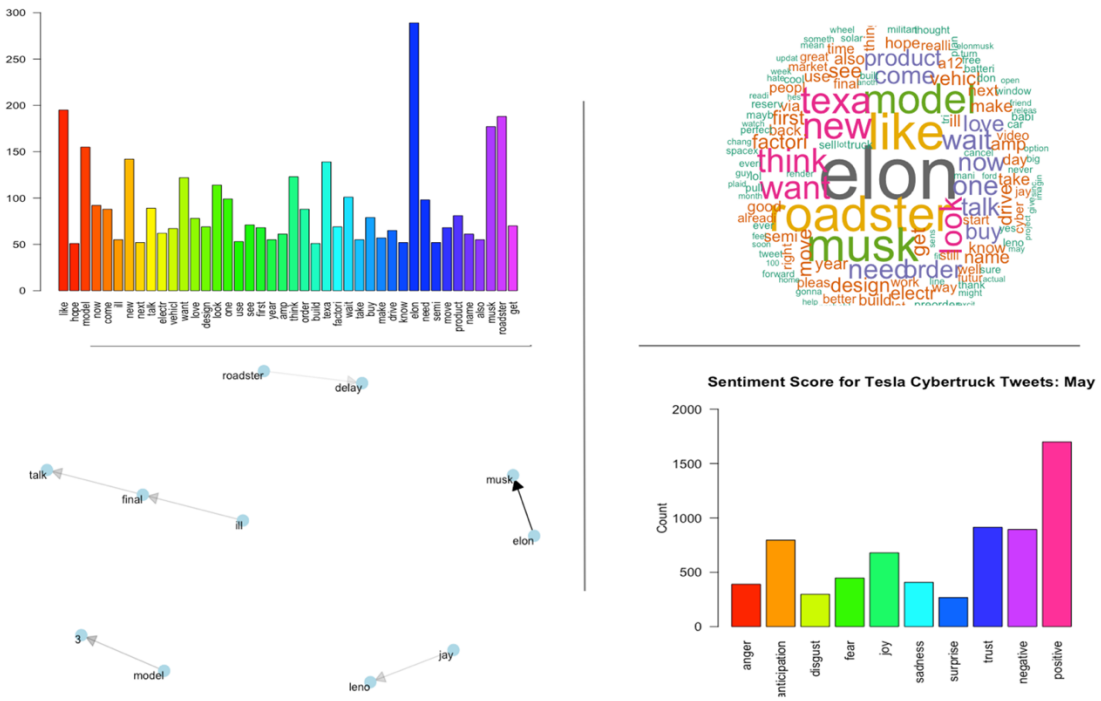
One of the main topics of the month the one that concerned the choice of the location for a new Gigafactory in central USA. That was the content of Musk's tweet in March, adding that the new factory will produce both Cybertruck and Model Y for the East USA market (Levy, 2020). Not surprisingly, the news created a buzz in the Internet, with a real positive response by the users as shown in the NRC analysis. In addition to having positive attitude towards Cybertruck tweets in March, they also showed a high level of anticipation, probably linked to the suspense derived from the decision of the location for the new Gigafactory.

Figure 21: Cybertruck Word Frequency, Word cloud, Bigram Relationship and Sentiment Analysis, April 2020

Retrieved from *R Studio*

For what concerns April, no big changes needed to be reported compared to what happened in the previous month: the sentiment analysis trend, with positive reactions and great anticipation arguably due to the Gigafactory topic, was confirmed, with only changes in frequency and thus proportions. News can be retrieved from the bigram graph: the fact that Joplin, a small city in Missouri, offered Tesla $1 billion in tax incentives for constructing the Gigafactory there (Rapier, 2020). In addition, in April, the new season of "Jay Leno's Garage", a known American web and television series about motors and even new model's test drives, was announced. This show will include a special and really awaited episode on Tesla Cybertruck (Lambert, 2020).

Figure 22: Cybertruck Word Frequency, Word cloud, Bigram Relationship and Sentiment Analysis, May 2020



Retrieved from *R Studio*

May did not produced that much additional information: sentiment trend is still the same, with positives overtaking negatives; some of the main topics are still linked to Jay Leno's Garage and Gigafactory. For the latter, the May news can be the one that, according to reliable sources, Tesla had decided to set the Gigafactory location in Austin, Texas (Lambert, 2020). Finally, a relevant news is the announcement of the delay for the new Tesla Roadster, that will be produced in 2022 and after Cybertruck (Szymkowski, 2020). This news can explain the lowering in the emotion of anticipation that, despite still higher in frequency than the others, is considerably lower than the last month.

The month of May concludes the monthly analysis for the tweets regarding Tesla Cybertruck. First findings however show promising results for the model, but a deeper analysis is necessary to understand more about the content of the tweets taken into account.

Findings of round 2

This set of findings is focused on extracting the most relevant topics from tweets by using the unsupervised learning technique of LDA. As already said in previous sections, Latent Dirichlet Allocation is a generative probabilistic model that can allow to assign a topic to a specific word, knowing the distribution of topics in the documents. The generative element refers to the fact that the new samples are generated from the same distribution and ultimately from the same data frame. (Blei, et al., 2003) (Manwani, 2018). Furthermore, a prerequisite of LDA is that each document has to share the same topic, whose number is pre-specified, but distribution differs. In this way, the process ultimately assigns a certain score to the words and allocates them into the different sets of topics created. That is a useful way to group similar texts and identify relevant groups of words that can be analysed from a sentiment point of view afterwards. This is how LDA is inserted in this

study. It was essentially used as a first step to extract topics and words that compose them, with the following step being their analysis of sentiment. In other words, this phase merges LDA and sentiment analysis to understand the impact of specific arguments of discussion in the sentiment of tweets, to produce a deeper finding for the entire study.

From a computational point of view, the methodology used for this phase is the same as the first round, but the Cybertruck tweets were unified, thus there is no month distinction. So, the cleaning phase and the creation of a document-term matrix are steps that were repeated here. The differences started when the package "topicmodels" was implemented (Grün, et al., 2020), that allows to compute probabilistic models based on the data structure obtained with the package "tm". However, an additional point needs to be addressed: in fact, here a precise method for LDA is specified, which is Gibbs sampling. As seen, Latent Dirichlet Allocation allows to assign a distribution of topic in each document and a distribution of words in topics: what Gibbs does is to try optimizing the conditional distribution of these variables by repeating the classification over time. Since the allocation to the exact topic is a probability matter, there is a higher chance for the word to be classified correctly, but no certainty. Gibbs maximizes the likelihood of a good classification by levering on the conditional probability distribution of a word's topic assignment that is conditioned by the rest of the topic assignments. In other words, to get the topic k of a word w that belongs to a document d, LDA with Gibbs method analyses, among other factors, the number of times that document d used topic k, $n(d,k)$, and the number of times topic k used the word w, $n(k,w)$. A Dirichlet parameter allows to still derive an assignment if these two values are equal to 0. Finally, as mentioned, to better perform this probability, a higher number of trials contribute to more precise results (Tomar, 2018) (Grün & Hornik, 2011). The following codes were the ones used for this study and were retrieved from (Hossain, 2018):

```
dtm = DocumentTermMatrix(cleanset1)
dtm

dim(dtm)

doc.length = apply(dtm, 1, sum)
dtm = dtm[doc.length > 0,]
dtm

inspect(dtm[1:2,10:15])

freq = colSums(as.matrix(dtm))
length(freq)

ord = order(freq, decreasing = TRUE)
freq[head(ord, n = 20)]

# Creating a subset of words having more than 200 frequency

plot = data.frame(words = names(freq), count = freq)

plot = subset(plot, plot$count > 200)
```

```
str(plot)

ggplot(data = plot, aes(words, count)) +
  geom_bar(stat = 'identity') +
  ggtitle('Words used more than 200 times')+
  coord_flip()

# HERE IT COMES LDA

library(topicmodels)

# LDA model with 5 topics selected

lda_5 = LDA(dtm, k = 5, method = 'Gibbs',
            control = list(nstart = 5, seed = list(1505,99,36,56,88), best = TRUE,
                           thin = 500, burnin = 4000, iter = 2000))

# LDA model with 2 topics selected

lda_2 = LDA(dtm, k = 2, method = 'Gibbs',
            control = list(nstart = 5, seed = list(1505,99,36,56,88), best = TRUE,
                           thin = 500, burnin = 4000, iter = 2000))

# LDA model with 10 topics selected

lda_10 = LDA(dtm, k = 10, method = 'Gibbs',
             control = list(nstart = 5, seed = list(1505,99,36,56,88), best = TRUE,
                            thin = 500, burnin = 4000, iter = 2000))

# Top 10 terms or words under each topic

top10terms_5 = as.matrix(terms(lda_5,10))
top10terms_2 = as.matrix(terms(lda_2,10))
top10terms_10 = as.matrix(terms(lda_10,10))

top10terms_5

top10terms_2

top10terms_10

lda.topics_5 = as.matrix(topics(lda_5))
lda.topics_2 = as.matrix(topics(lda_2))
lda.topics_10 = as.matrix(topics(lda_10))
```

What was done here is a previous cleaning that led to "cleanset1". Then, this was converted into a document-text matrix. Inspections were conducted to have a first glance of the dataset. After that, a top 20 frequency words plot was created, followed by an identification of words that are used more than 200 times. Since this is not the main focus of this phase, that plot was not reported. After these preliminary steps, LDA with "topicmodels" took place. The *k* factor was the number of topics that had to be pre-set, knowing that selecting an higher number of *k* leads to a more fine-grained division, but the differences may get blurred, while choosing little number allows to have a clear idea of the differences but it would lead to a loss of information in the form of topics. To tackle this optimization problem, the code chose to apply three different *k* values. As

showed, Gibbs was the method chosen, where "burnin" are the number of first iteration to discard, "thin" represents the number of iterations omitted during the processing and "iter" are the number of iterations to perform. In particular, the "thin" argument is used to prevent sample correlation during the iteration (Liu, 2015). Finally, seeds are a set of casual numbers that allows to fix the results arising from probability samplings and distributions, that otherwise could differ.

Nevertheless, that all still does not give back the results as it still does not give back real implementable knowledge: that is why the top 10 terms were filtered from the samples in order to retain only valuable information. The rest of the code serves as an explanation of the probability distribution implemented through Gibbs sampling. What matters for the continuation of the analysis was the result of the top 10 terms obtained after computing "LDA" function: these are shown below:

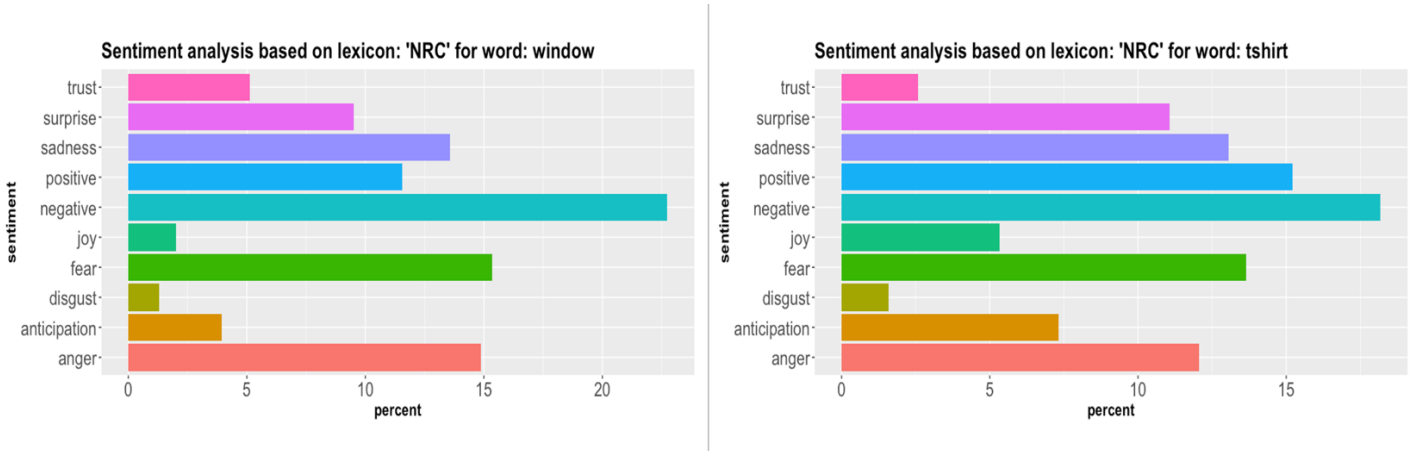
```
> top10terms_5
      Topic 1    Topic 2       Topic 3  Topic 4  Topic 5
 [1,] "model"    "need"        "like"   "elon"   "new"
 [2,] "now"      "factori"     "look"   "musk"   "come"
 [3,] "wheel"    "gigafactori" "think"  "design" "wait"
 [4,] "hot"      "product"     "want"   "via"    "love"
 [5,] "roadster" "texa"        "one"    "electr" "order"
 [6,] "window"   "build"       "vehicl" "day"    "buy"
 [7,] "preorder" "pleas"       "drive"  "futur"  "year"
 [8,] "sell"     "great"       "know"   "news"   "use"
 [9,] "market"   "locat"       "peopl"  "chang"  "also"
[10,] "semi"     "central"     "thing"  "free"   "next"

> top10terms_2
      Topic 1       Topic 2
 [1,] "model"       "elon"
 [2,] "now"         "musk"
 [3,] "wheel"       "like"
 [4,] "need"        "new"
 [5,] "hot"         "look"
 [6,] "via"         "come"
 [7,] "factori"     "think"
 [8,] "gigafactori" "one"
 [9,] "roadster"    "want"
[10,] "product"     "design"

> top10terms_10
      Topic 1  Topic 2  Topic 3    Topic 4 Topic 5       Topic 6      Topic 7    Topic 8
 [1,] "one"    "think"  "model"    "new"   "elon"        "like"       "electr"   "wheel"
 [2,] "want"   "order"  "now"      "come"  "musk"        "look"       "better"   "hot"
 [3,] "vehicl" "wait"   "roadster" "love"  "factori"     "design"     "ford"     "via"
 [4,] "get"    "buy"    "sell"     "year"  "gigafactori" "amp"        "news"     "window"
 [5,] "thing"  "time"   "car"      "next"  "build"       "cybertruck" "ever"     "preorder"
 [6,] "realli" "peopl"  "way"      "take"  "product"     "chang"      "show"     "two"
 [7,] "first"  "good"   "mode"     "use"   "texa"        "garag"      "plant"    "toy"
 [8,] "real"   "still"  "tshirt"   "talk"  "locat"       "might"      "watch"    "remotecontrol"
 [9,] "sure"   "reserv" "right"    "video" "central"     "featur"     "missouri" "version"
[10,] "test"   "well"   "semi"     "hope"  "plan"        "cool"       "best"     "unveil"

      Topic 9   Topic 10
```

```
 [1,] "drive"    "need"
 [2,] "day"      "pleas"
 [3,] "batteri"  "also"
 [4,] "solar"    "know"
 [5,] "option"   "great"
 [6,] "rang"     "work"
 [7,] "free"     "back"
 [8,] "motor"    "thank"
 [9,] "mile"     "let"
[10,] "long"     "bring"
```

The topics generated went through some news that were shown in the previous section, but now there is a chance to select a word that represents the topic, extract the tweets that contain it and conduct a sentiment analysis on it. That is the continuation of the analysis, as subsets of some of the tweets containing these words were created and a sentiment summary was plotted for each of them. The lexicon chosen was still NRC, but this time the frequency outcomes were presented in a percentage form.

The words were selected manually by looking at which of them are of most valuable interest. For instance, it can be useful to understand the sentiment that is linked to the word "Elon", to assess the impact that the CEO had on Twitter users in the time range considered. Especially, LDA with five topics was inspected. The words "tshirt", the only one not present in *k*=5, "window", "gigafactori", "electr" that is the stem form for "electric", "design", "elon" and "order" were closely analysed. The results are shown below:

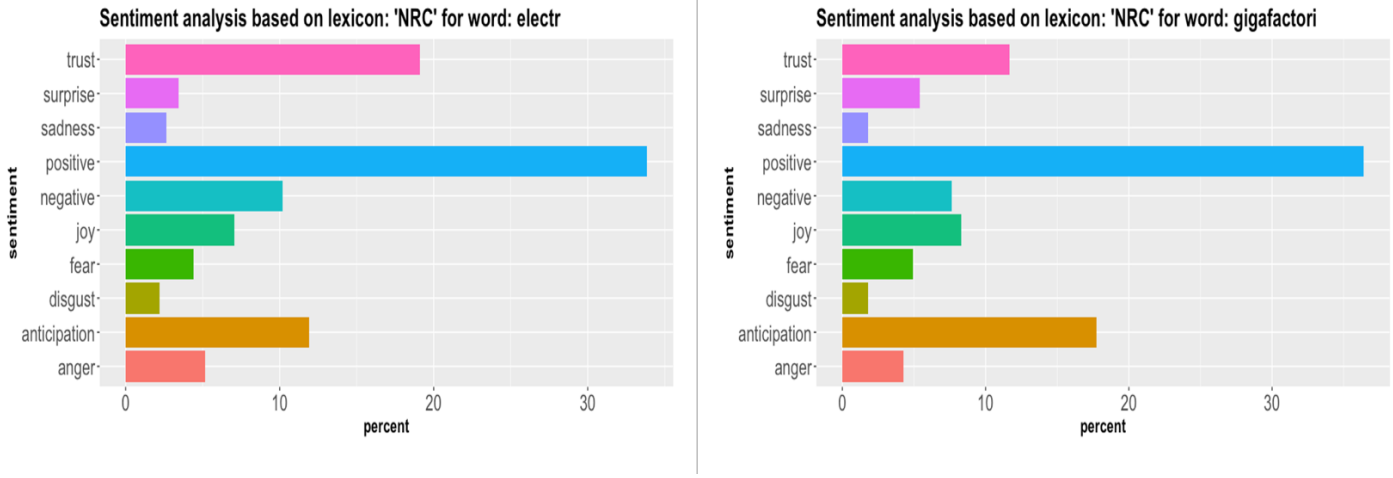
Figure 23: NRC sentiment analysis for "tshirt" and "window"



Retrieved from *R Studio*

The first two sentiment analysis were conducted on two of the set of tweets that were identified by LDA and refer to the episode of the unveiling. As shown, the predominant sentiment association is negative, with emotions like "anger", "fear" and "sadness" that represents the highest percentages. However, when looking at the "tshirt" sample, it is possible to notice that the value for "positive" and "surprise" are higher, probably meaning that the move made by Elon Musk to lever on emotional intelligence and not hiding the mistake resulted to be a successful one. Despite that, overall, the results showed a tendency to bad associations, so from a brand perception and reputation point of view, Tesla paid its mistake.

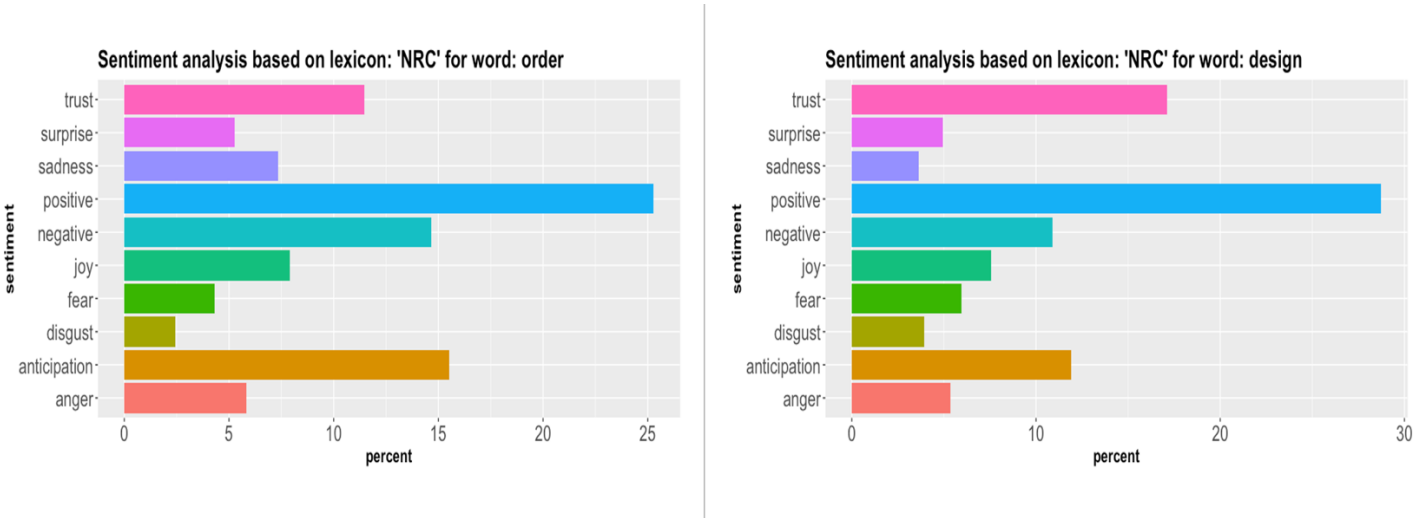
Figure 24: NRC sentiment analysis for "electr" and "gigafactori"



Retrieved from *R Studio*

The two plots above told a different story, as the values for two of the main elements of not only Cybertruck, but Tesla in general, were positively perceived by Twitter users. Some remarkable outcomes arose from the trust perceived for the electric automotive sector, that represents the highest emotion for the set, and the positive sentiment reaction overall. It was already mentioned the increased importance that electric vehicles are having in nowadays society, so this outcome represents a confirmation in terms of perception of this trend. For what concerns the Gigafactory set, high values in "anticipation" and "positive" can be explained by the great communication made by Elon Musk around the decision of building a new factory, that made Tesla fans more enthusiast overall.

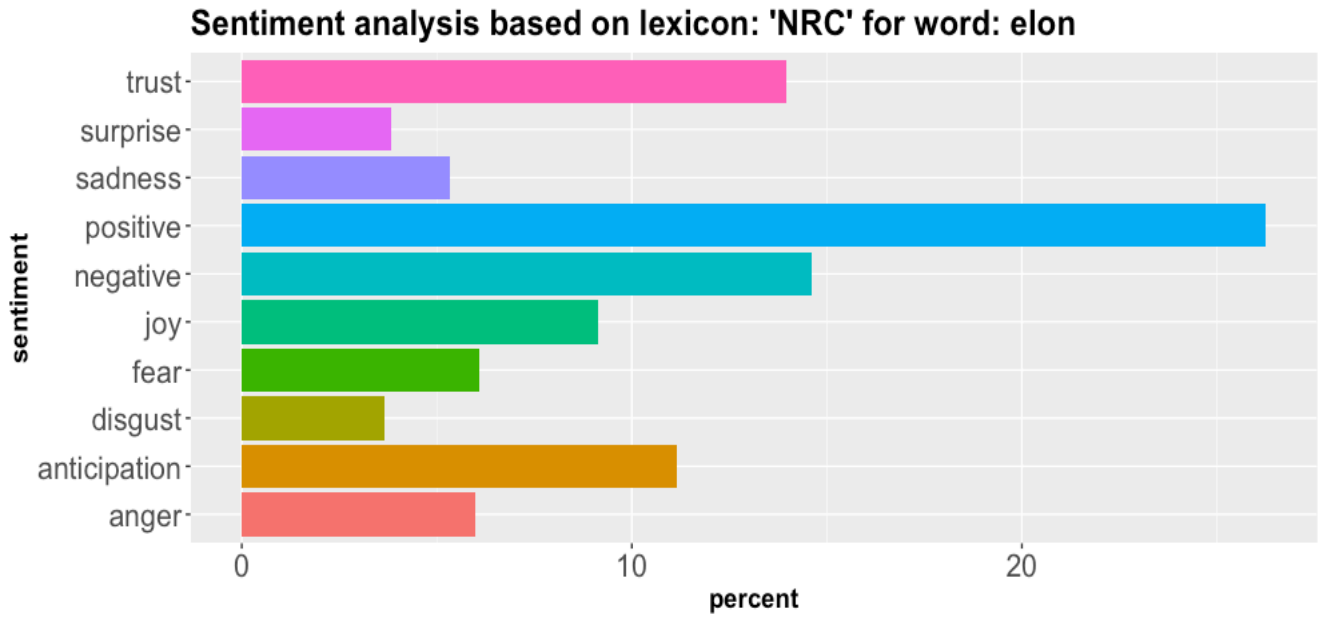
Figure 25: NRC sentiment analysis for "order" and "design"



Retrieved from *R Studio*

These two plots are probably the ones that are linked the most to Cybertruck as a product, in particular the set of "design". The images show a similar trend, with "positive" representing the highest sentiment and "trust" and "anticipation" as highest emotion values in percentage. The order set presents a slightly higher bad values: that can be linked to the uncertainty that sometimes surround Tesla production, with Cybertruck not making an exception. However, the good perceptions are higher with a consistent differential.

Figure 26: NRC sentiment analysis for "elon"



Retrieved from *R Studio*

Last but not least, the plot above shows the perception of tweets containing Elon Musk. The CEO confirms to be a relevant figure from a perception side too, as positive sentiment was higher than negative, and "trust" was the leading emotion. Despite the critiques that Musk had received for its controversial behaviours, he still remains the main person man that drives Tesla, Inc. and his approach and preparation had made him to be perceived as one of the most influential person not only in the automotive industry, but in the business worldwide.

That was the analysis of tweets subset identified by LDA. Moving on to the next round, the most relevant changes concerns the different lexicon chosen for the sentiment analysis.

Findings of round 3

The third set of findings for this study are focused in the application of specific functions retrieved from R package "sentimentr" (Rinker, 2019). The application of this package was made with the intention of deepening the results of this sentiment analysis study. In fact, the implementation of "sentimentr" allowed to include, in the calculation of the sentiment score, the context surrounding a polarized word, like negators, amplifiers or de-amplifiers, adversatives, emojis, emoticons and slang language: in other words, the package allows to compute a sentiment analysis at a sentence-level. As seen above, these are some element that are

very commonly used in a social media domain like Twitter. Thus, including them can definitely give back a more precise outcome of Cybertruck's tweets sentiment. Before going into details of the code used for this analysis, some premises about this round need to be formulated. Firstly, for the specific code used, the package allowed to choose among different lexicons, with the majority of them being versatile for a lot of domains. The following table resumes them:

Table 3: Retrievable dictionaries in the implemented function of "sentimentr"

Dictionary	Number of classified words & sentences	Score Range
senticnet	23626	-0.98; 0.982
sentiword	20093	-1; 1
jockers_rinker	11710	-1; 1
jockers	10738	-1; 1
huliu	6874	-2; 1
nrc	5468	-1; 1
socal_google	3290	-30.160085; 30.738912
loughran_mcdonald	2702	-1; 1

The lexicon showed have different characteristics both in their structure and creation: "senticnet" is a dictionary for concept-level sentiment analysis covering multiple discipline like statistics and semantics and used in different domains and tasks, social media marketing included. The "sentiword" opens up one of the biggest collections of classified words and synset and is largely used in literature for sentiment analysis tasks. The "jockers_rinker" and "jockers" retrieve the "syuzhet" dictionary, created from a corpus of contemporary novels. The lexicon "huliu" opens up the Bing dictionary, created by implementing a bootstrapping approach from WordNet and a set of product reviews. The "nrc" specific recalls the dictionary used for the analysis of the first two rounds: it was created through crowdsourcing and, in here, the results are only shown in a -1;1 range. Moreover, "socal_google" is a semi-supervised lexicon that was created by using a small set of words through a mixed approach of Pointwise Mutual Information and search engine queries. Lastly, "loughran_mcdonald" is a dictionary that is mainly targeted for finance domains, that is why it is also presented as "finance-specific". (Barnes, et al., 2019) (Naldi, 2019) (SenticNet, s.d.) (FBK-ICT, s.d.) (R Package Documentation, 2019). For this study, due to its size and fit for social media domains, the "senticnet" dictionary was used.

Moreover, since this analysis took into account the context of the tweet, the cleaning phase differed from the ones applied before. In fact, it just featured the lower-case conversion, usernames, stop words, links, duplication removal and white spaces stripping. So, the punctuation was retained, as necessary for the system to read and recognize the emoticons.

What the function performed with emoticon and emoji was a conversion of them into readable words: this means that there is a pre-defined dictionary that associates a word with a specific combination of punctuations. For instance, here is an example of how this function works:

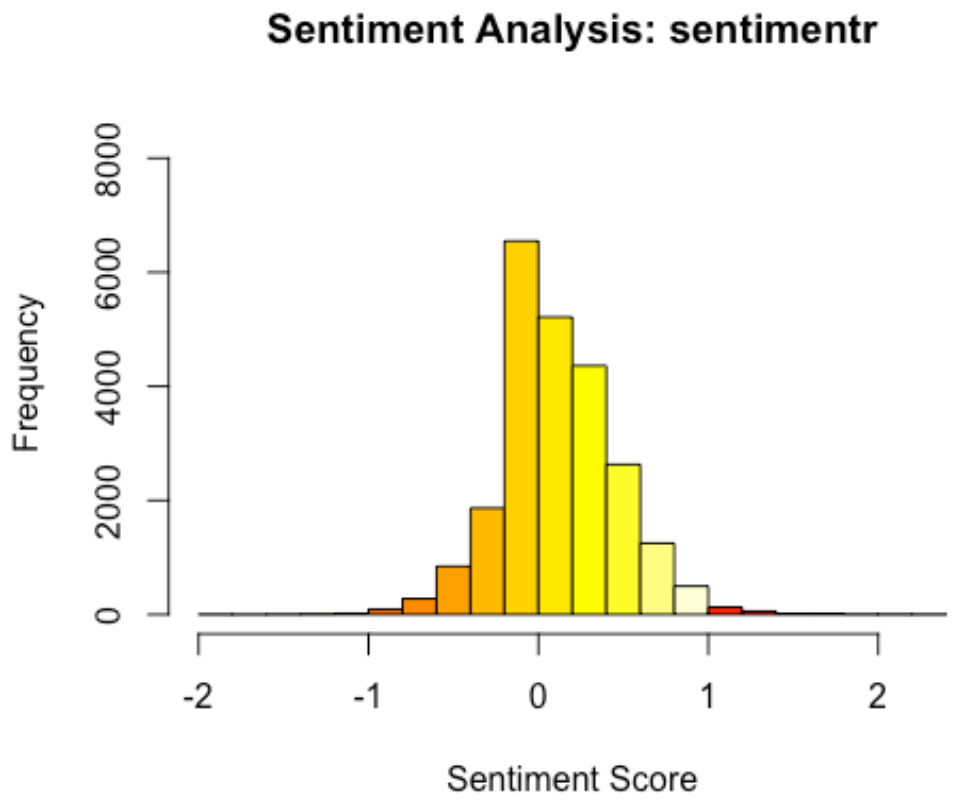
```
> replace_emoticon(":D")
[1] " laughing "
```

That is a representation of the fact that, combining the special lexicon with the dictionary adopted, more elements can be taken into account for the sentiment analysis. Nevertheless, little modification needed to be made. In fact, the correspondent word for ":)", arguably one of the most used emoticons in text messages, is "smiley". However, when inspecting the dictionaries, it was noticed that the word was not classified, but the word "smile" was present instead. This required an extra-step to be implemented. In fact, the word "smiley" was substituted with "smile", in order to tackle the problem. This is an ulterior evidence of the fact that a verification of the method implemented can be of relevant importance.
The core part of the code that gave back the outcome of interest was the one presented below (Rinker, 2019):

```
Cybertruck_Sentiment <-sentiment(Cybertruck_Clean_Sentences,
                                 polarity_dt = lexicon::hash_sentiment_senticnet,
                                 emoticon_dt = lexicon::hash_emoticons,
                                 valence_shifters_dt = lexicon::hash_valence_shifters,
                                 slang = lexicon::hash_sentiment_slangsd,
                                 emoji_dt= lexicon::hash_sentiment_emojis,
                                 hyphen = "",
                                 amplifier.weight = 0.8, n.before = 5, n.after = 2,
                                 question.weight = 1, adversative.weight = 0.25,
                                 neutral.nonverb.like = FALSE, missing_value = 0)
```

The code highlights the mix of lexicons used for calculating the score, together with clarify how the tweet context is taken into account. As said, "senticnet" was the dictionary chosen and was reported in the "polarity-dt" argument. Following that, the emoticon, valence shifters, slang and emoji dictionaries are activated and substituted as showed above. The argument "hyphen" just substitutes the hyphen with what indicated in the equation, in this case a blank space. The weights that follow are calculated by multiplying the polarized terms by 1 + this weight values, which were used at their default mode and so were the number of words to consider as surrounding the term: five before and two after it. This means that the system evaluates a cluster composed by 5 + polarized term + 2 number of words. A special mention needs to be addressed to the adversative words: if they are located before the polarized term in the cluster, the latter is increased in value by $(1 + N_{adversative\ conjunction} \times z_2)$, where z_2 is the weight of the adversative word. Conversely, if the word is found after the polarized term, it decreases the value of the cluster by $(1 + N_{adversative\ conjunction} \times -1 \times z_2)$.
The "neutral.nonverb.like" argument just strips the word "like" that can be usually found after some linking verbs: since this is rarely polarized, it is set as false, meaning not to consider that in the computation. Finally, "missing values" replaced 0 value to tweets where no terms were found. This means that a tweet whose value is 0 has no polarizing potential, so it was considered as neutral. Since it was likely that the majority of them can result as being not polarized, it was expected neutrals tweets to be higher in frequency value.
This function was applied only to Tesla Cybertruck data frame with no differences in months. The following is the histogram that allows to visualize the outcomes:

Figure 27: Cybertruck sentiment analysis results using the package "sentimentr" in numbers



Retrieved from R Studio

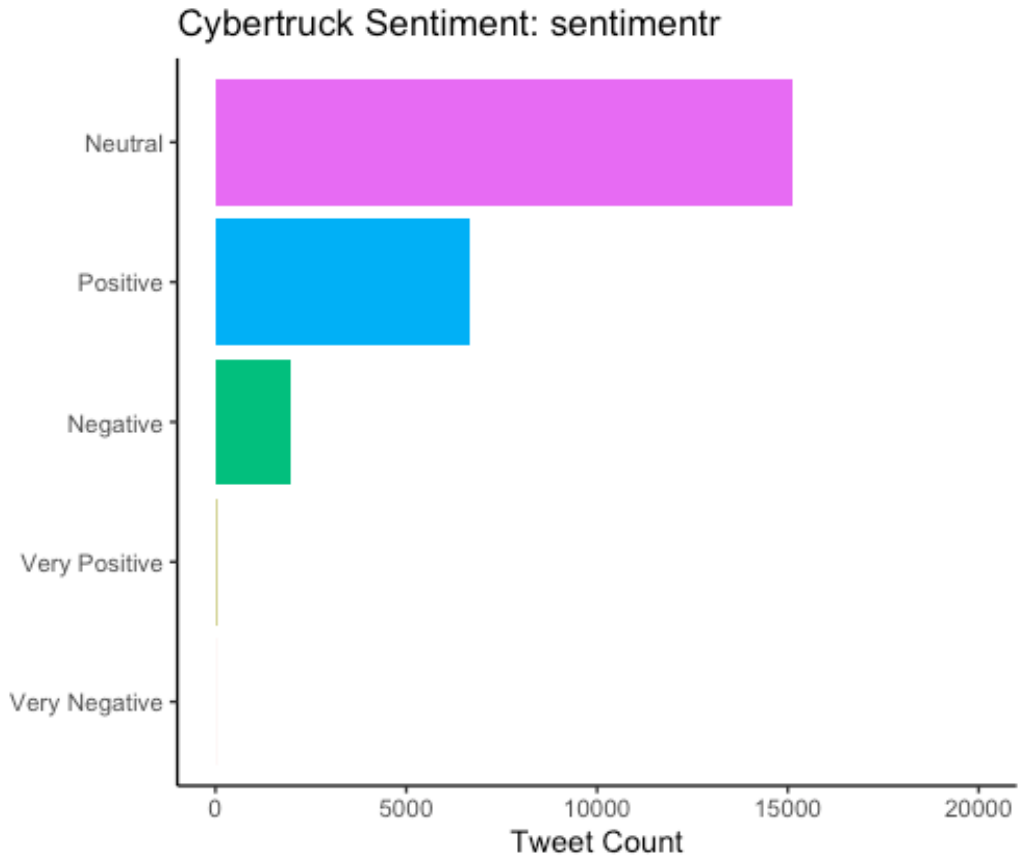
As expected, a lot of the tweets were close to the 0 value, meaning that neutrality is predominant. However, moving away from the central zero point, it is possible to notice that the bar plots are taller in the positive side, namely the one that is composed by values higher than 0. This can give an idea of the fact that the perception and sentiment associated to Cybertruck tweets is more positive than negative, even taking all the above-mentioned factors into account.

To have a clearer vision of that, another visualization tool was implemented by manually convert numeric values into ranges that went from "very negative" to "very positive" in the following order:

1. -2.5 to -1.51: "very negative"
2. -1.50 to -0.31: "negative"
3. -0.30 to 0.3: "neutral"
4. 0.31 to 1.5: "positive"
5. 1.51 to 2.5: "very positive"

The following is the graph obtained:

Figure 28: Cybertruck sentiment analysis results using the package "sentimentr" in value ranges



Retrieved from R Studio

This division was made for the sake of a better visualization of the results already noticed from the previous plot. In addition, the ranges were enlarged due to the effect of the above-mentioned calculations made by the system. It is now clear that the number of positive tweets were higher than the negatives, despite neutrals having the highest frequency. These results confirmed what already found before, that is the feedbacks for Cybertruck and Tesla in general are more positive than negative. This is due to several factors like the uniqueness of the product, the organization and mission of Tesla and the excitement arising from the CEO Elon Musk and his initiatives like the constructions of the huge Gigafactory. This model recalls to more than a pickup vehicle, making it a product that leaves space to imagination and wondering.
A final sentiment study for Tesla Cybertruck is the one that compares it to some of its main competitor in the sample's time range and it is presented in the following round.

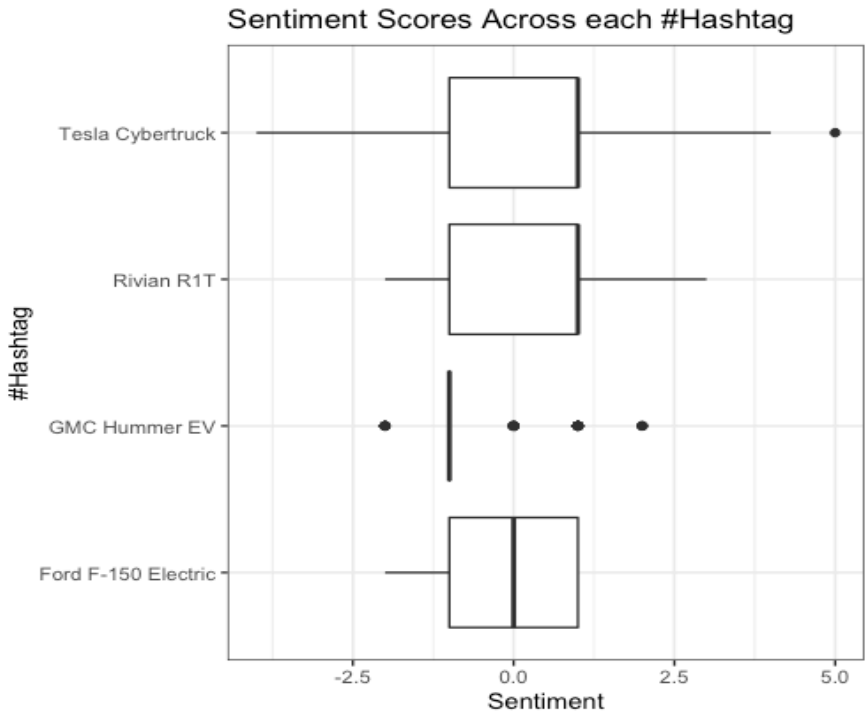
Findings of round 4

The final findings took into account Tesla Cybertruck with some of its already mentioned competitors, namely Rivian R1T, Ford F-150 Electric and GMC Hummer EV. To process this phase of the analysis, another special R package was used and it is called "saotd" (Munson, et al., 2019), an acronym that stands for Sentiment Analysis of Twitter Data. Similar to other packages, "saotd" allows to compute a complete sentiment analysis, from the acquisition of data to visualization tools, by using the already known Bing dictionary. However, one of its main application is the comparison of tweets with different "hashtags", that in this case represent

different brands. In fact, to work properly, a creation of a dedicated column called "hashtag" is necessary to identify the subject of each tweets. That is why, in a new data frame appositely created for this study, tweets of different electric pickups were divided by their brand in the column "hashtag". Despite the gap in the collection of data, a meaningful over time evaluation of the samples was implemented and gave back the evolution of the sentiment throughout the months, levering specific visualizations tools like box and violin plots. In addition, other tools like the list of the positive and negative words, together with bigram and trigram correlation were analysed: however, the scope of this round was to compute a competitive analysis, so the focus was on visualization that allowed the comparison of brands. Moreover, the time range taken into account referred to the trimester of March-May 2020.

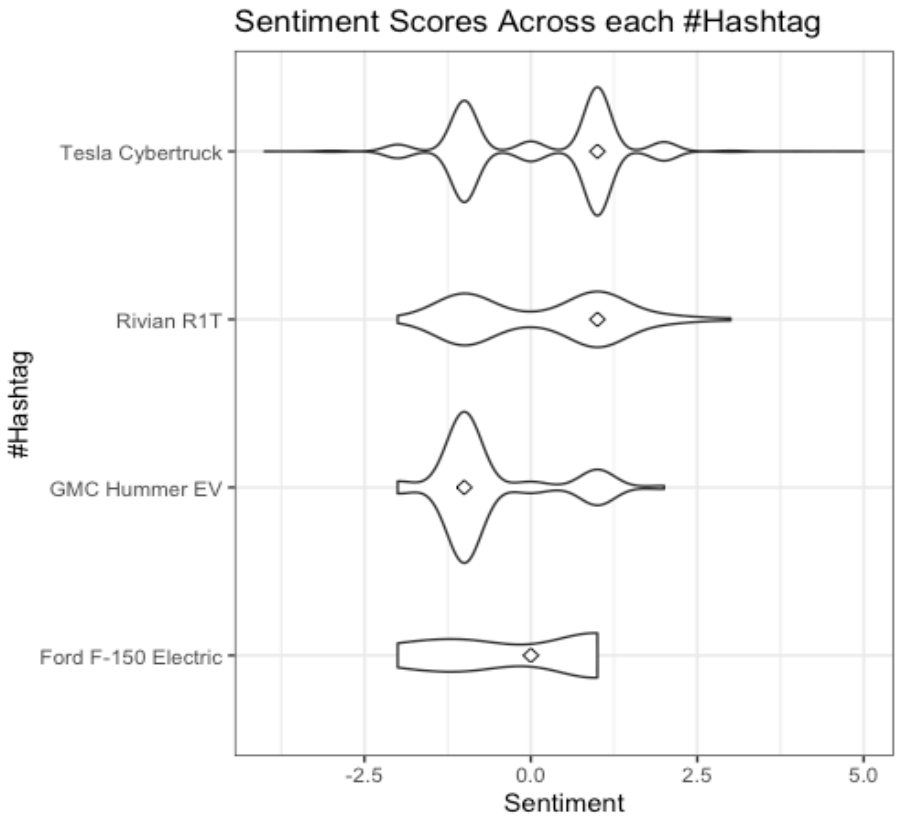
After having cleaned the data as usual, the following outcomes were retrieved (CRAN, 2019):

Figure 29: Box plot of sentiment score for different EVs brands using "saotd"



Retrieved from R Studio

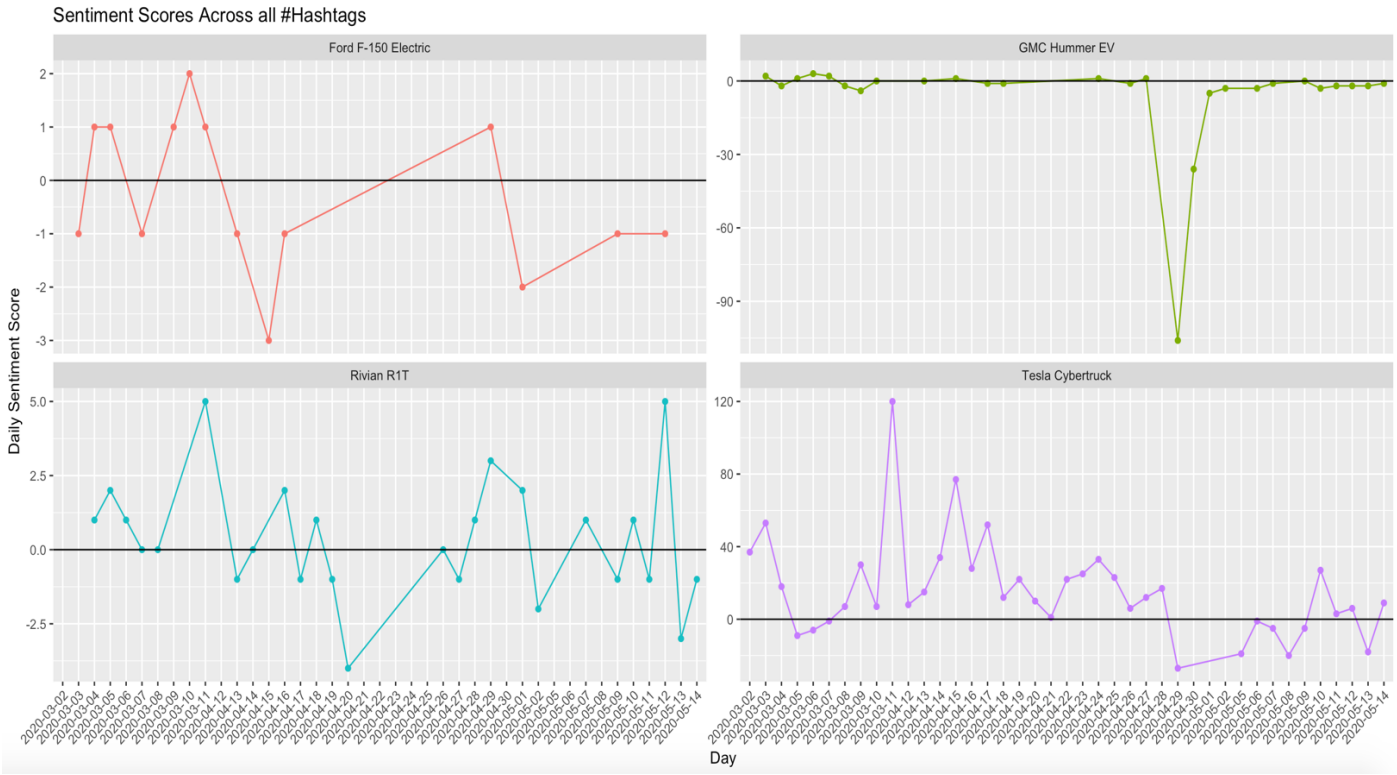
Figure 30: Violin plot of sentiment score for different EVs brands using "saotd"



Retrieved from R Studio

A first analysis of the data frame allowed to understand the general trend for each brand. For instance, the more positive tendency for Cybertruck was confirmed. Rivian R1T followed that path, with a mean that is similar to the Tesla Model. F-150 seemed to be generally neutral, thus the retrieved tweets were not polarized. Conversely, GMC Hummer EV shows the lowest average, in addition to the fact that an exact box plot could not be retrieved. The trends were confirmed by the following figure, that shows the violin plots for each brand. Having understood the differences across the models, an exact evaluation of the sentiment evolution allowed to inspect why the scores were different and to what extent. To find this out, the following plot showed the daily evolution of the samples' sentiment scores. The graph showed a sum of the lexicon sentiment results; thus, it did not take into account a sentiment range, but rather a daily sum of the scores:

Figure 31: Overtime sentiment score across each EVs brand using "saotd"



Retrieved from R Studio

This graph gave back meaningful insights even if the tweets collected for each brand differed in number. It showed the days of tweets samples in year-month-day order on the axis and the daily sentiment score sum in the ordinates. As expected, the trends showed value that were close to zero in most of the cases, except from the Cybertruck that could also count on higher sentiment sums. Moreover, going into the breakdown of the results across each electric pickup model, it is possible to point out that Rivian R1T had more positive peaks than negative, thus confirming what was seen before. Nevertheless, the results were somehow not volatile, with a sum range going from about -3 to 5 in score. The same applied for Ford, which had the smallest set of tweets and sentiment range, with the latter that went from -3 to 2. A narrower score range for this study, not taking into account the size of the dataset, can either mean that no big rumours or news were made about the car model or that the reaction towards were almost neutral overall. That can be particularly true for the Ford electric pickup, as not many information about the model are already in place. For what concerns Rivian R1T, the data changed more in the time range taken into account but still not with meaningful peaks and lows. What captured the attention the most were the scores for the models in the right side of the graph, namely GMC Hummer EV and Tesla Cybertruck. While the peaks of Tesla's electric pickups on March 11th 2020 and April 15th 2020 were already identified, since the former date is the one in which Elon Musk tweeted about scouting a new location in Central USA for the new Gigafactory and the latter coincides with the $1 billion offer made by Missouri for building the factory on their state (Crum, 2020) (HDMotori.it, 2020), the clear low for GMC Hummer EV registered on April 29th 2020 still needed to be inspected. On that date, in fact, GMC officially stated that the 20th May GMC Hummer EV's reveal date was postponed to an undetermined later date due to the Coronavirus pandemic (GMC Pressroom, 2020) (Capparella, 2020). That was also the reason why the

reveal of Rivian R1T was postponed too, thus eventually justifying the drop in April (Raia, 2020). Apart from that news, GMC maintained a pretty neutral score, so this can mean that the perception of product quality was not the reason why the GMC electric pickup had scored slightly more negative compared to the competitors overall.

Deriving this ulterior graph allowed to give a better explanation behind the sentiment scores of the pickups, thus understanding the reason behind them can gave more complete and precise insights about the companies and their products.

Discussion and Managerial Implications

The insights and findings of this study surrounding the emerging market of the electric pickups can be useful for not only understand its evolution so far, but also to try predicting future trends and providing useful implications that eventually drove the success of some brands compared to others. Nowadays there are too many user-generated data about companies to be ignored. Not levering them can make the difference between a good or bad product and this was valid even before the spreading of social media. As seen, sentiment analysis can be a powerful tool for deriving a lot of meaningful results for business, especially when dealing with brand and product perception, detection of trends and consumer engagement (Altexsoft, 2018) (Marketing Inside, 2017). These particular points were some of the addressed outcomes that arose from this analysis. In fact, what was emerged is that levering social media wisely is a low-cost way to drive an increase in company value in the long run. Thus, another implication of this study is that social listening potentially covers a crucial rule for the company's business. In fact, it can allow to amplify positive perception factors and understand the reason behind certain setbacks or flaws that can be then tackled. In addition, if the study is targeted to the competitors, it is possible to derive the reasons behind their peaks and lows, thus not only having a more complete competitive overview but also to align or improve one's own offering in comparison to the other operators in the market. As emerged from the study, this is something that Tesla as a company did and is still doing and explains a considerable part of its success both on financial and marketing terms. In fact, it was able to individuate great trends like the green economy and the potential revolution of electric transport without renouncing to quality, that is arguably one of the factors that is valued the most by consumers. Having taken all of this into account, it is not a surprise that the Cybertruck entered the electric pickup market as first-comer or at least as the product to compare others' with: the moves made by Elon Musk's company in that sense not only anticipated the competition, but also gave a big amplification to the product that then was translated in company's value and improved brand perception, even, and in a way especially, in adverse circumstances. That is an implication that companies can derive out of the Cybertruck and Tesla case from a marketing perspective.

In addition, understanding this can be an opportunity for both incumbent and newcomers in every sector to rely on data to constantly improve their business without stopping to listen what the final customer has to say. As a matter of fact, text mining can be seen as an analytical study about words and emotions, that in other

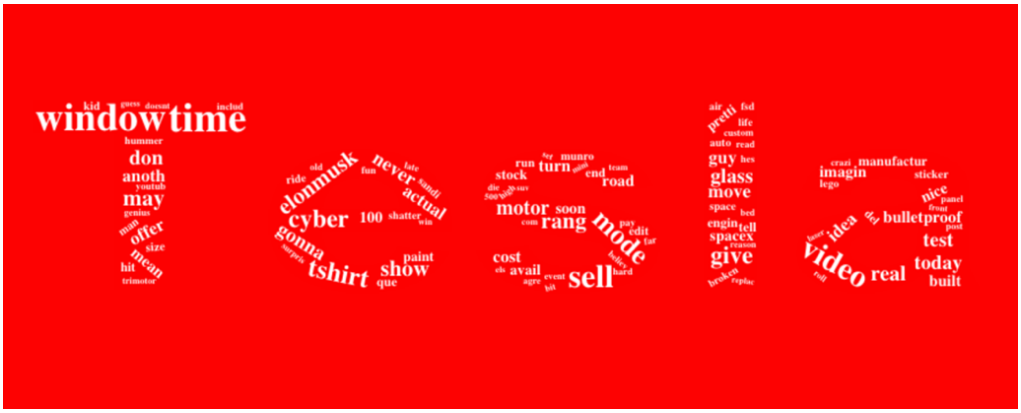
words can mean to get "the best of both worlds". Nevertheless, these tools do not have to substitute entirely the human work, as it is necessary to address the study at issue and to inspect for imprecision and errors that even machines can make. Again, levering both parts is a successful way for conducting an analysis. Moreover, a study that takes into account different approaches can lead to more compete outcomes and better findings, especially when two different ways of doing things are complementary.

Research limits and future perspectives

This research presents some limitations, that are especially related to the gathering of implementable data. In fact, the code could retrieve very few tweets for Tesla' competitors and this factor, as seen, limited the findings in the competitive analysis part. In addition, the code and Twitter authorizations allowed to retrieve not only a limited number of tweets, but also in a limited time range. That is why the gathering phase for Cybertruck started in January, two months after its unveiling, thus not allowing to capture the initial thoughts and opinions of Twitter users. For the same reason, not each tweet posted monthly could be gathered, but only a sample of them. Moreover, the electric pickup market in the USA is still immature as it is developing right in these times, thus adding an ulterior reason of the fact that the data are limited. Despite being official, some of the major Cybertruck's competitors still has not shared a lot of information about their new products, thus the research had to be narrowed to just three of them. In addition, there could have been different ways of conducting this sentiment study: the methods used were seen as adapt for this kind of project since it was more focused and addressed on deriving results rather than create and evaluate new computational models. Moreover, their usage on previous studies was seen as a confirmation of their value. Then, only English tweets were considered, as the majority of text mining tools cannot be applied to every language and very few tweets were available in other languages.

Moving back to the electric pickup, despite being a limitation now, the fact that this market is emerging represents one of the most exciting future perspective, as the evolution of the sector can still be tracked together with the general megatrend of green economy. In other words, in the future there will arguably be more data available for studying the perception of users and customers for a text mining study, and similarly, new computational, machine learning and Natural Language Processing tools developments, for instance by perfecting the use of tools in different languages, can broaden the field of research and add new researches to the ones already made.

Figure 32: Tesla Word Cloud



Retrieved from R Studio

Conclusion

The study has gone through the identification of the elements that characterizes the text mining field of study, describing some of its main form of implementation and techniques used to process large amount of textual information, highlighting also the relationship it has with machine learning and Natural Language Processing. Moreover, the subfields of sentiment and social media analysis were presented, highlighting how the tools arising from text mining can be adapted for this specific objective and domain, respectively. In particular, these two composed the structure of this study, as sentiment analysis was applied to the Twitter domain.

Furthermore, the object of this study was presented. Firstly, the company Tesla was introduced, with its totally electric transport vision and its ability to derive value out communication efficiently, especially thanks to its CEO Elon Musk. Therefore, the challenges in the change of the automotive ecosystem towards the total electrification of the vehicles were showed and discussed, evidencing the important role that Tesla had and still has in this conversation. Then, the actual product of interest for the study was discussed: this is the Tesla Cybertruck, a polarizing electric pickup truck that has made itself heard for its peculiar shape and its controversial unveiling event. As seen, these contributed to make this model really popular and talked about, thus making interesting to be analysed from a sentiment perspective. The last chapter was then focused on the sentiment analysis of this product, introducing programming and computational tools to conduct the study in the chosen domain Twitter. In particular, the findings were divided into four rounds: the first three were aimed at studying the Cybertruck with ever deeper levels of analysis, divided into different time ranges. The first was focused on a monthly analysis of single words, assessing their frequency and sentiment value through a lexicon-emotion association. The second was conducted on the general dataset focused on the identification of relevant topics of discussion in tweets through the use of a probabilistic method called Latent Dirichlet Allocation or LDA. After having individuated them, a lexicon-based sentiment analysis of the tweets containing some of these most relevant topics was conducted, to understand the opinion value for these specific arguments. The third was again not monthly and the aim was to adopt a lexicon that can retain as more textual information form tweets as possible. A special dictionary was then used that could give value to slangs, emoticon, emoji, valence shifters, amplifiers, adversatives and question weights, thus capturing more

information contained in a sentence and deepening the level of analysis. Lastly, the tweets of Cybertruck were compared to the ones of three competitors, namely GMC Hummer EV, Rivian R1T, Ford F-150 Electric, in the fourth round. This analysis was conducted though special functions that allowed a sentiment analysis comparison for different set of tweets. In fact, the score for each brand and their sentiment evolution over time were then confronted. For all these rounds, specific visualization tools were implemented to better highlight the findings of these analysis.

R Script

```
########################################TESLA CYBERTRUCK R SCRIPT#########################################


## ROUND "0" IS ABOUT GETTING THE TWEETS AND CLEANING THEM

# Packages to install and general packages to activate, plus citations

install.packages(c("tidyr","dplyr","xlsx","openxlsx","ggplot2","twitteR","tm","wordcloud","wordcloud2",
                   "tidytext","devtools","reshape2","syuzhet","lubridate","scales","igraph","ggraph",
                   "widyr","topicmodels","textdata","quanteda","sentimentr","saotd"))

library(tidyr)
library(dplyr)
library(xlsx)
library(openxlsx)
library(ggplot2)

citation("tidyr")
citation("dplyr")
citation("xlsx")
citation("openxlsx")
citation("ggplot2")
citation("twitteR")
citation("tm")
citation("wordcloud")
citation("wordcloud2")
citation("tidytext")
citation("devtools")
citation("reshape2")
citation("syuzhet")
citation("lubridate")
citation("scales")
citation("igraph")
citation("ggraph")
citation("widyr")
citation("topicmodels")
citation("textdata")
citation("quanteda")
citation("sentimentr")
citation("saotd")


# Getting Tweets: the action was repeated for each brand with a monthly frequence
```

```
library(twitteR)

Consumer_key <- "xxxxxxxxxxx"
Consumer_secret <- "xxxxxxxxxxxxxxx"
Access_token <- "xxxxxxxxxxxxxxxxx"
Access_token_secret <- "xxxxxxxxxxxxxxxxx"

setup_twitter_oauth(Consumer_key,
                    Consumer_secret,
                    Access_token,
                    Access_token_secret)

Cybertruck_English_Mentions <- searchTwitter("cybertruck -filter:retweets",
                                             n = 5000,
                                             lang = "en",
                                             since = NULL,
                                             until = NULL,
                                             retryOnRateLimit = 120)

Cybertruck_English_Mentions_dataframe <- twListToDF(Cybertruck_English_Mentions)

Cybertruck_English_Hashtag <- searchTwitter("#cybertruck -filter:retweets",
                                            n=5000,
                                            lang="en",
                                            since=NULL,
                                            until=NULL,
                                            retryOnRateLimit=120)

Cybertruck_English_Hashtag_dataframe <- twListToDF(Cybertruck_English_Hashtag)

# The following codes were repeated for each month sample: January, February, March, April and May

Cybertruck_May <- rbind(Cybertruck_English_Mentions_dataframe,
                                  Cybertruck_English_Hashtag_dataframe)

Cybertruck_May <- distinct(Cybertruck_May)

write.xlsx(Cybertruck_May,"Cybertruck_May.xlsx")

Cybertruck <- Cybertruck_May

# Build Corpus

library(tm)

corpus <- iconv(Cybertruck$text)
corpus <- Corpus(VectorSource(corpus))
inspect(corpus[1:5])

# Cleaning Phase

corpus <- tm_map(corpus, tolower)

removeUsernames <- function(x)gsub("@\\w+", " ",x)
corpus <- tm_map(corpus, content_transformer(removeUsernames))

corpus <- tm_map(corpus, removePunctuation)
```

```
removeURL <- function(x) gsub("http[[:alnum:]]*","",x)
removeURL2 <- function(x) gsub("http[[:alnum:]]*","",x)
corpus <- tm_map(corpus, content_transformer(removeURL))
corpus <- tm_map(corpus, content_transformer(removeURL2))

removeSymbols <- (function(x) gsub("[^[:alnum:]]", " " ,x))
corpus <- tm_map(corpus, content_transformer(removeSymbols))

cleanset <- tm_map(corpus, removeWords, stopwords("english"))
inspect(cleanset[1:5])

cleanset <- tm_map(cleanset, removeWords, c("tesla","cybertruck","keep", "check", "can","just","isnt","hey",
                                            "ask","theyr","dont","theyre","cmon","htt","everything",
                                            "even","enough","fuck", "wont", "rt", "tsla", "shit",
                                            "car","truck","pickup","mais","get","la","im","thats",
                                            "fevercureheres","here's","heres","theres","earning",
                                            "ch","whic","ive","debi","teslas","cybrtrk","wai","will","put",
                                            "much","able","make","made","says","say","said","cant","got",
                                            "come"))

cleanset <- tm_map(cleanset, content_transformer(function(s){
  gsub(pattern = '[^a-zA-Z0-9\\s]+',
       x = s,
       replacement = "",
       ignore.case = TRUE,
       perl = TRUE)}))

cleanset <- tm_map(cleanset, stemDocument)

removeDup <- function(x) unique(x)
cleanset <- tm_map(cleanset, removeDup)

inspect(cleanset[1:5])

Cybertruck_Clean <- do.call("rbind", lapply(cleanset, as.data.frame))

Cybertruck_Clean <- Cybertruck_Clean[Cybertruck_Clean$`X[[i]]` != "", ]

Cybertruck_Clean <- do.call("rbind", lapply(Cybertruck_Clean, as.data.frame))

Cybertruck_Clean <- na.omit(Cybertruck_Clean)

Cybertruck_Clean <- distinct(Cybertruck_Clean)

names(Cybertruck_Clean)[1] <- "word"



#FINDINGS ROUND 1: DATA VISUALIZATION: MONTHLY

tdm <- TermDocumentMatrix(cleanset)
tdm
tdm <- as.matrix(tdm)
tdm[1:10, 1:20]

# BarPlot: the proportions were adjusted according to the dataset, w >50 or w >100
```

```
w <- rowSums(tdm)
w <- subset(w,w>50)

ylim <- c(0, 1.1*max(w))
barplot(w, las=2, col = rainbow(50), ylim = ylim)

# WordCloud

library(wordcloud)

w <- sort(rowSums(tdm), decreasing = TRUE)
set.seed(222)
wordcloud(words = names(w), freq = w,
          max.words = 150,
          random.order = F,
          min.freq = 5,
          colors = brewer.pal(8,"Dark2"),
          scale = c(7,0.3),
          rot.per = 0.1)


# Sentiment Analysis

library(syuzhet)
library(lubridate)
library(scales)

# Read File

tweets <- iconv(Cybertruck_Clean$word, to = "utf-8-mac")
tweets <- Cybertruck_Clean$word

# Obtain Sentiment Scores

sentiment_nrc <- get_nrc_sentiment(tweets)
head(sentiment_nrc)

# BarPlot: the proportions were adjusted according to the dataset

barplot(colSums(sentiment_nrc),
        las = 2,
        col = rainbow(10),
        ylab = "Count",
        main = "Sentiment Score for Tesla Cybertruck Tweets: May",
        ylim = c(0,600))




# Analyzing the relationship between words: monthly bigrams

cybertruck.bigrams <- Cybertruck_Clean %>%
  unnest_tokens(bigram, word, token = "ngrams", n = 2)

bigrams.separated <- cybertruck.bigrams %>%
  separate(bigram, c("word1", "word2"), sep = " ")

bigrams.filtered <- bigrams.separated %>%
```

```
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word)

bigram.counts <- bigrams.filtered %>%
  count(word1, word2, sort = TRUE)

bigram.counts

bigrams.united <- bigrams.filtered %>%
  unite(bigram, word1, word2, sep = " ")

bigrams.united

bigram.tf.idf <- bigrams.united %>%
  count(bigram, bigram) %>%
  bind_tf_idf(bigram, bigram, n) %>%
  arrange(desc(tf_idf))

bigrams.separated %>%
  filter(word1 == "not") %>%
  count(word1, word2, sort = TRUE)

library(igraph)

# original counts

bigram.counts

# filter for only relatively common combinations

bigram.graph <- bigram.counts %>%
  filter(n > 20) %>%
  graph_from_data_frame()

bigram.graph

library(ggraph)
set.seed(2017)

ggraph(bigram.graph, layout = "fr") +
  geom_edge_link() +
  geom_node_point() +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1)

set.seed(2016)

a <- grid::arrow(type = "closed", length = unit(.15, "inches"))

ggraph(bigram.graph, layout = "fr") +
  geom_edge_link(aes(edge_alpha = n), show.legend = FALSE,
                 arrow = a, end_cap = circle(.07, 'inches')) +
  geom_node_point(color = "lightblue", size = 5) +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1) +
  theme_void()
```

```
## FINDINGS ROUND 2: THE THIRD ROUND COMBINES LDA WITH SENTIMENT ANALYSIS
## TO ASSESS THE IMPACT OF THE MOST RELEVANT WORDS OF TOPICS

# REPROPOSING THE SAME CLEANING PHASE BUT FOR THE COMPLETE SET OF CYBERTRUCK TWEETS, SO NOT MONTHLY

# Build Corpus

Cybertruck <- rbind(Cybertruck_January, Cybertruck_February,
                    Cybertruck_March, Cybertruck_April,Cybertruck_May)

corpus1 <- iconv(Cybertruck$text)
corpus1 <- Corpus(VectorSource(corpus1))
inspect(corpus1[1:5])

corpus1 <- tm_map(corpus1, tolower)

removeUsernames <- function(x)gsub("@\\w+", " ",x)
corpus1 <- tm_map(corpus1, content_transformer(removeUsernames))

corpus1 <- tm_map(corpus1, removePunctuation)

removeURL <- function(x) gsub("http[[:alnum:]]*","",x)
removeURL2 <- function(x) gsub("http[[:alnum:]]*","",x)
corpus1 <- tm_map(corpus1, content_transformer(removeURL))
corpus1 <- tm_map(corpus1, content_transformer(removeURL2))

removeSymbols <- (function(x) gsub("[^[:alnum:]]", " " ,x))
corpus1 <- tm_map(corpus1, content_transformer(removeSymbols))

cleanset1 <- tm_map(corpus1, removeWords, stopwords("english"))
inspect(cleanset1[1:5])

cleanset1 <- tm_map(cleanset1, removeWords, c("tesla","cybertruck","keep", "check", "can","just","isnt","hey",
                                              "ask","theyr","dont","theyre","cmon","htt","everything",
                                              "even","enough","fuck", "wont", "rt", "tsla", "shit",
                                              "car","truck","pickup","mais","get","la","im","thats",
                                              "fevercureheres","here's","heres","theres","earning",
                                              "ch","whic","ive","debi","teslas","cybrtrk","wai","will","put",
                                              "much","able","make","made","says","say","said","cant","got",
                                              "see"))

cleanset1 <- tm_map(cleanset1, content_transformer(function(s){
  gsub(pattern = '[^a-zA-Z0-9\\s]+',
       x = s,
       replacement = "",
       ignore.case = TRUE,
       perl = TRUE)}))

cleanset1 <- tm_map(cleanset1, stemDocument)

removeDup <- function(x) unique(x)
cleanset1 <- tm_map(cleanset1, removeDup)

inspect(cleanset1[1:5])

Cybertruck_Clean_1 <- do.call("rbind", lapply(cleanset1, as.data.frame))
names(Cybertruck_Clean_1)[1] <- "word"
```

```
dtm = DocumentTermMatrix(cleanset1)
dtm

dim(dtm)

doc.length = apply(dtm, 1, sum)
dtm = dtm[doc.length > 0,]
dtm

inspect(dtm[1:2,10:15])

freq = colSums(as.matrix(dtm))
length(freq)

ord = order(freq, decreasing = TRUE)
freq[head(ord, n = 20)]

# Creating a subset of words having more than 200 frequency

plot = data.frame(words = names(freq), count = freq)

plot = subset(plot, plot$count > 200)

str(plot)

ggplot(data = plot, aes(words, count)) +
  geom_bar(stat = 'identity') +
  ggtitle('Words used more than 200 times')+
  coord_flip()

# HERE IT COMES LDA

library(topicmodels)

# LDA model with 5 topics selected

lda_5 = LDA(dtm, k = 5, method = 'Gibbs',
            control = list(nstart = 5, seed = list(1505,99,36,56,88), best = TRUE,
                           thin = 500, burnin = 4000, iter = 2000))

# LDA model with 2 topics selected

lda_2 = LDA(dtm, k = 2, method = 'Gibbs',
            control = list(nstart = 5, seed = list(1505,99,36,56,88), best = TRUE,
                           thin = 500, burnin = 4000, iter = 2000))

# LDA model with 10 topics selected

lda_10 = LDA(dtm, k = 10, method = 'Gibbs',
             control = list(nstart = 5, seed = list(1505,99,36,56,88), best = TRUE,
                            thin = 500, burnin = 4000, iter = 2000))

# Top 10 terms or words under each topic

top10terms_5 = as.matrix(terms(lda_5,10))
top10terms_2 = as.matrix(terms(lda_2,10))
top10terms_10 = as.matrix(terms(lda_10,10))
```

```
top10terms_5

top10terms_2

top10terms_10

lda.topics_5 = as.matrix(topics(lda_5))
lda.topics_2 = as.matrix(topics(lda_2))
lda.topics_10 = as.matrix(topics(lda_10))

# write.csv(lda.topics_5,file = paste('LDAGibbs',5,'DocsToTopics.csv'))
# write.csv(lda.topics_2,file = paste('LDAGibbs',2,'DocsToTopics.csv'))
# write.csv(lda.topics_10,file = paste('LDAGibbs',10,'DocsToTopics.csv'))

summary(as.factor(lda.topics_5[,1]))

summary(as.factor(lda.topics_2[,1]))

summary(as.factor(lda.topics_10[,1]))

topicprob_5 = as.matrix(lda_5@gamma)
topicprob_2 = as.matrix(lda_2@gamma)
topicprob_10 = as.matrix(lda_10@gamma)

# write.csv(topicprob_5, file = paste('LDAGibbs', 5, 'DoctToTopicProb.csv'))
# write.csv(topicprob_2, file = paste('LDAGibbs', 2, 'DoctToTopicProb.csv'))
# write.csv(topicprob_10, file = paste('LDAGibbs', 10, 'DoctToTopicProb.csv'))

head(topicprob_5,4)

# Tokenizing character vector file 'tweets'.

library(textdata)
library(tidytext)

tweets1 <- Cybertruck_Clean_1$word

# ASSESSING SENTIMENT OF MOST IMPORTANT TOPICS EXTRACTED FROM LDA

#Note: the corpus was repeated with different words than "elon" that is the one showed below.
# the other words for which this analysis was conducted were
#tshirt, window, gigafactori, electr, design, elon, order

library(quanteda)

corpus_elon = subset(tweets1, grepl("elon", texts(tweets1)))
writeLines(as.character(corpus_elon[[150]]))

#Tokenizing character vector file 'tweets'.

token_elon = data.frame(text=corpus_elon, stringsAsFactors = FALSE) %>% unnest_tokens(word, text)

#Matching sentiment words from the 'NRC' sentiment lexicon

senti_elon = inner_join(token_elon, get_sentiments("nrc")) %>%
  count(sentiment)

senti_elon$percent = (senti_elon$n/sum(senti_elon$n))*100
```

```
#Plotting the sentiment summary

ggplot(senti_elon, aes(sentiment, percent)) +
  geom_bar(aes(fill = sentiment), position = 'dodge', stat = 'identity')+
  ggtitle("Sentiment analysis based on lexicon: 'NRC' for word: elon")+
  coord_flip() +
  theme(legend.position = 'none', plot.title = element_text(size=18, face = 'bold'),
        axis.text=element_text(size=16),
        axis.title=element_text(size=14,face="bold"))




## FINDINGS ROUND 3: THE FOURTH ROUND CONDUCTS A SENTIMENT ANALYSIS AT A DEEPER LEVEL, SINCE EMOTICONS,
NEGATIONS
## AND AMPLIFIERS ARE TAKEN INTO ACCOUNT IN THE CALCULATION OF THE SENTIMENT SCORE: THIS IS POSSIBLE THANKS
## TO THE SENTIMENTR PACKAGE: NOT MONTLHY

library(sentimentr)

show(lexicon::hash_sentiment_jockers) # Syuzhet lexicon. Range: -1; 1

show(lexicon::hash_sentiment_jockers_rinker) # Syuzhet Lexicon upgraded. Range: -1; 1

show(lexicon::hash_sentiment_huliu) # Bing Lexicon. Lexicon for general tasks. WordNet Bootstrapping + Product
Reviews
#Range: -2; -1.5; -1; 0; 1

show(lexicon::hash_sentiment_loughran_mcdonald) # Loughran-McDonald finance-specific Dictionary.
# Range: -1; 1

show(lexicon::hash_sentiment_nrc) # The lexicon Used for the previous analysis. Range: -1;1

show(lexicon::hash_sentiment_senticnet) # Used Concept-level sentiment analysis and largest and largest
dataset.
#Range: -0.98; 0.982

show(lexicon::hash_sentiment_sentiword) # Contains a large set of words and synset. Range: -1; 1

show(lexicon::hash_sentiment_socal_google) # Semi-supervised lexicon with small set of seed words
#Range: -30.160085; 30.738912

Jockers <-lexicon::hash_sentiment_jockers
Jockers_Rinker <-lexicon::hash_sentiment_jockers_rinker
Huliu <-lexicon::hash_sentiment_huliu
Loughran <- lexicon::hash_sentiment_loughran_mcdonald
NRC <-lexicon::hash_sentiment_nrc
Senticnet <-lexicon::hash_sentiment_senticnet
Sentiword <-lexicon::hash_sentiment_sentiword
Socal <-lexicon::hash_sentiment_socal_google

# The following code line was repeated for each dictionary to inspect the sentiment score range
summary(Jockers$y)

show(lexicon::hash_sentiment_emojis)
show(lexicon::hash_valence_shifters)
show(lexicon::hash_emoticons)
```

```
show(lexicon::hash_sentiment_slangsd)
show(lexicon::emojis_sentiment)

emoji <-lexicon::hash_sentiment_emojis
slang <-lexicon::hash_sentiment_slangsd
sentiment<-lexicon::hash_sentiment_jockers_rinker
emoticon <-lexicon::hash_emoticons

replace_emoticon(":D")

Cybertruck <- rbind(Cybertruck_January, Cybertruck_February,
                    Cybertruck_March, Cybertruck_April,Cybertruck_May)

Cybertruck_text<- Cybertruck$text

# Cleaning Phase

corpus1 <- iconv(Cybertruck_text)
corpus1 <- Corpus(VectorSource(corpus1))
inspect(corpus1[1:5])

corpus1 <- tm_map(corpus1, tolower)

removeUsernames <- function(x)gsub("@\\w+", " ",x)
corpus1 <- tm_map(corpus1, content_transformer(removeUsernames))

corpus1 <- tm_map(corpus1, removeWords, c("rt"))

cleanset1 <- tm_map(corpus1, removeWords, stopwords("english"))
inspect(cleanset1[1:5])

remove <- function(x)gsub("http.+ |http.+$", " ", x)
cleanset1 <- tm_map(cleanset1, content_transformer(remove))

inspect(cleanset1[1:5])

removeDup <- function(x) unique(x)
cleanset1 <- tm_map(cleanset1, removeDup)

cleanset1 <- tm_map(cleanset1, stripWhitespace)

Cybertruck_Clean_1<- do.call("rbind", lapply(cleanset1, as.data.frame))
names(Cybertruck_Clean_1)[1] <- "word"

Cybertruck_Clean_1<-Cybertruck_Clean_1$word
Cybertruck_Clean_1<-replace_emoticon(Cybertruck_Clean_1)
Cybertruck_Clean_1<-replace_emoji(Cybertruck_Clean_1)
Cyberturck_Clean_1<-replace_internet_slang(Cybertruck_Clean_1)

corpus2 <- iconv(Cybertruck_Clean_1)
corpus2 <- Corpus(VectorSource(corpus2))

smile <- function(x)gsub("smiley", "smile", x)
corpus2 <- tm_map(corpus2, content_transformer(smile))

Cybertruck_Clean_1<- do.call("rbind", lapply(corpus2, as.data.frame))
names(Cybertruck_Clean_1)[1] <- "word"
```

```
Cybertruck_Clean_1 <- Cybertruck_Clean_1$word

Cybertruck_Clean_Sentences <- get_sentences(Cybertruck_Clean_1)

Cybertruck_Sentiment <-sentiment(Cybertruck_Clean_Sentences,
                                 polarity_dt = lexicon::hash_sentiment_senticnet,
                                 emoticon_dt = lexicon::hash_emoticons,
                                 valence_shifters_dt = lexicon::hash_valence_shifters,
                                 slang = lexicon::hash_sentiment_slangsd,
                                 emoji_dt= lexicon::hash_sentiment_emojis,
                                 hyphen = "",
                                 amplifier.weight = 0.8, n.before = 5, n.after = 2,
                                 question.weight = 1, adversative.weight = 0.25,
                                 neutral.nonverb.like = FALSE, missing_value = 0)


Cybertruck_Sentiment<-na.omit(Cybertruck_Sentiment)

Bar_1 <- hist(Cybertruck_Sentiment$sentiment,
              main = "Sentiment Analysis: sentimentr",
              xlab = "Sentiment Score",
              ylab = "Frequency", beside = T,
              legend.text = rownames(Cybertruck_Sentiment),
              col=heat.colors(15),
              ylim = c(0,8500))


install.packages("SentimentAnalysis")
library(SentimentAnalysis)


Cybertruck_Sentiment_Score <-cut(Cybertruck_Sentiment$sentiment, breaks= c(-2.5,-1.51,-0.31,0.30,1.5,2.5),
                                 labels = c("Very Negative","Negative",
                                            "Neutral","Positive","Very Positive"))

Cybertruck_Sentiment_Score <- do.call("rbind", lapply(Cybertruck_Sentiment_Score , as.data.frame))

attach(Cybertruck_Sentiment_Score )
names(Cybertruck_Sentiment_Score )[1] <- "sentiment"

Sentiment_plot <- Cybertruck_Sentiment_Score  %>%
  group_by(sentiment) %>%
  summarise(word_count = n()) %>%
  ungroup() %>%
  mutate(sentiment = reorder(sentiment, word_count)) %>%
  ggplot(aes(sentiment, word_count, fill = sentiment)) +
  geom_col() +
  guides(fill = FALSE) +
  theme_classic() +
  labs(x = NULL, y = "Tweet Count") +
  scale_y_continuous(limits = c(0, 20000)) +
  ggtitle("Cybertruck Sentiment: sentimentr") +
  coord_flip()

Sentiment_plot
```

```
## FINDINGS ROUND 4: COMPETITIVE SENTIMENT ANALYSIS FOR ELECTRIC PICKUP CYBERTRUCK, F-150 ELECTRIC, GMC
# HUMMER EV AND RIVIAN R1T AND EVOLUTION OF SCORE OVERTIME

library(saotd)

Cybertruck_trimester <- rbind(Cybertruck_March, Cybertruck_April,Cybertruck_May)
F_150_electric_trimester <- rbind(F_150_electric_March, F_150_electric_April,F_150_electric_May)
GMC_Hummer_Ev_trimester <- rbind(GMC_Hummer_Ev_March, GMC_Hummer_Ev_April,GMC_Hummer_Ev_May)
R1T_trimester <- rbind(R1T_March, R1T_April,R1T_May)

names(Cybertruck_trimester) [11] <- "key"
names(F_150_electric_trimester) [11] <- "key"
names(GMC_Hummer_Ev_trimester) [11] <- "key"
names(R1T_trimester) [11] <- "key"

Cybertruck_trimester$hashtag="Tesla Cybertruck"
F_150_electric_trimester$hashtag="Ford F-150 Electric"
GMC_Hummer_Ev_trimester$hashtag="GMC Hummer EV"
R1T_trimester$hashtag="Rivian R1T"

Electric.Cars <- rbind(Cybertruck_trimester,
                       F_150_electric_trimester,
                       GMC_Hummer_Ev_trimester,
                       R1T_trimester)

Electric.Cars <- distinct(Electric.Cars)

# Analysis Phase

TD_Tidy <- saotd::tweet_tidy(DataFrame = Electric.Cars)

TD_Tidy <- distinct(TD_Tidy)

TD_Tidy$Token[3:8] %>%
  knitr::kable("html")

saotd::unigram(DataFrame = Electric.Cars) %>%
  dplyr::top_n(10) %>%
  knitr::kable("html", caption = "Twitter data Uni-Grams")

saotd::bigram(DataFrame = Electric.Cars) %>%
  dplyr::top_n(10) %>%
  knitr::kable("html", caption = "Twitter data Bi-Grams")

saotd::trigram(DataFrame = Electric.Cars) %>%
  dplyr::top_n(10) %>%
  knitr::kable("html", caption = "Twitter data Tri-Grams")

TD_Bigram <- saotd::bigram(DataFrame = Electric.Cars)

TD_Trigram <- saotd::trigram(DataFrame = Electric.Cars)

saotd::bigram_network(BiGramDataFrame = TD_Bigram,
                      number = 50,
                      layout = "fr",
                      edge_color = "blue",
                      node_color = "black",
```

```
                      node_size = 3,
                      set_seed = 1234)

TD_Corr <- saotd::word_corr(DataFrameTidy = TD_Tidy,
                            number = 100,
                            sort = TRUE)

saotd::word_corr_network(WordCorr = TD_Corr,
                         Correlation = .5,
                         layout = "fr",
                         edge_color = "blue",
                         node_color = "black",
                         node_size = 1)

TD_Scores <- saotd::tweet_scores(DataFrameTidy = TD_Tidy,
                                 HT_Topic = "hashtag")

saotd::posneg_words(DataFrameTidy = TD_Tidy,
                    num_words = 10)

saotd::posneg_words(DataFrameTidy = TD_Tidy,
                    num_words = 10,
                    filterword = "shit")

saotd::tweet_max_scores(DataFrameTidyScores = TD_Scores,
                        HT_Topic = "hashtag")


saotd::tweet_corpus_distribution(DataFrameTidyScores = TD_Scores,
                                 color = "black",
                                 fill = "white")

saotd::tweet_min_scores(DataFrameTidyScores = TD_Scores,
                        HT_Topic = "hashtag")

saotd::tweet_max_scores(DataFrameTidyScores = TD_Scores,
                        HT_Topic = "hashtag",
                        HT_Topic_Selection = "f150")

saotd::tweet_distribution(DataFrameTidyScores = TD_Scores,
                          binwidth = 1,
                          HT_Topic = "hashtag",
                          color = "black",
                          fill = "white")

saotd::tweet_box(DataFrameTidyScores = TD_Scores,
                 HT_Topic = "hashtag")

saotd::tweet_violin(DataFrameTidyScores = TD_Scores,
                    HT_Topic = "hashtag")

saotd::tweet_time(DataFrameTidyScores = TD_Scores,
                  HT_Topic = "hashtag")

utils::vignette('saotd')
```

```
# A special wordcloud made for the general dataset

tdm_final <- TermDocumentMatrix(cleanset1)
tdm_final
tdm_final <- as.matrix(tdm_final)

# BarPlot

w <- rowSums(tdm_final)
w <- subset(w,w>50)

# WordCloud

w <- data.frame(names(w), w)
colnames(w) <- c("word", "frequency")

w <- w %>%
  anti_join(stop_words)


library(devtools)
devtools::install_github("lchiffon/wordcloud2")
library(wordcloud2)
library(reshape2)

letterCloud(w,
            word = "Tesla",
            color = "white",
            backgroundColor = "red")
```

Bibliography

Bibliography

Öztürk, N. & Ayvaz, S., 2018. Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis. *Telematics and Informatics,* pp. 136-147.

Abu-Alhaija, A. S., Nerina, R., Hashim, H. & Jaharuddin, N. S., 2018. Determinants of Customer Loyalty: A Review and Future Directions. *Australian Journal of Basic and Applied Sciences,* 12(7), pp. 106-111.

Aggarwal, C. & Zhai, C., 2012. *Mining Text Data.* s.l.:Springer Science & Business Media.

Altexsoft, 2018. *Sentiment Analysis: Types, Tools, and Use Cases.* [Online] Available at: https://www.altexsoft.com/blog/business/sentiment-analysis-types-tools-and-use-cases/

Anderson, M., 2018. *Tesla Cars Are Great - Their Ecosystem Strategy Not So Much.* [Online] Available at: https://www.forbes.com/sites/babson/2018/01/27/tesla-cars-are-great-their-ecosystem-strategy-not-so-much/

Anna Jurek, M. D. M. Y. B., 2015. Improved lexicon-based sentiment analysis for social media analytics. *Security Informatics,* 9 December.4(9).

Asiri, S., 2018. *Machine Learning Classifiers.* [Online] Available at: https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623

Azad, P., Navimipour, N. J., Rahmani, A. M. & Sharifi, A., 2019. The role of structured and unstructured data managing mechanisms in the Internet of things. *Cluster Computing.*

Bail, C., n.d. *Dictionary-Based Text Analysis in R.* [Online] Available at: https://cbail.github.io/SICSS_Dictionary-Based_Text_Analysis.html

Barba, P., 2019. *Machine Learning for Natural Language Processing.* [Online] Available at: https://www.lexalytics.com/lexablog/machine-learning-natural-language-processing

Bariso, J., 2020. *Elon Musk Announced a New Product Celebrating Tesla's Epic Fail--and It's a Major Lesson in Emotional Intelligence.* [Online] Available at: https://www.inc.com/justin-bariso/elon-musk-announced-a-new-product-celebrating-teslas-epic-fail-and-its-a-major-lesson-in-emotional-intelligence.html [Accessed 30 April 2020].

Barnes, J., Touileb, S., Øvrelid, L. & Velldal, E., 2019. *Lexicon information in neural sentiment analysis: a multi-task learning approach.* Turku, Finland, Linköping University Electronic Press, pp. 175-186.

Biswas, N., 2014. *Text Mining and its Business Applications.* [Online] Available at: https://www.codeproject.com/Articles/822379/Text-Mining-and-its-Business-Applications#_articleTop

Blei, D. M., 2012. Probabilistic topic models. *Communications of the acm,* 55(4).

Blei, D. M., Ng, A. Y. & Jordan, M. I., 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research ,* Volume , pp. 993-1022.

Bock, T., 2018. *What is Hierarchical Clustering?.* [Online] Available at: https://www.displayr.com/what-is-hierarchical-clustering/

Boost Labs, 2014. *Word Clouds & the Value of Simple Visualizations.* [Online] Available at: https://boostlabs.com/what-are-word-clouds-value-simple-visualizations/ [Accessed 21 May 2020].

Bradley, M. M. & Lang, P. J., 2017. *ANEW Message.* [Online] Available at: https://csea.phhp.ufl.edu/media/anewmessage.html

Brems, M., 2017. *A One-Stop Shop for Principal Component Analysis.* [Online] Available at: https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c

Brownlee, J., 2019. *A Tour of Machine Learning Algorithms.* [Online] Available at: https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

BSC, n.d. *Sparse Matrix.* [Online] Available at: http://www.btechsmartclass.com/data_structures/sparse-matrix.html [Accessed 19 May 2020].

Buckley, M. T. K. & Paltoglou, G., 2011. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology,* pp. 163-173.

Business Strategy Hub, 2020. *Tesla Business Model (2020) | Tesla Business Model Canvas.* [Online] Available at: https://bstrategyhub.com/tesla-business-model-tesla-business-model-canvas/ [Accessed 28 April 2020].

Business Strategy Hub, 2020. *Tesla SWOT Analysis (2020).* [Online] Available at: https://bstrategyhub.com/tesla-swot-analysis/

Businesswire, 2015. *New Crowdtap Study Looks Under the Hood at What Drives Automotive Buying Behaviors.* [Online] Available at: https://www.businesswire.com/news/home/20150409006096/en/New-Crowdtap-Study-Hood-Drives-Automotive-Buying

Cambridge Dictionary, n.d. *Meaning of analyse in English.* [Online] Available at: https://dictionary.cambridge.org/dictionary/english/analyse

Cambridge Dictionary, n.d. *Pickup truck.* [Online] Available at: https://dictionary.cambridge.org/dictionary/english/pickup-truck [Accessed 1 May 2020].

Cao□, J. et al., 2009. A density-based method for adaptive LDA model selection. *Neurocomputing,* Volume 72, p. 1775–1781.

Capparella, J., 2020. *GMC Hummer EV Reveal Delayed Due to COVID-19 Pandemic.* [Online] Available at: https://www.caranddriver.com/news/a32317151/gmc-hummer-ev-reveal-delayed/ [Accessed 25 May 2020].

Carrasco, O. C., 2019. *Support Vector Machines for Classification.* [Online] Available at: https://towardsdatascience.com/support-vector-machines-for-classification-fc7c1565e3

Castañón, J., 2019. *10 Machine Learning Methods that Every Data Scientist Should Know.* [Online] Available at: https://towardsdatascience.com/10-machine-learning-methods-that-every-data-scientist-should-know-3cc96e0eeee9

Chang, J. et al., 2018. *Text mining 101.* [Online] Available at: https://www.fosteropenscience.eu/content/text-mining-101

Cho, H., Kim, S., Lee, J. & Lee, J.-S., 2014. Data-driven integration of multiple sentiment dictionaries for lexicon-based sentiment classification of product reviews. *Knowledge-Based Systems,* pp. 61-71.

Choi, Y. & Wiebe, J., n.d. +/-EffectWordNet: Sense-level Lexicon Acquisition for Opinion Inference.

Cho, Y.-J., Fu, P.-W. & Wu, C.-C., 2017. Popular Research Topics in Marketing Journals, 1995–2014. *Journal of Interactive Marketing,* Volume 40, pp. 52-72.

CleanTechnica, 2019. *Tesla Gigafactory Bigger Than You Expect — Truly Gigantic.* [Online] Available at: https://cleantechnica.com/2019/07/21/tesla-gigafactory-bigger-than-you-expect-truly-gigantic/

Cox Automotive, 2019. *Car Buyers Visiting Fewer Dealerships, Making Faster Decisions as Online Engagement Rises.* [Online]

Available at: https://www.coxautoinc.com/news/car-buyers-visiting-fewer-dealerships-making-faster-decisions-as-online-engagement-rises/

CRAN, 2019. *Sentiment Analysis of Twitter Data.* [Online] Available at: https://cran.r-project.org/web/packages/saotd/vignettes/saotd.html

Crum, R., 2020. *Elon Musk: Tesla to build Cybertruck in 'central USA'.* [Online] Available at: https://www.mercurynews.com/2020/03/11/elon-musk-tesla-to-build-cybertruck-in-central-usa/ [Accessed 25 May 2020].

De Moura, E. S., 2009. Text Indexing Techniques. In: *Encyclopedia of Database Systems.* Boston, MA: Springer.

DeBord, M., 2019. *The Tesla Cybertruck is the first stainless-steel vehicle since the ill-fated DeLorean — here's a closer look at both.* [Online] Available at: https://www.businessinsider.com/tesla-cybertruck-versus-delorean-stainless-steel-vehicles-2019-12?IR=T

[Accessed 30 April 2020].

Devopedia, 2019. *Text Corpus for NLP.* [Online] Available at: https://devopedia.org/text-corpus-for-nlp [Accessed 19 May 2020].

Duel, 2018. *9 Advantages of Customer Advocacy Marketing.* [Online] Available at: https://www.duel.tech/blog/9-advantages-customer-advocacy-marketing [Accessed 2020 April 28].

Edureka, 2019. *Who uses R?.* [Online] Available at: https://www.edureka.co/blog/who-uses-r/ [Accessed 18 May 2020].

Edwards, G., 2018. *Machine Learning | An Introduction.* [Online] Available at: https://towardsdatascience.com/machine-learning-an-introduction-23b84d51e6d0

Eising, P., 2017. *What exactly IS an API?.* [Online] Available at: https://medium.com/@perrysetgo/what-exactly-is-an-api-69f36968a41f [Accessed 18 May 2020].

Electrek, 2020. *Tesla Cybertruck.* [Online] Available at: https://electrek.co/guides/tesla-cybertruck/ [Accessed 29 April 2020].

Electric Car Home, n.d. *BEV, PHEV, HEV, ICE – what on earth do they mean?.* [Online] Available at: https://electriccarhome.co.uk/electric-cars/bev-phev-hev-ice/

Elrhoul, M., 2015. *Research: Four ways brands can build customer service relationships on Twitter.* [Online] Available at: https://blog.twitter.com/en_us/a/2015/research-four-ways-brands-can-build-customer-service-relationships-on-twitter.html

Evannex, 2019. *Tesla Is The King Of Social Media, No Need For An Advertising Budget.* [Online] Available at: https://insideevs.com/news/365726/tesla-no-advertising-social-media-wins/ [Accessed 28 April 2020].

EVgo, n.d. *Types of Electric Vehicles.* [Online] Available at: https://www.evgo.com/why-evs/types-of-electric-vehicles/

Expert System Team, 2017. *Natural Language Process semantic analysis: definition.* [Online] Available at: https://expertsystem.com/natural-language-process-semantic-analysis-definition/

Expert System Team, 2017. *What is Machine Learning? A definition.* [Online] Available at: https://expertsystem.com/machine-learning-definition/

Fawcett, T., 2015. *The Basics of Classifier Evaluation: Part 1.* [Online] Available at: https://www.svds.com/the-basics-of-classifier-evaluation-part-1/

FBK-ICT, n.d. *SentiWords.* [Online] Available at: https://hlt-nlp.fbk.eu/technologies/sentiwords [Accessed 24 May 2020].

Feinerer, I. & Hornik, K., 2019. *tm: Text Mining Package.* s.l.:s.n.

Feinerer, I., n.d. *tm v0.7-7.* [Online] Available at: https://www.rdocumentation.org/packages/tm/versions/0.7-7 [Accessed 19 May 2020].

Folschette, C., 2019. *Tesla's marketing strategy shows that it's time for CEOs to get social.* [Online] Available at: https://www.talkwalker.com/blog/tesla-marketing-strategy-social-ceo [Accessed 29 April 2020].

Forbes, 2020. *How Does Tesla Spend Its Money?.* [Online] Available at: https://www.forbes.com/sites/greatspeculations/2020/01/03/how-does-tesla-spend-its-money/#2bc9a61425da [Accessed 2020 April 28].

Fröhlich, J., 2004. *Supervised and unsupervised learning.* [Online] Available at: https://www.nnwj.de/supervised-unsupervised.html

Gentry, J., 2014. *Twitter client for R.* [Online] Available at: http://geoffjentry.hexdump.org/twitteR.pdf

Gentry, J., 2015. *twitteR: R Based Twitter Client.* [Online] Available at: https://CRAN.R-project.org/package=twitteR

Gentry, J., n.d. *twitteR.* [Online] Available at: https://www.rdocumentation.org/packages/twitteR/versions/1.1.9 [Accessed 19 May 2020].

Gertner, J., 2020. *Rivian Wants to Bring Electric Trucks to the Masses.* [Online] Available at: https://www.wired.com/story/rivian-race-design-all-electric-pickup-truck/ [Accessed 1 May 2020].

GMC Pressroom, 2020. *Update on GMC HUMMER EV Reveal.* [Online] Available at: https://media.gmc.com/media/us/en/gmc/home.detail.html/content/Pages/news/us/en/2020/apr/0429-hummer.html
[Accessed 25 May 2020].

GMC, 2020. *The Quiet Revolution Charges On First Ever Gmc Hummer Ev.* [Online] Available at: https://www.gmc.com/electric-truck/hummer-ev
[Accessed 1 May 2020].

GNU Operating System, 2020. *What is GNU?.* [Online] Available at: http://www.gnu.org/
[Accessed 17 May 2020].

Gorzelany, J., 2019. *Here's Which New Electric Pickups Will Compete With Tesla's 'Cybertruck' By 2021.* [Online]
Available at: https://www.forbes.com/sites/jimgorzelany/2019/12/11/heres-which-new-electric-pickups-will-compete-with-teslas-cybertruck-by-2021/#53ea306c5fca
[Accessed 1 May 2020].

Gorzelany, J., 2019. *How Tesla's Rock-Solid Resale Values Can Make Buying A New One A Better Deal Than Buying A Used One.* [Online]
Available at: https://www.forbes.com/sites/jimgorzelany/2019/11/05/why-buying-a-new-tesla-can-be-a-better-deal-than-choosing-a-used-one/#41dabcb0322d

Govoni, L., n.d. *Sentiment Analysis per migliorare l'immagine aziendale.* [Online] Available at: https://lorenzogovoni.com/sentiment-analysis/

Grün, B. & Hornik, K., 2011. topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software,* 40(13), pp. 1-30.

Grün, B. et al., 2020. *Package 'topicmodels'.* [Online] Available at: https://cran.r-project.org/web/packages/topicmodels/topicmodels.pdf
[Accessed 22 May 2020].

Guru99, n.d. *What is R Programming Language? Introduction & Basics.* [Online] Available at: https://www.guru99.com/r-programming-introduction-basics.html

Hardeniya, T. & Borikar, D. A., 2016. An Approach To Sentiment Analysis Using Lexicons With Comparative Analysis of Different Techniques. *IOSR Journal of Computer Engineering,* 18(3), pp. 53-57.

Hartmann, J., Huppertz, J., Schamp, C. & Heitmann, M., 2018. Comparing automated text classification methods. *International Journal of Research in Marketing.*

Hawkins, A. J. & O'Kane, S., 2019. *Tesla Model Y Announced: Release Set For 2020, Price Starts At $47,000.* [Online]
Available at: https://www.theverge.com/2019/3/14/18264446/tesla-model-y-suv-compact-announcement-price-release-date-features-elon-musk

HDMotori.it, 2020. *Il Missouri vuole il Tesla Cybertruck | Sul piatto 1 miliardo di dollari.* [Online] Available at: https://www.hdmotori.it/tesla/articoli/n519611/tesla-cybertruck-missouri-fabbrica-gigafactory/ [Accessed 25 May 2020].

Hemmatian, F. & Sohrabi, M., 2019. A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review,* pp. 1495-1545.

Hernández, I. C., 2016. *The Pros and Cons of using Twitter for your Online Marketing.* [Online] Available at: https://www.matrixinternet.ie/the-pros-and-cons-of-twitter/

Hoffman, C., 2017. *What Is OAuth? How Those Facebook, Twitter, and Google Sign-in Buttons Work.* [Online]
Available at: https://www.howtogeek.com/53275/exchanging-data-safely-with-oauth/ [Accessed 18 May 2020].

Hong, J.-W. & Park, S.-B., 2019. The Identification of Marketing Performance Using Text Mining of Airline Review Data. *Mobile Information Systems.*

Hornik, 2020. *The R FAQ.* [Online] Available at: https://CRAN.R-project.org/doc/FAQ/R-FAQ.html [Accessed 17 May 2020].

Hossain, A., 2018. *Text analysis on the tweets about Bangladesh.* [Online] Available at: https://rpubs.com/arafath/twitter_analysis

Hotho, A., Nürnberger, A. & Paaß, G., 2005. A Brief Survey of Text Mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology,* pp. 19-62.

Howe, J. & Robinson, M., 2006. *Crowdsourcing: A definition.* [Online].

Hsu, C.-W., Chang, C.-C. & Lin, C.-J., 2016. A Practical Guide to Support Vector Classification. 19 May.

Hughes, J., 2019. *Anti-Ads: The Elon Musk Approach To Marketing In The Digital Age.* [Online] Available at: https://velocitize.com/2019/11/14/anti-ads-the-elon-musk-approach-to-marketing-in-the-digital-age/
[Accessed 29 April 2020].

Hull, D. & Zhang, C., 2019. *Elon Musk Set Up His Shanghai Gigafactory in Record Time.* [Online] Available at: https://www.bloomberg.com/news/articles/2019-10-23/elon-musk-opened-tesla-s-shanghai-gigafactory-in-just-168-days

Humphreys, A. & Wang, R. J.-H., 2018. Automated Text Analysis for Consumer Research. *Journal of Consumer Research,* 44(6), p. 1274–1306.

Hutto, C. & Gilbert, E., 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Eighth International AAAI Conference on Weblogs and Social Media.*

Iberdrola, n.d. *Electric Vehicle Positioning.* [Online] Available at: https://www.iberdrola.com/wcorp/gc/prod/en_US/conocenos/docs/Vehiculo_electrico.pdf

Imanuel, n.d. *What is text analytics?.* [Online] Available at: https://www.predictiveanalyticstoday.com/text-analytics/#bigdatatextanalyticsandpredictiveanalytics

Jockers, M., 2017. *Introduction to the Syuzhet Package.* [Online] Available at: https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html [Accessed 21 May 2020].

Jothilakshmi, S. & Gudivada†, V., 2016. Chapter 10 - Large Scale Data Enabled Evolution of Spoken Language Research and Applications. In: *Cognitive Computing : Theory and Applications.* s.l.:North Holland, pp. 301-340.

Jr., V. B., 2015. *On Twitter, retail events and auto shows are marketing gold.* [Online] Available at: https://www.autonews.com/article/20150612/BLOG06/150619958/on-twitter-retail-events-and-auto-shows-are-marketing-gold

Jurek, A., Mulvenna, M. & Bi, Y., 2015. Improved lexicon-based sentiment analysis for social media analytics. *Security Informatics,* 4(9).

K10, 2019. *26K comments from the Twitter community about the CyberTruck, helps create a Truck fit for a Hero..* [Online] Available at: https://medium.com/@K10teslakitten/26k-comments-from-the-twitter-community-about-the-cybertruck-helps-create-a-truck-fit-for-a-hero-ed5123fa663c [Accessed 30 April 2020].

Kabacoff, R. I., 2017. *Packages.* [Online] Available at: https://www.statmethods.net/interface/packages.html [Accessed 17 May 2020].

Kabir, A. I., Karim, R., Newaz, S. & Hossain, M. I., 2018. The Power of Social Media Analytics: Text Analytics Based on Sentiment Analysis and Word Clouds on R. *Informatica Economică,* 22(1), pp. 25-38.

Kahl, J., 2017. *Che cos'è l'Information Retrieval e di cosa si occupa.* [Online] Available at: https://jacopokahl.com/ir-information-retrieval-di-cosa-si-occupa/

Kane, M., 2020. *Plug-In Electric Car Sales In Europe: Record December And 2019.* [Online] Available at: https://insideevs.com/news/394870/plugin-sales-europe-record-december-2019/

Karani, D., 2018. *Introduction to Word Embedding and Word2Vec.* [Online] Available at: https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa

Katz, G., Ofek, N. & Shapira, B., 2015. ConSent: Context-based sentiment analysis. *Knowledge-Based Systems,* pp. 162-178.

Kauffmann, E. et al., 2019. Managing Marketing Decision-Making with Sentiment Analysis: An Evaluation of the Main Product Features Using Text Data Mining. *Sustainability,* 11(15), p. 4235.

Kennedy, A. & Inkpen, D., 2006. Sentiment Classification Of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence,* 22(2).

Khan, F. H., Qamar, U. & Bashir, S., 2016. SentiMI: Introducing point-wise mutual information with SentiWordNet to improve sentiment polarity detection. *Applied Soft Computing,* pp. 140-153.

Koehrsen, W., 2018. *Beyond Accuracy: Precision and Recall.* [Online] Available at: https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c

Kowshalya, A. M. & Valarmathi, M. L., 2018. Evaluating Twitter Data to Discover User's Perception About Social Internet of Things. *Wireless Personal Communications volume ,* p. 649–659.

Lambert, F., 2019. *Tesla Gigafactory 4 leaked plans reveal room for expansion.* [Online] Available at: https://electrek.co/2019/12/12/tesla-gigafactory-4-leaked-plans-reveal-room-for-expansions/

Lambert, F., 2020. *Elon Musk: Tesla Cybertruck will come standard with 'laser blade lights,' be ~82" wide.* [Online]
Available at: https://electrek.co/2020/02/21/tesla-cybertruck-standard-laser-blade-ligths-elon-musk/ [Accessed 21 May 2020].

Lambert, F., 2020. *New Tesla Cybertruck and electric ATV footage on new Jay Leno's Garage.* [Online] Available at: https://electrek.co/2020/04/28/tesla-cybertruck-electric-atv-footage-jay-leno-garaga/ [Accessed 21 May 2020].

Lambert, F., 2020. *Tesla's next factory is going to be in Austin, Texas, and it's going to happen quickly.* [Online]
Available at: https://electrek.co/2020/05/15/tesla-factory-austin-texas/ [Accessed 21 May 2020].

Lawer, C., 2006. Customer advocacy and brand development. *Journal of Product & Brand Management,* 15(2), pp. 121-129.

Leanse, A., 2019. *13 Tesla Cybertruck Features You Probably Missed While Being Distracted By Its Wild Looks.* [Online]
Available at: https://www.motortrend.com/news/tesla-cybertruck-features-missed-details/ [Accessed 30 April 2020].

Lee, T. B., 2020. *Tesla Is Now Worth More Than Ford and GM—Combined.* [Online] Available at: https://www.wired.com/story/tesla-worth-more-than-ford-gm-combined/

Levy, A., 2020. *Elon Musk says he's looking to open a Tesla Cybertruck factory in 'central USA'.* [Online] Available at: https://www.cnbc.com/2020/03/10/elon-musk-scouting-tesla-cybertruck-gigafactory-in-central-usa.html
[Accessed 21 May 2020].

Liao, S., 2019. *Elon Musk says Tesla has received 200,000 orders for the Cybertruck since its reveal.* [Online] Available at: https://edition.cnn.com/2019/11/23/cars/cybertruck-tesla-preorders/index.html [Accessed 30 April 2020].

Lienert, P., 2019. *Eight electric pickup truck manufacturers to load up U.S. market by 2021.* [Online] Available at: https://www.reuters.com/article/us-autos-electric-trucks/eight-electric-pickup-truck-

manufacturers-to-load-up-u-s-market-by-2021-idUSKBN1XZ1W7
[Accessed 1 May 2020].

Lin, Y., 2019. *10 Twitter Statistics Every Marketer Should Know in 2020 [Infographic].* [Online] Available at: https://www.oberlo.com/blog/twitter-statistics

Liu, B., 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions.* s.l.:Cambridge University Press.

Liu, E., 2015. *Latent Dirichlet Allocation Using Gibbs Sampling.* [Online] Available at: https://ethen8181.github.io/machine-learning/clustering_old/topic_model/LDA.html
[Accessed 22 May 2020].

MacroTrends, 2020. *Tesla - Stock Price History | TSLA.* [Online] Available at: https://www.macrotrends.net/stocks/charts/TSLA/tesla/stock-price-history

Mangram, M. E., 2012. The globalization of Tesla Motors: a strategic marketing plan analysis. *Journal of Strategic Marketing,* 20(4), pp. 289-312.

Manwani, N., 2018. *Generative Deep Learning : Let's seek how AI Extending, not Replacing Creative Process.* [Online]
Available at: https://towardsdatascience.com/generative-deep-learning-lets-seek-how-ai-extending-not-replacing-creative-process-fded15b0561b
[Accessed 22 May 2020].

Marinova, D., Singh, S. K. & Singh, J., 2018. Frontline Problem-Solving Effectiveness: A Dynamic Analysis of Verbal and Nonverbal Cues. *Journal of Marketing Research,* Volume 55.

Marketing Examples, 2019. *The genius of Tesla's marketing strategy.* [Online] Available at: https://marketingexamples.com/referral/tesla-marketing-strategy

Marketing Inside, 2017. *Sentiment Analysis.* [Online] Available at: https://www.insidemarketing.it/glossario/definizione/sentiment-analysis/

Marshall, A., 2019. *Want a Tax Credit for Buying an Electric Vehicle? Move Fast.* [Online] Available at: https://www.wired.com/story/want-tax-credit-buying-electric-vehicle-move-fast/

Marta, 2020. *A comprehensive guide to social media analysis.* [Online] Available at: https://brand24.com/blog/guide-to-social-media-analysis/

Mathur, M., 2018. *Sentiment Analysis.* [Online] Available at: https://rpubs.com/Malvika30/Brand-Perception-Sentiment-Analysis-R

Matousek, M., 2019. *I pre-ordered a $76,900 Tesla Cybertruck. Here's each step I took to reserve one..* [Online]
Available at: https://www.businessinsider.com/how-to-order-tesla-cybertruck-2019-12?IR=T
[Accessed 29 April 2020].

Matousek, M., 2019. *The 10 best-selling electric vehicles in the US this year so far.* [Online] Available at: https://www.businessinsider.com/best-selling-electric-vehicles-united-states-so-far-2019-2019-7?IR=T

Matousek, M., 2020. *Tesla just bested Toyota, Volkswagen, and Lincoln to become one of drivers' favorite brands.* [Online]
Available at: https://www.businessinsider.com/tesla-makes-big-leap-consumer-reports-car-brand-survey-ranking-2020-2?IR=T

Mayer, J. D., Salovey, P. & Caruso, D. R., 2008. Emotional intelligence: New ability or eclectic traits?. *American Psychologist,* September, 63(6), pp. 503-517.

McCarthy, N., 2019. *America's best-Selling Pickup Trucks.* [Online]
Available at: https://www.statista.com/chart/20067/us-sales-of-new-pickup-trucks/
[Accessed 1 May 2020].

McKinsey&Company, 2017. Electrifying insights: How automakers can drive electrified vehicle sales and profitability. *Advanced Industries.*

Mohammad, S. M., n.d. *NRC Word-Emotion Association Lexicon (aka EmoLex).* [Online]
Available at: https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
[Accessed 21 May 2020].

Mohammad, S. M. & Turney, P. D., 2010. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text.*

Mohammad, S. M. & Turney, P. D., 2013. Crowdsourcing a Word–Emotion Association Lexicon.

Mohammad, S. M. & Turney, P. D., 2013. Crowdsourcing a Word–Emotion Association Lexicon. *Institute for Information Technology, National Research Council Canada.,* 29(3), pp. 436-465.

MonkeyLearn, 2019. *Text Analysis.* [Online]
Available at: https://monkeylearn.com/text-analysis/

MonkeyLearn, 2019. *Text Mining: The Beginner's Guide.* [Online]
Available at: https://monkeylearn.com/text-mining/
[Accessed 14 May 2020].

Monkeylearn, 2020. *Sentiment Analysis.* [Online]
Available at: https://monkeylearn.com/sentiment-analysis/

Monsters, D., 2017. *Sentiment Analysis Tools Overview, Part 1. Positive and Negative Words Databases.* [Online]
Available at: https://medium.com/@datamonsters/sentiment-analysis-tools-overview-part-1-positive-and-negative-words-databases-ae35431a470c

Munro & Associates, Inc., 2020. *Back by Popular Demand: Sandy Munro Discusses Tesla Cybertruck on Autoline.* [Online]
Available at: https://leandesign.com/back-by-popular-demand-sandy-munro-discusses-tesla-cybertruck-on-autoline/
[Accessed 21 May 2020].

Munro & Associates, Inc., n.d. *Evolving How You Manufacture.* [Online] Available at: https://leandesign.com/ [Accessed 21 May 2020].

Munson, E. L., Smith, C. M., Boehmke, B. C. & Freels, J. K., 2019. *{saotd}: Sentiment Analysis of Twitter Data.* [Online] Available at: https://github.com/evan-l-munson/saotd

Musk, E., 2006. *The Secret Tesla Motors Master Plan (just between you and me).* [Online] Available at: https://www.tesla.com/it_IT/blog/secret-tesla-motors-master-plan-just-between-you-and-me

Musk, E., 2014. *All Our Patent Are Belong To You.* [Online] Available at: https://www.tesla.com/it_IT/blog/all-our-patent-are-belong-you?redirect=no

Musk, E., 2014. *Tesla and SpaceX: Elon Musk's industrial empire* [Interview] (30 March 2014).

Musk, E., 2018. *Elon Musk: The Recode interview* [Interview] (5 November 2018).

Musk, E., 2019. Los Angeles: s.n.

Musk, E., 2019. *Cybertruck design influenced partly by The Spy Who Loved Me.* [Online] Available at: https://twitter.com/elonmusk/status/1197638937109336066 [Accessed 30 April 2020].

Musto, C., Semeraro, G. & Polignano, M., 2014. *A comparison of Lexicon-based approaches for Sentiment Analysis of microblog posts.* [Online] Available at: http://ceur-ws.org/Vol-1314/paper-06.pdf

Naldi, M., 2019. *A review of sentiment computation methods with R packages.* [Online] Available at: https://arxiv.org/pdf/1901.08319.pdf [Accessed 24 May 2020].

Namugera, F., Wesonga, R. & Jehopio, P., 2019. Text mining and determinants of sentiments: Twitter social media usage by traditional media houses in Uganda. *Computational Social Networks volume,* 6(3).

Nigam, N. & Yadav, D., 2018. Lexicon-Based Approach to Sentiment Analysis of Tweets Using R Language. *Advances in Computing and Data Sciences.*

O'Kane, S., 2019. *Tesla still isn't getting enough batteries from Panasonic.* [Online] Available at: https://www.theverge.com/2019/4/11/18305976/tesla-panasonic-gigafactory-batteries-model-3

O'Kane, S., 2020. *Hot Wheels announces a remote-controlled Tesla Cybertruck.* [Online] Available at: https://www.theverge.com/2020/2/21/21147004/hot-wheels-tesla-cybertruck-rc-car-remote-controlled-toy [Accessed 21 May 2020].

Ordenes, F. V. et al., 2019. Cutting through Content Clutter: How Speech and Image Acts Drive Consumer Sharing of Social Media Brand Messages. *Journal of Consumer Research,* Volume 45.

Ordenes, F. V. et al., 2017. Unveiling What Is Written in the Stars: Analyzing Explicit, Implicit, and Discourse Patterns of Sentiment in Social Media. *Journal of Consumer Research.*

Ordenes, F. V. et al., 2014. Analyzing Customer Experience Feedback Using Text Mining: A Linguistics-Based Approach. *Journal of Service Research,* 17(3), pp. 278-295.

Pandarachalil, R., Sendhilkumar, S. & Mahalakshmi, G., 2015. Twitter Sentiment Analysis for Large-Scale Data: An Unsupervised Approach.. *Cogn Comput,* Volume 7, p. 254–262.

Paukert, C., 2020. *Tesla Cybertruck Lego kit is adorable, may see production.* [Online] Available at: https://www.cnet.com/roadshow/news/tesla-cybertruck-lego-ideas-elon-musk-von-holzhausen-project/

[Accessed 21 May 2020].

Pierre, S., 2019. *What does Twitter think of the New Tesla Cybertruck? Sentiment Analysis in Python.* [Online] Available at: https://towardsdatascience.com/what-does-twitter-think-of-the-new-tesla-cybertruck-sentiment-analysis-in-python-afa22a9aefce

[Accessed 16 June 2020].

Pierre, S., 2019. *What does Twitter think of the New Tesla Cybertruck? Sentiment Analysis in Python.* [Online] Available at: https://towardsdatascience.com/what-does-twitter-think-of-the-new-tesla-cybertruck-sentiment-analysis-in-python-afa22a9aefce

[Accessed 16 June 2020].

Popescu, A.-M. & Etzioni, O., 2005. *Extracting Product Features and Opinions from Reviews.* s.l., HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 339-346.

Popkin, H. A. S., 2018. *Elon Musk's Twitter Account Is Tesla's $40 Million Marketing Platform. 'Worth It'.* [Online]

Available at: https://www.forbes.com/sites/helenpopkin/2018/10/30/elon-musks-twitter-account-is-teslas-40-million-marketing-platform-worth-it/#5538ccc47873

[Accessed 29 April 2020].

Princeton University, 2010. *Princeton University "About WordNet.".* [Online] Available at: https://wordnet.princeton.edu

R Package Documentation, 2019. *get_sentiment: Get Sentiment Values for a String.* [Online] Available at: https://rdrr.io/cran/syuzhet/man/get_sentiment.html

[Accessed 24 May 2020].

Raia, J., 2020. *Rivian EV debut delayed by Covid-19 crisis.* [Online] Available at: https://theweeklydriver.com/2020/04/rivian-ev-debut-delayed-because-of-covid-10/

[Accessed 25 May 2020].

Rai, B., 2017. *R - Sentiment Analysis and Wordcloud with R from Twitter Data | Example using Apple Tweets.* [Online]

Available at: https://www.youtube.com/watch?v=otoXeVPhT7Q

Rapier, G., 2020. *A small Missouri city wants Tesla to build its Cybertruck factory there — and is offering $1 billion in incentives to make it happen.* [Online]

Available at: https://www.businessinsider.com/joplin-missouri-pitch-musk-tesla-cybertruck-factory-1-billion-incentives-2020-4?IR=T
[Accessed 21 May 2020].

Rathan, M., Hulipalled, V. R., Venugopal, K. & Patnaikc, L., 2018. Consumer insight mining: Aspect based Twitter opinion mining of mobile phone reviews. *Applied Soft Computing,* Volume 68, pp. 765-773.

Reed, E., 2020. *History of Tesla: Timeline and Facts.* [Online] Available at: https://www.thestreet.com/technology/history-of-tesla-15088992

Rinker, T., 2019. *Package 'sentimentr'.* [Online] Available at: https://cran.r-project.org/web/packages/sentimentr/sentimentr.pdf

Rokad, B., 2019. *Machine Learning Approaches and Its Applications.* [Online] Available at: https://medium.com/datadriveninvestor/machine-learning-approaches-and-its-applications-7bfbe782f4a8

Root, A., 2019. *Tesla's Pickup Truck Was a Stock Market Flop. How Investors Missed the Point..* [Online] Available at: https://www.barrons.com/articles/tesla-stock-cybertruck-launch-51574437479
[Accessed 30 April 2020].

Rouse, M., 2019. *sentiment analysis (opinion mining).* [Online] Available at: https://searchbusinessanalytics.techtarget.com/definition/opinion-mining-sentiment-mining

Saif, H., He, Y., Fernandez, M. & Alani, H., 2016. Contextual semantics for sentiment analysis of Twitter. *Information Processing & Management,* 52(1), pp. 5-19.

Sailunaz, K. & Alhajj, R., 2019. Emotion and sentiment analysis from Twitter text. *Journal of Computational Science,* Volume 36.

Salian, I., 2018. *SuperVize Me: What's the Difference Between Supervised, Unsupervised, Semi-Supervised and Reinforcement Learning?.* [Online] Available at: https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/

Sathees Kumar, B. & Karthika, R., 2014. A survey on text mining process and techniques. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET),* July, Volume 3, pp. 2279-2284.

Schreiber, B. A. & Gregersen, E., 2018. *Tesla, Inc..* [Online] Available at: https://www.britannica.com/topic/Tesla-Motors

Schubert, L., 2019. *Computational Linguistics.* [Online] Available at: https://plato.stanford.edu/archives/spr2019/entries/computational-linguistics/

SenticNet, n.d. *SenticNet.* [Online] Available at: https://sentic.net/
[Accessed 24 May 2020].

Serna, L., 2018. *Tesla SWOT Analysis As a Great Business Model.* [Online] Available at: https://www.luckscout.com/tesla-swot-analysis/

Shetty, B., 2018. *Natural Language Processing (NLP) for Machine Learning.* [Online] Available at: https://towardsdatascience.com/natural-language-processing-nlp-for-machine-learning-d44498845d5b

Shipley, L., 2020. How Tesla Sets Itself Apart. *Harvard Business Review* .

Silver, S., 2020. *Elon Musk Tweets He's Looking at "Central USA" Factory for Tesla Cybertruck.* [Online] Available at: https://nationalinterest.org/blog/buzz/elon-musk-tweets-hes-looking-central-usa-factory-tesla-cybertruck-133307

[Accessed 29 April 2020].

Smith, B., 2019. *Why Tesla's weird new Cybertruck could be a hit.* [Online] Available at: https://edition.cnn.com/2019/11/29/perspectives/cybertruck-tesla-elon-musk/index.html

[Accessed 30 April 2020].

Squatriglia, C., 2009. *Tesla's Movin' Up... To Palo Alto.* [Online] Available at: https://www.wired.com/2009/08/tesla-palo-alto/

Statista, 2020. *Pickup Trucks.* [Online] Available at: https://www.statista.com/outlook/1400000/109/pickup-trucks/united-states#market-revenue

[Accessed 1 May 2020].

Straubel, J. B., 2018. *Q1 2018 Financial Report* [Interview] 2018.

Stringham, E. P., Miller, J. K. & Clark, J., 2015. Overcoming Barriers To Entry In An Established Industry: Tesla Motors. *California Management Review ,* 57(4).

Sun, S., Luo, C. & Chen, J., 2017. A review of natural language processing techniques for opinion mining systems. *Information Fusion,* pp. 10-25.

Szymkowski, S., 2020. *Tesla Roadster delayed, Elon Musk says.* [Online] Available at: https://www.cnet.com/roadshow/news/elon-musk-tesla-roadster-delay-cybertruck-semi-joe-rogan/

[Accessed 21 May 2020].

Taboada, M. et al., 2011. Lexicon-Based Methods for Sentiment Analysis.

Talib, R., Hanif, M., Ayesha, S. & Fatima, F. §., 2016. Text Mining: Techniques, Application and Issues. *International Journal of Advanced Computer Science and Applications,* 7(11), pp. 414-418.

Taylor, C., 2017. *Unstructured Data.* [Online] Available at: https://www.datamation.com/big-data/unstructured-data.html

Taylor, C., 2018. *Structured vs. Unstructured Data.* [Online] Available at: https://www.datamation.com/big-data/structured-vs-unstructured-data.html

Tesla Motors, Inc., 2011. *Form S-1 Registration Statement.* [Online] Available at: https://www.sec.gov/Archives/edgar/data/1318605/000119312511149963/ds1.htm

[Accessed 2020 April 27].

Tesla, 2011. *Panasonic Enters into Supply Agreement with Tesla Motors to Supply Automotive-Grade Battery Cells.* [Online]

Available at: https://www.tesla.com/it_IT/blog/panasonic-enters-supply-agreement-tesla-motors-supply-automotivegrade-battery-c?redirect=no

Tesla, n.d. *About Tesla.* [Online]
Available at: https://www.tesla.com/about

Tesla, n.d. *Cybertruck.* [Online]
Available at: https://www.tesla.com/it_it/cybertruck
[Accessed 29 April 2020].

Tesla, n.d. *Tesla Gigafactory.* [Online]
Available at: https://www.tesla.com/gigafactory

Tesla, n.d. *Tesla Gigafactory 2.* [Online]
Available at: https://www.tesla.com/gigafactory2

The R Foundation, n.d. *The Comprehensive R Archive Network.* [Online]
Available at: https://cran.r-project.org/index.html
[Accessed 17 May 2020].

The R Project for Statistical Computing, n.d. *What is R?.* [Online]
Available at: https://www.r-project.org/about.html
[Accessed 17 May 2020].

Timoshenko, A. & Hauser, J. R., 2019. Identifying Customer Needs from User-Generated Content. *Marketing Science,* 38(1), pp. 1-20.

Tomar, A., 2018. *Topic modeling using Latent Dirichlet Allocation(LDA) and Gibbs Sampling explained!.* [Online]
Available at: https://medium.com/analytics-vidhya/topic-modeling-using-lda-and-gibbs-sampling-explained-49d49b3d1045
[Accessed 22 May 2020].

V12, 2016. *10 Social Media Strategies for Car Dealerships.* [Online]
Available at: https://v12data.com/blog/10-social-media-strategies-car-dealerships/

Valdes-Dapena, P., 2020. *The electric pickup wars are about to begin.* [Online]
Available at: https://edition.cnn.com/2020/02/14/cars/electric-pickup-truck-wars/index.html
[Accessed 1 May 2020].

Wagner, 2020. *Pickup trucks - U.S. sales 2015-2019 Published by I. Wagner, Jan 23, 2020 In 2019, a over three million pickup trucks were sold to customers in the United States, as sales figures continue to increase. The Ford F-Series remains the most popular pickup tr.* [Online]
Available at: https://www.statista.com/statistics/746742/number-of-new-pickup-sales-in-the-united-states/
[Accessed 1 May 2020].

Wagner, 2020. *Tesla - Statistics & Facts.* [Online]
Available at: https://www.statista.com/topics/2086/tesla/

Wagner, I., 2020. *U.S. new and used car sales 2000-2018 Published by I. Wagner, Apr 20, 2020 This statistic represents the number of new and used light vehicle sales in the United States from 2000 through 2018. In 2018, sales of used light vehicles in the United States ca.* [Online] Available at: https://www.statista.com/statistics/183713/value-of-us-passenger-cas-sales-and-leases-since-1990/

Warriner, A. B., Kuperman, V. & Brysbaert, M., 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behav Res.*

Wayland, M., 2019. *Tesla Cybertruck: 5 important questions about the polarizing EV.* [Online] Available at: https://www.cnbc.com/2019/11/22/tesla-cybertruck-5-important-questions-about-the-polarizing-ev.html

[Accessed 30 April 2020].

Wealth Inflator, 2018. *What Is Crowdtap? A Review.* [Online] Available at: https://wealthinflator.com/what-is-crowdtap-a-review

Webharvy, n.d. *What is Web Scraping ?.* [Online] Available at: https://www.webharvy.com/articles/what-is-web-scraping.html

[Accessed 19 May 2020].

Whissel, C., 2009. Using The Revised Dictionary Of Affect In Language To Quantify The Emotional Undertones Of Samples Of Natural Language. *Psychological Reports,* Volume 105, pp. 509-521.

Williams, M., 2019. *What Will the Pickup Truck World Think of the Tesla Cybertruck?.* [Online] Available at: https://www.motortrend.com/news/tesla-cybertruck-electric-pickup-reception/

[Accessed 1 May 2020].

Wittmeyer, A. P., 2012. *The FP Top 100 Global Thinkers.* [Online] Available at: https://foreignpolicy.com/2012/11/26/the-fp-top-100-global-thinkers-2/

[Accessed 29 April 2020].

Wonderflow, 2019. *How Consumer Insights Teams Can Leverage Text Mining Techniques.* [Online] Available at: https://www.wonderflow.co/blog/how-consumer-insights-teams-can-leverage-text-mining-techniques

Worldz, 2018. *WOM: il passaparola è più efficace dell'advertising.* [Online] Available at: https://tools.worldz.net/wom-passaparola/

Xu, K., Shaoyi Liao, S., Li, J. & Song, Y., 2011. Mining comparative opinions from customer reviews for Competitive Intelligence. *Decision Support Systems,* pp. 743-754.

Y.K.Lau, R., Li, C. & S.Y.Liao, S., 2014. Social analytics: Learning fuzzy product ontologies for aspect-oriented sentiment analysis. *Decision Support Systems,* Volume 65, pp. 80-94.

Yao, L., Mimno, D. & McCallum, A., 2009. Efficient Methods for Topic Model Inference on Streaming Document Collections. *KDD.*

Yao, R., 2018. *What Is The "API Economy" And How Brands Can Benefit From It.* [Online] Available at: https://medium.com/ipg-media-lab/what-is-the-api-economy-and-how-brands-can-benefit-

from-it-b46210d0434d
[Accessed 18 May 2020].
Yeomans, M., 2015. What Every Manager Should Know About Machine Learning. *Harvard Business Review.*
Young, C., 2019. *6 Pop Culture Inspirations for Tesla's Shiny New Cybertruck.* [Online] Available at: https://interestingengineering.com/6-pop-culture-inspirations-for-teslas-shiny-new-cybertruck [Accessed 30 April 2020].
Yse, D. L., 2019. *Your Guide to Natural Language Processing (NLP).* [Online] Available at: https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1
YS, T., 2018. *Creating brand value on social media the Elon Musk way.* [Online] Available at: https://yourstory.com/2018/01/elon-musk-social-media-brand-value-creation [Accessed 29 April 2020].
Zhai, C. & Massung, S., 2016. *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining.* Morgan & Claypool: s.n.
Zhang, W., Xu, M. & Jiang, Q., 2018. Opinion Mining and Sentiment Analysis in Social Media: Challenges and Applications. In: *Lecture Notes in Computer Science.* s.l.:s.n.
Zhao, Y., 2013. *Analysing Twitter Data with Text Mining and Social Network Analysis.* s.l.:s.n.

Executive Summary

Introduction

Sharing opinions is an activity that defines social life of humans. People express thoughts to communicate their feelings about events, other people and objects that surround them. These can assume the shape of different emotions, according to how the input impacted and was elaborated by the subject. Thus, an opinion can be seen as a form of elaboration of taste made by a subject towards an object to which he found himself in contact with. Once an opinion is formed, sharing is a subsequent step that allows people to be in contact and thus giving sense to human as a social living being. Thus, expressing an opinion is sometimes a way to express oneself, to tray relating to people in the social fabric. The evolution that sharing have had in modern era is unprecedented, and the main point of this progress is given by the exponential increase of the audience that can elaborate a certain message expressed by one subject. In other terms, nowadays one sender has millions and millions of receivers: if this is multiplied by each subject expressing a thought, the number generated would be just huge. The tool that mostly influenced the change in people's communication and society in general is Internet, particularly in the form of social media. These are actual platforms that were built for sharing a multitude of information. Everything is this media is focused on allowing each user to share contents, from posts to comments. In some case and for some entities, an opinion can represent a source of

value, as it is inserted into a process where its role is giving directions towards the achievement of a certain output. This is the case in which an opinion turns into a feedback. In an economical sense, the subject turns into a potential customer and the feedback is directed to the evaluation of a product or service. This information is then received by the company that provided these elements and can used to understand the reactions towards them. Naturally, this system is prior to the advent of the social network era, but the changes brought by them are what made the real difference. In particular, two factors determined the revolution: the way the information is amplified and the way it is shared unfiltered. The former allows to think in terms of quantity, the latter in terms of quality. The amplification is what makes the process big and so why the businesses are more and more prone to analyse bigger and bigger data: for some aspects, in terms of opinions, the user needs to be thought as a cross-cultural and cross-national entity nowadays, with all its pros, cons and distinctions. In the era of social network domains, the communication is potentially worldwide, especially for big and renewed companies. Moreover, there is no "business filtering" when expressing feedbacks: in social networks, in most of the cases, users are not pushed by companies to express their opinions, neither are participating to an experiment: what they share is somewhat as close as possible to what they actually think about the object of discussion. These are what make social media information such valuable, in quantitative and qualitative terms, for businesses. That is also why the attention towards their analysis and so the implementation of specific computational developments is more and more increasing.

In particular, some programming tools were already in place that analysed big amount of information in form of comments and opinions, in particular by understanding the meaning of a word or sentence in a scalable way. However, these new kinds of sources have pushed the programs to rethink some aspects, especially in terms of language expressions, in order to adapt the system to the reception of the above described value. In particular, a form of computational analysis called text mining, that will be further described, have been adapted to the new and valuable source of textual information namely social media, leading to the creation of a new field of study whose name is social media analysis. Therefore, when the aim is to capture the feelings that lay behind a text, the focus of the mining shift towards the study of a specific emotion expressed by the user: this is when sentiment analysis comes into play. These aspects are then applied to a specific business or product, thus giving a structure and business value to the entire process of analysis. Ultimately, this is what the following study is about. The first chapter focuses on identifying the main elements and factors that characterize text mining, narrowing the field of research to the sentiment analysis and opinion mining in social media contexts like Twitter. The second chapter presents the product to which the analysis is addressed, namely the Tesla Cybertruck, an electric pickup presented by Tesla Inc. on November 21st, 2019, that polarized people, critics and fans for its peculiar shape and features, together with its controversial unveiling. The third chapter applies some of sentiment analysis techniques to the Cybertruck, with the findings that were divided into four round and take into account different aspects and levels of study.

Elements of Text Mining

The study and analysis of written text allows to extract important knowledge that can be used to derive meaningful findings for businesses and researches. To achieve this goal, different approaches can be implemented, but they should all share some common steps that are fundamental in order to select relevant contents and discard non useful information. The chapter firstly highlights the importance that these types of information, known as unstructured data, are gaining throughout the years, considering also the vast development of different platforms where users, companies and brands in general can share knowledges through text. Moreover, some crucial aspects of the process of the extraction of information are presented, focusing on the main steps of analysis from data collection to final visualization. In addition, examples of algorithms and approaches, that are taken from machine learning fields and dictionary-based ones, are shown in a wider focus on Natural Language Processing, by this defining the ways through which human language is actually understood and then processed by machines and computers. One of the main scopes of the main analysis of text mining, namely opinion mining or sentiment analysis, that deals with capturing the actual emotion and thoughts about different topics, is discussed. The focus then shifts to the discussion over particular domain for text mining studies, namely social network. These in fact represents one of the abovementioned platforms where user generated contents can be usefully and meaningfully levered, if not the main one, as its growth in popularity and subscriptions rose exponentially in the last decade. In them, a huge amount of written information is posted and shared every day by people and potential customers, thus explaining its importance for this field of study. Although there are different social media sources, one of the main relevant for text mining in undoubtedly Twitter, since its core functionality is strongly linked to the diffusion posts and comments in the written, that in this domain are also knowns as “tweets”.

Object of study: product characteristics

A sentiment analysis needs to be directed on an entity, namely the object over which the emotion or sentiment is expressed (Sun, et al., 2017). Moreover, one of the main areas of research in Twitter Sentiment Analysis or TSA is product reviews (Pandarachalil, et al., 2015). So, the object of this study is introduced in the present section and comes from the automotive sector. Along with the reasons why Twitter is a valuable source of information for brands and consumers’ purchasing process (Lin, 2019) (Elrhoul, 2015), there are also additional evidences highlighting the importance that social media cover particularly in cars and vehicles purchase decisions. The product at issue is a vehicle that has made the news for its distinctive design and its controversial unveiling. This is the electric pickup truck of the American company Tesla, called Cybertruck.

The market-edged nature of Cybertruck

At a first glance of the model at issue, it is straightforward to understand why it is considered a unique product. For this purpose, the following is an image of the Cybertruck:

Figure 1: Tesla Cybertruck

Retrieved from *businessinsider.com,* 2020

What firstly captures the attention and makes a strong visual impact is undoubtedly the peculiar design of the car, where angular and edgy shapes make it not only easily distinguishable, but also to be perceived as a futuristic and in a way more focused on essentiality and functionality, making it look resistant and captivating at the same time. The shapes of the Cybertruck contributed to make a breach in pop culture, since many people saw in the vehicle some references of past futuristic cars that appeared in cult movies: for instance, Elon Musk itself tweeted that the "*Cybertruck design influenced partly by The Spy Who Loved Me"* (Musk, 2019), a 1977 James Bond movie; in addition, others saw resemblances with model cars from cult movies like Blade Runner, Mad Max: Fury Road and Akira, and even from the videogame Cyberpunk 2077 (Young, 2019). Another relevant comparison was the one that saw the Tesla pick-up truck to the 1981-1983 DMC DeLorean, a vehicle made iconic by the movie "Back to The Future".

Tesla Cybertruck is thus defined as a full-fledged polarizing car model, as some people were enthusiastic about its futuristic appearance and technological-advanced software and features while other were really sceptical about both the vehicle, which seemed uncomplete and so looked more like a prototype, and its impact on the American pickup market. Nevertheless, in order to get a more complete overview of the scenario, it also has to be considered the business strategy of the Cybertruck and the segment Tesla is targeting with the pickup. It is undoubtedly true that this vehicle wants to represent a disruptive element in its market, as Elon Musk himself stated during the unveiling: *"Trucks have been the same for a very long time […]. Like a hundred years, trucks have been basically the same. We want to try something different."* (Musk, 2019). However, that truck proposition can be seen as a starting point for a broader strategy of Tesla to enter the market, starting with a

product whose objective is not be targeted for mass-market, but rather for a niche of enthusiasts and eventually curious consumers who want to change its choices and persevere in the attention towards the environment without renouncing to the standard performances and functions of a truck. An analysis (Smith, 2019) for CNN highlights this point: what it is stated here is that for Tesla to enter in the pickup market, a lookalike product would not have been noteworthy plus not reflecting the business style of Tesla and its CEO. Rather, the proposition of a distinctive product would help to reach a niche a target of 5% to 10% of potential pickup buyers, which can be still a profitable percentage considering the size of the market. Once having established with the vehicle, the next step would be to build credibility, that could then lead to the proposition of another pickup model that will try to capture a higher user base and potential buyers. This was in a way confirmed by Musk during an interview, where he admitted that the Cybertruck could represent part of a bigger strategy in the sector: *"You know, I actually don't know if a lot of people will buy this pickup truck or not, but I don't care […]. If there's only a small number of people that like that truck, I guess we'll make a more conventional truck in the future."* (Musk, 2018). In other words, Tesla seems to re-propose its general strategy in the reduced scale of the pickup market: start as a niche, then try enlarging the market to target the mass with a more standard product, all while converting the network into the electrification and the adoption and installation of "in-car" software.

Application of Sentiment Analysis on product

Once having introduced both the characteristics of text mining and the product, the following section will aim at applying some of the features of sentiment analysis to Tesla Cybertruck, in order to assess the impact that this car had on the audience at an emotional and opinion level in the domain of one of the most important and most-used social media that is Twitter. Moreover, an overtime evaluation and evolution would be implemented to see how the focus of the audience shifted across different themes, that changed according to the eventual happenings and events that surround the product and its company. In fact, the analysis will also identify major elements, like relevant figures, tweets, strategic company decisions and general contingencies, that in a way were and are still associated to the Cybertruck. That step goes beyond a mere sentiment analysis, thus allowing to capture a useful and larger range of information and eventually to explain the reasons behind certain audience responses. Moreover, the analysis provides a combination of approaches to inspect the features on a deeper level and so to retain pieces of information that come from different sources.

Furthermore, understanding the audience's response about a changing trend is also an important aspect to take into account, as it goes beyond a single company's strategy and it is part of the elements that surround the Cybertruck. In addition, the social media dynamism of Tesla and its main figures like Elon Musk contribute to provide a great amount of information to be analysed, representing a factor that not only differentiate them from competitors, but also make them more easily searchable and in a way analysable, thus making a relevant impact on the available sources for this study. In other words, the more the information are made, posted and shared about an object, the more are the data available for the analysis, especially for text mining where large data sets conducts to large quantitative results and the possible application and more appropriate approach

depend on the type of information available (MonkeyLearn, 2019). Having stated these premises, the following sections will go through the steps implemented for the sentiment analysis and the detection of some relevant topics, from data collection on Twitter to the visualization and inspection of findings, both considering the data as a general entity and splitting them according to specific reference periods.

Findings

The study conduces to diverse findings that vary basing on the method and time range chosen. In detail, the findings are structured in this way:

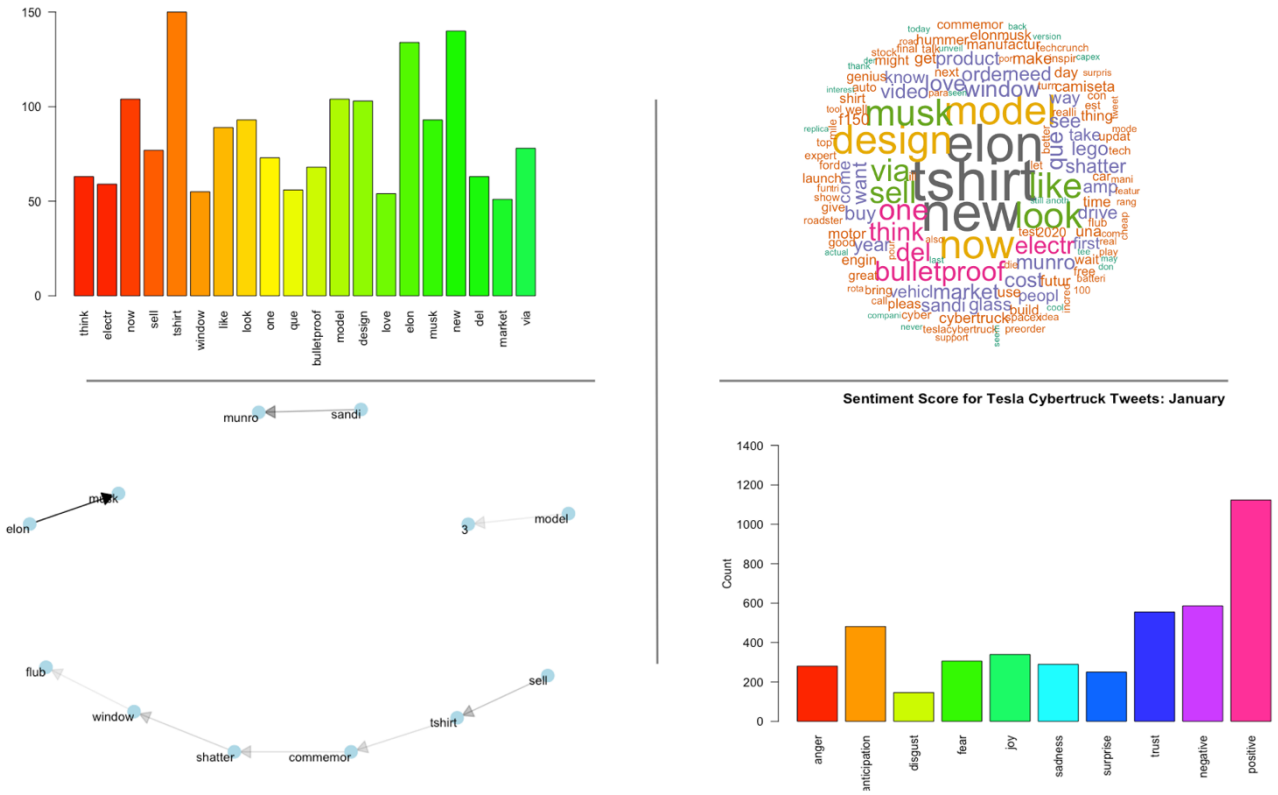
1. Findings of round 1: A broaden inspection of the dataset mainly based on general visualization tools. This round was repeated for each monthly dataset of Cybertruck and it is useful to have a first glance of the word present in tweets and thus individuate potential point of discussion. Moreover, still starting from the monthly datasets, a subsequent phase deepens the comprehension of the words, as it examines the relationships that can occur between them through the decomposition of tweets into the most frequent bigrams. Again, this can be a good way of inspecting the dataset as it goes beyond a single word inspection and thus captures more context. A month's lexicon-based sentiment analysis is then conducted.
2. Findings of round 2: This is the round in which an unsupervised learning method taken from NLP field is implemented: Latent Dirichlet Allocation or LDA. This one is used to clearly find the most relevant topics in the tweets and further deepens the level of the analysis. This phase is conducted for the entire dataset of Cybertruck tweets, without monthly division. After having completed it, a set of topics were given back: their content is then evaluated to manually extract topics that are judged to be more relevant. Finally, a singular sentiment analysis is conducted over the arguments chosen. This round allows to analyse the single features and drivers of the sentiment, thus going beyond document-level to inspect the score of precise characteristics that surround Cybertruck.
3. Findings of round 3: The final round that was entirely conducted on just Cybertruck dataset is the one that implemented the special dictionary that detects particular elements of Twitter's language, like emoticons, emoji and slangs, together with taking into account negators, amplifiers, question weights and adversatives. A numerical value range that went from -2.5 to 2.5 was then assigned to each tweet and graphs and other visualization tools are used to show the final score of the Cybertruck dataset. This phase is conducted by taking the data frame of Cybertruck in its entirety.
4. Findings round 4: The very last round is the one that compares Cybertruck to its abovementioned competitors by levering a special R package that allows to specifically conduct a competitive sentiment analysis of Twitter data. Again, the data frames were used with no monthly division.

Findings of round 1

There were retained the words that were cited more than 50 times or 100 if the bar plot resulted too crowded. In addition, a specific kind of image was generated from that dataset: this is the word cloud, a type of visualization that shows text data organized in a shape of a cloud, whose size varies based on words' recurrence

on the dataset: the higher the frequency of the word, the bigger is it scape occupied in the cloud. In other words, *"a word cloud is a collection, or cluster, of words depicted in different sizes. The bigger and bolder the word appears, the more often it's mentioned within a given text and the more important it is"* (Boost Labs, 2014). The command for word cloud allowed to show words with a minimum frequency of 5 and a maximum number of showable words equal to 150. The other set of findings focused examined how the word are linked to each other: this was made possible after having tokenized the text and inspected the frequent bigrams, namely words that occur together the most. Specifically, only the bigrams with a frequency higher than 20 were retained. In some cases, the bigrams were even linked to each other, building a path that resembles a sentence structure. The final monthly analysis conducted was a sentiment analysis based on the lexicon called NRC Word-Emotion Association Lexicon, or EmoLex (Mohammad & Turney, 2010) (Mohammad & Turney, 2013), that is a dictionary composed by 14,182 unigrams divided in eight basic emotions, namely anger, fear, anticipation, trust, surprise, sadness, joy and disgust, and two sentiments, namely negative and positive. The classification of EmoLex was manually conducted through crowdsourcing in Mechanical Turk platform (Mohammad, s.d.). The dictionary is retrievable on R by using the package "syuzhet" (Jockers, 2017).
The following plots were the results of this first monthly analysis and were all retrieved from the R Studio code:

Figure 2: Cybertruck Word Frequency, Word cloud, Bigram Relationship and Sentiment Analysis, January 2020

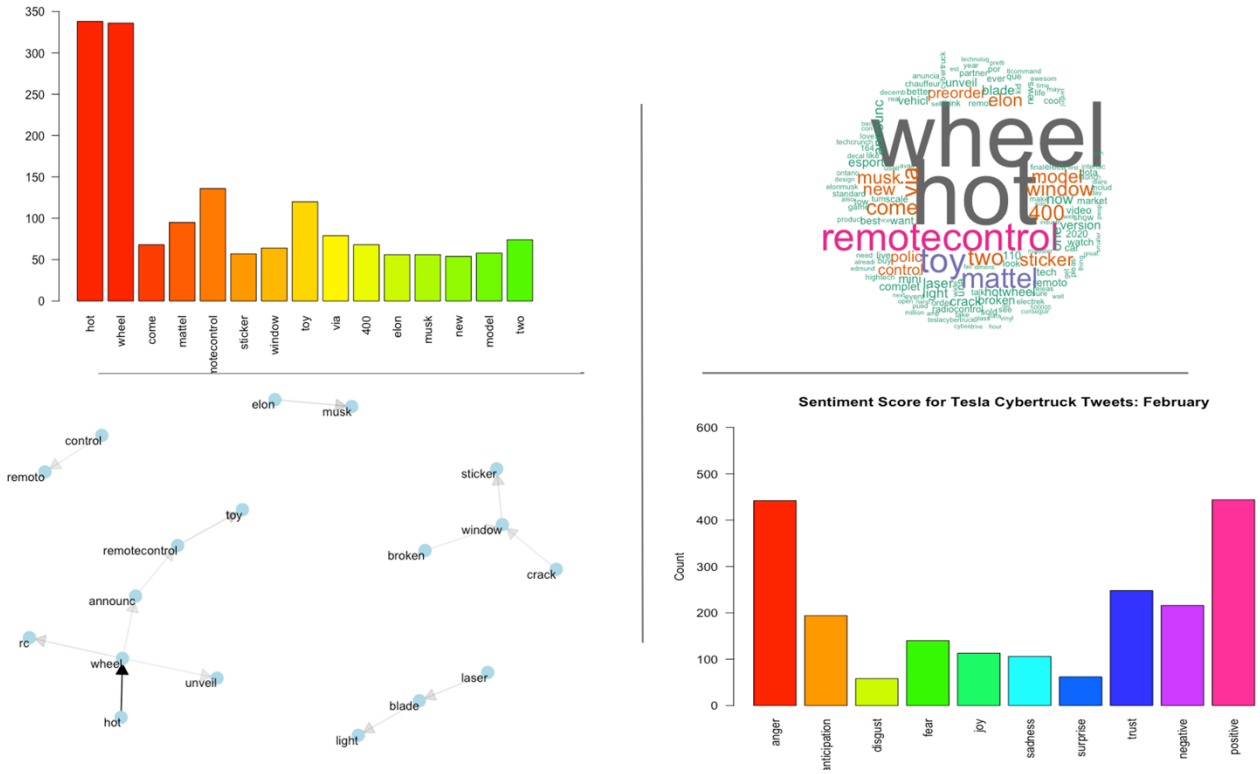


Retrieved from *R Studio*

January word plots highlights some relevant topics: firstly, it is possible to still find referrals to the shatter of Cybertruck's window happened during the pickup's unveiling on November 21st, 2019. In particular, the

words "window" and "bulletproof" are clear referrals, but also "tshirt" is one of them. In fact, after the episode of the broken glass Musk started the selling of t-shirt depicting the broken window. Moreover, an expected mention to the CEO Elon Musk himself are in place: this is a particular recurrent referral in each part of the study, as Musk is inextricably linked to its company and their product, Cybertruck included. In addition, in the word cloud it is possible to find even more details: a proof of that is the word "shatter". Other curious information retrieved in the cloud arose, for instance, from the word "lego", in the upper mid right section of the circle: in fact, after making some researches, it was found that a group of Tesla enthusiasts submitted a concept for a Lego Cybertruck to Lego Ideas, the site that collects new and original ideas for the creation and commercialization of new Lego pieces. This sort of request was sustained by about 10,000 people, pushing Lego to start the project review process, beginning on May 2020 (Paukert, 2020). From what concerns the bigram relationship, these confirm the information retrieved in the figures before, adding another frequent bigram: "sandi munro". This refers to the CEO of Munro & Associates Sandy Munro, whose company's core business is *"helping companies reduce "time to market", R&D, engineering and manufacturing costs all while increasing the quality of our customers products, processes and systems".* (Munro & Associates, Inc., s.d.). Munro provided a review of the Cybertruck after having accurately analysed it from an engineering perspective, even receiving a positive feedback from Elon Musk (Munro & Associates, Inc., 2020). Lastly, the sentiment analysis conducted through NRC showed a stronger presence of positive words compared to the negative ones. For what concerns the emotion analysis, trust and anticipation were the dominant ones, probably meaning that, since the attention for the Cybertruck was kept high, users are still impatient to know more about the product and have in a way forgiven the mistakes made during the unveiling.

Figure 3: Cybertruck Word Frequency, Word cloud, Bigram Relationship and Sentiment Analysis, February 2020



Retrieved from *R Studio*

The analysis for February clearly highlights a word trend and most discussed topic: this is about the collaboration that Tesla made with Mattel's Hot Wheels for the realization and commercialization of a remote-controlled Tesla Cybertruck in both 1:64 scale, at the price of $20, and 1:10 scale, that would cost about $400. The ship is set to start on December 2020 in limited edition (O'Kane, 2020). The only adding that can be retrieved from the bigram relationship graph is the one that concerns the laser blade light feature on Cybertruck, announced, as usual, by Elon Musk in his Twitter account (Lambert, 2020). For what concerns the sentiment analysis, an unexpected outcome is shown in the high frequency for the "anger" emotion: however, this can be clearly justified by the fact that it is a misclassification of the dictionary, that recognizes the word "hot" as referring to the "anger" category, while, as abovementioned, it is just associated to the Mattel's Hot Wheels brand. This case clearly explains the reason why it is necessary to go beyond a mere word classification by inspecting how most relevant words are used and in which context. In this case, the combination of computational analysis and human observation was useful to detect this factor.

Figure 4: Cybertruck Word Frequency, Word cloud, Bigram Relationship and Sentiment Analysis, March 2020

Retrieved from *R Studio*

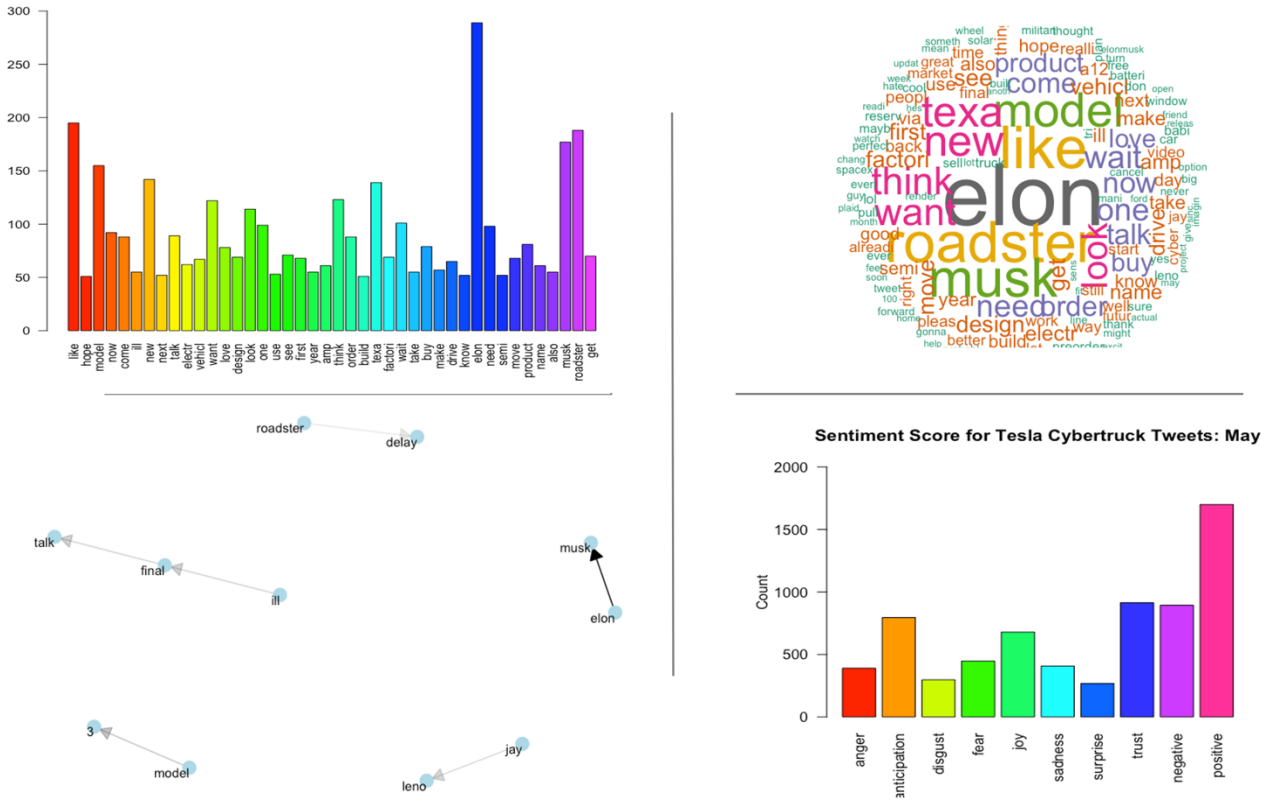
One of the main topics of the month the one that concerned the choice of the location for a new Gigafactory in central USA. That was the content of Musk's tweet in March, adding that the new factory will produce both Cybertruck and Model Y for the East USA market (Levy, 2020). Not surprisingly, the news created a buzz in the Internet, with a real positive response by the users as shown in the NRC analysis. In addition to having positive attitude towards Cybertruck tweets in March, they also showed a high level of anticipation, probably linked to the suspense derived from the decision of the location for the new Gigafactory.

Figure 5: Cybertruck Word Frequency, Word cloud, Bigram Relationship and Sentiment Analysis, April 2020

Retrieved from *R Studio*

For what concerns April, no big changes needed to be reported compared to what happened in the previous month: the sentiment analysis trend, with positive reactions and great anticipation arguably due to the Gigafactory topic, was confirmed, with only changes in frequency and thus proportions. News can be retrieved from the bigram graph: the fact that Joplin, a small city in Missouri, offered Tesla $1 billion in tax incentives for constructing the Gigafactory there (Rapier, 2020). In addition, in April, the new season of "Jay Leno's Garage", a known American web and television series about motors and even new model's test drives, was announced. This show will include a special and really awaited episode on Tesla Cybertruck (Lambert, 2020).

Figure 6: Cybertruck Word Frequency, Word cloud, Bigram Relationship and Sentiment Analysis, May 2020



Retrieved from *R Studio*

May did not produced that much additional information: sentiment trend is still the same, with positives overtaking negatives; some of the main topics are still linked to Jay Leno's Garage and Gigafactory. For the latter, the May news can be the one that, according to reliable sources, Tesla had decided to set the Gigafactory location in Austin, Texas (Lambert, 2020). Finally, a relevant news is the announcement of the delay for the new Tesla Roadster, that will be produced in 2022 and after Cybertruck (Szymkowski, 2020). This news can explain the lowering in the emotion of anticipation that, despite still higher in frequency than the others, is considerably lower than the last month.

Findings of round 2

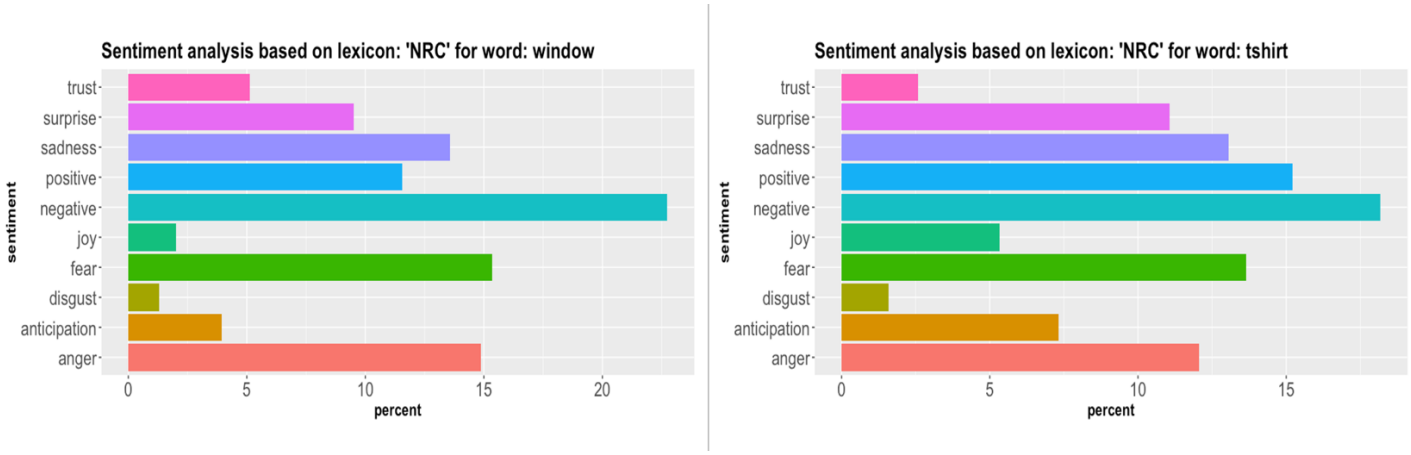
This set of findings is focused on extracting the most relevant topics from tweets by using the unsupervised learning technique of Latent Dirichlet Allocation or LDA, that is a generative probabilistic model that can allow to assign a topic to a specific word, knowing the distribution of topics in the documents. The generative element refers to the fact that the new samples are generated from the same distribution and ultimately from the same data frame (Blei, et al., 2003) (Manwani, 2018). Furthermore, a prerequisite of LDA is that each document has to share the same topic, whose number is pre-specified, but distribution differs. In this way, the process ultimately assigns a certain score to the words and allocates them into the different sets of topics created. That is a useful way to group similar texts and identify relevant groups of words that can be analysed from a sentiment point of view afterwards. This is how LDA is inserted in this study. It was essentially used as a first step to extract topics and words that compose them, with the following step being their analysis of

sentiment. In other words, this phase merges LDA and sentiment analysis to understand the impact of specific arguments of discussion in the sentiment of tweets, to produce a deeper finding for the entire study.

From a computational point of view, the methodology used for this phase is the same as the first round, but the Cybertruck tweets were unified, thus there is no month distinction. The differences started when the package "topicmodels" was implemented (Grün, et al., 2020), that allows to compute probabilistic models based on the data structure obtained with the package "tm". However, an additional point needs to be addressed: in fact, here a precise method for LDA is specified, which is Gibbs sampling. As seen, Latent Dirichlet Allocation allows to assign a distribution of topic in each document and a distribution of words in topics: what Gibbs does is to try optimizing the conditional distribution of these variables by repeating the classification over time. Since the allocation to the exact topic is a probability matter, there is a higher chance for the word to be classified correctly, but no certainty. Gibbs maximizes the likelihood of a good classification by levering on the conditional probability distribution of a word's topic assignment that is conditioned by the rest of the topic assignments. In other words, to get the topic *k* of a word *w* that belongs to a document *d*, LDA with Gibbs method analyses, among other factors, the number of times that document *d* used topic *k*, *n(d,k)*, and the number of times topic *k* used the word *w*, *n(k,w)*. A Dirichlet parameter allows to still derive an assignment if these two values are equal to 0. Finally, as mentioned, to better perform this probability, a higher number of trials contribute to more precise results (Tomar, 2018) (Grün & Hornik, 2011).

The words were selected manually by looking at which of them are of most valuable interest. For instance, it can be useful to understand the sentiment that is linked to the word "Elon", to assess the impact that the CEO had on Twitter users in the time range considered. The words "tshirt", "window", "gigafactori", "electr" that is the stem form for "electric", "design", "elon" and "order" were closely analysed. The results are shown below:

Figure 7: NRC sentiment analysis for "tshirt" and "window"

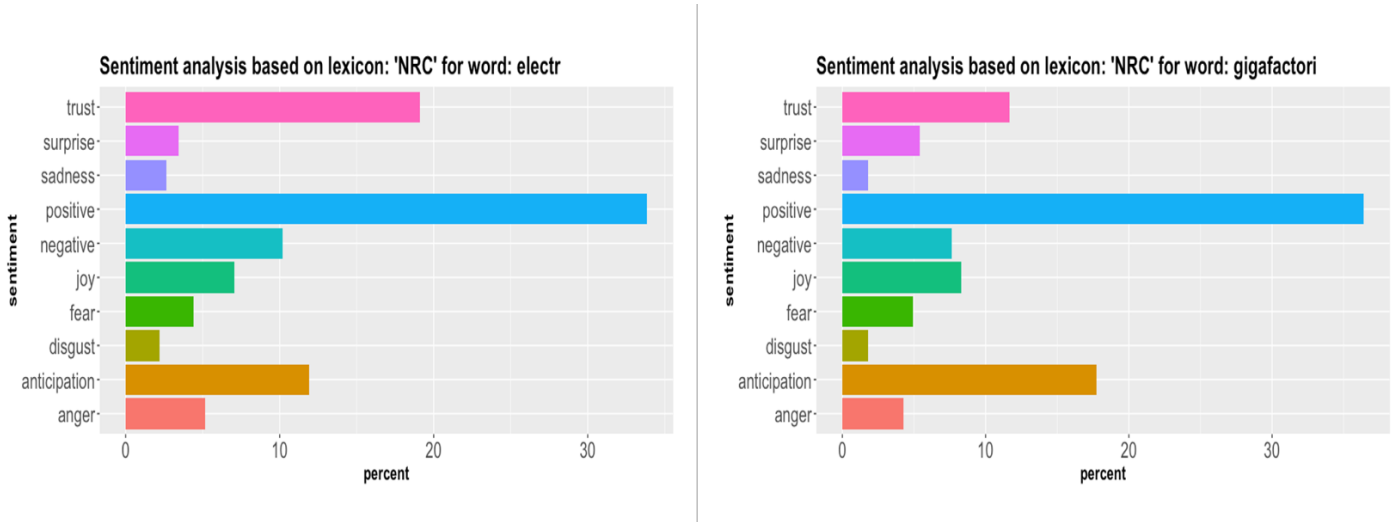


Retrieved from *R Studio*

The first two sentiment analysis were conducted on two of the set of tweets that were identified by LDA and refer to the episode of the unveiling. As shown, the predominant sentiment association is negative, with

emotions like “anger”, “fear” and “sadness” that represents the highest percentages. However, when looking at the “tshirt” sample, it is possible to notice that the value for “positive” and “surprise” are higher, probably meaning that the move made by Elon Musk to lever on emotional intelligence and not hiding the mistake resulted to be a successful one. Despite that, overall, the results showed a tendency to bad associations, so from a brand perception and reputation point of view, Tesla paid its mistake.

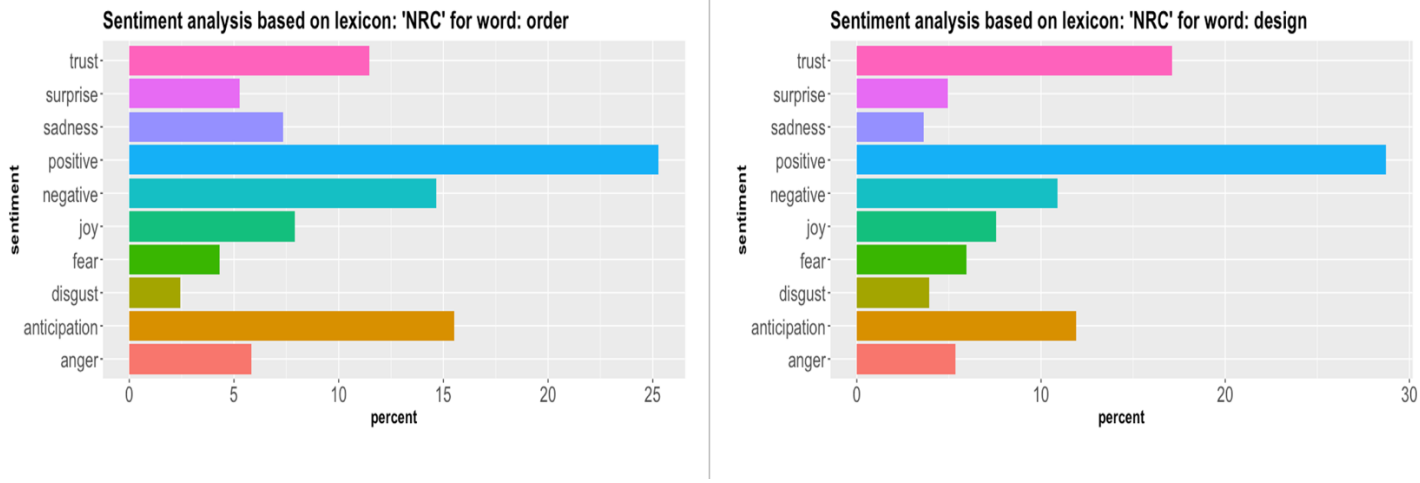
Figure 8: NRC sentiment analysis for “electr” and “gigafactori”



Retrieved from *R Studio*

The two plots above told a different story, as the values for two of the main elements of not only Cybertruck, but Tesla in general, were positively perceived by Twitter users. Some remarkable outcomes arose from the trust perceived for the electric automotive sector, that represents the highest emotion for the set, and the positive sentiment reaction overall. It was already mentioned the increased importance that electric vehicles are having in nowadays society, so this outcome represents a confirmation in terms of perception of this trend. For what concerns the Gigafactory set, high values in “anticipation” and “positive” can be explained by the great communication made by Elon Musk around the decision of building a new factory, that made Tesla fans more enthusiast overall.

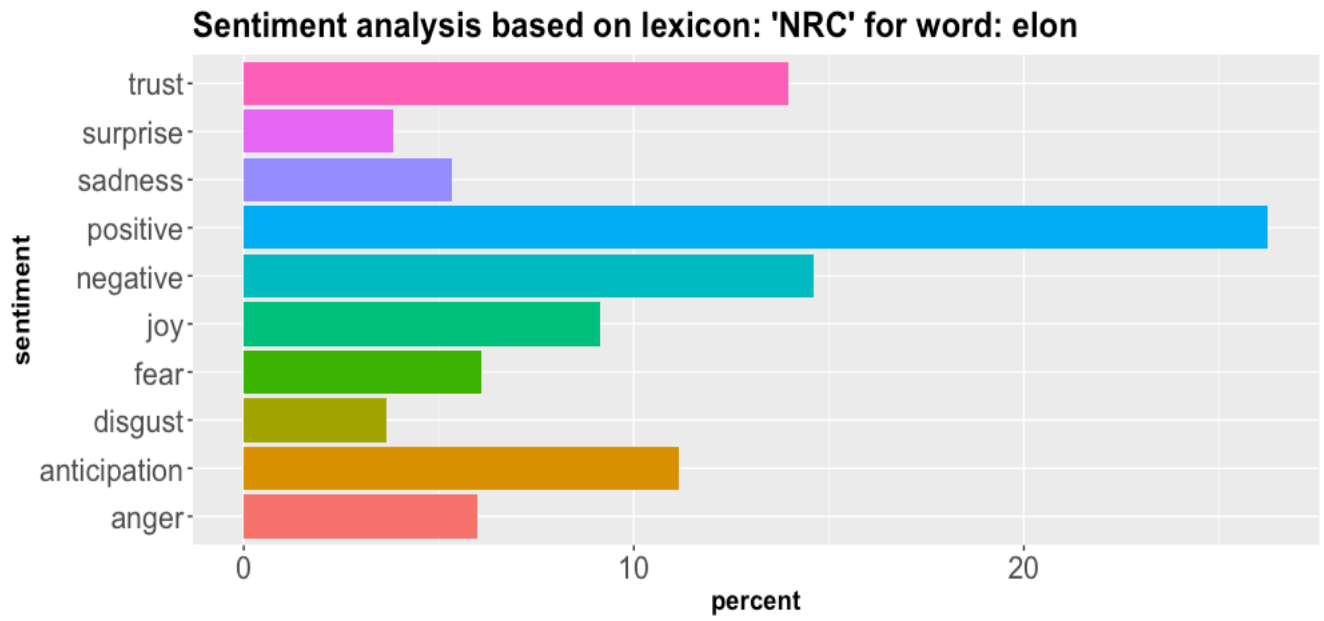
Figure 9: NRC sentiment analysis for “order” and “design”



Retrieved from *R Studio*

These two plots are probably the ones that are linked the most to Cybertruck as a product, in particular the set of "design". The images show a similar trend, with "positive" representing the highest sentiment and "trust" and "anticipation" as highest emotion values in percentage. The order set presents a slightly higher bad values: that can be linked to the uncertainty that sometimes surround Tesla production, with Cybertruck not making an exception. However, the good perceptions are higher with a consistent differential.

Figure 10: NRC sentiment analysis for "elon"



Retrieved from *R Studio*

Last but not least, the plot above shows the perception of tweets containing Elon Musk. The CEO confirms to be a relevant figure from a perception side too, as positive sentiment was higher than negative, and "trust" was the leading emotion. Despite the critiques that Musk had received for its controversial behaviours, he still remains the main person man that drives Tesla, Inc. and his approach and preparation had made him to be

perceived as one of the most influential person not only in the automotive industry, but in the business worldwide.

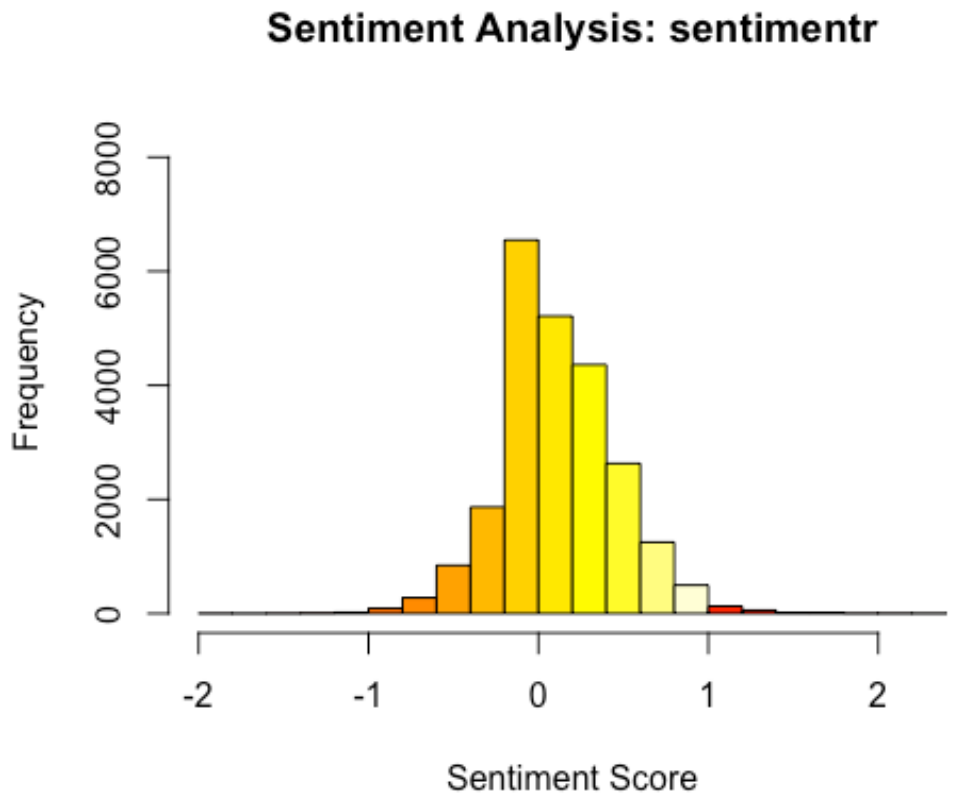
That was the analysis of tweets subset identified by LDA. Moving on to the next round, the most relevant changes concerns the different lexicon chosen for the sentiment analysis.

Findings of round 3

The third set of findings for this study are focused in the application of specific functions retrieved from R package "sentimentr" (Rinker, 2019). The application of this package was made with the intention of deepening the results of this sentiment analysis study. In fact, the implementation of "sentimentr" allowed to include, in the calculation of the sentiment score, the context surrounding a polarized word, like negators, amplifiers or de-amplifiers, adversatives, emojis, emoticons and slang language: in other words, the package allows to compute a sentiment analysis at a sentence-level.

These functions were applied only to Tesla Cybertruck data frame with no differences in months. The following is the histogram that allows to visualize the outcomes:

Figure 11: Cybertruck sentiment analysis results using the package "sentimentr" in numbers



Retrieved from R Studio

As expected, a lot of the tweets were close to the 0 value, meaning that neutrality is predominant. However, moving away from the central zero point, it is possible to notice that the bar plots are taller in the positive side,

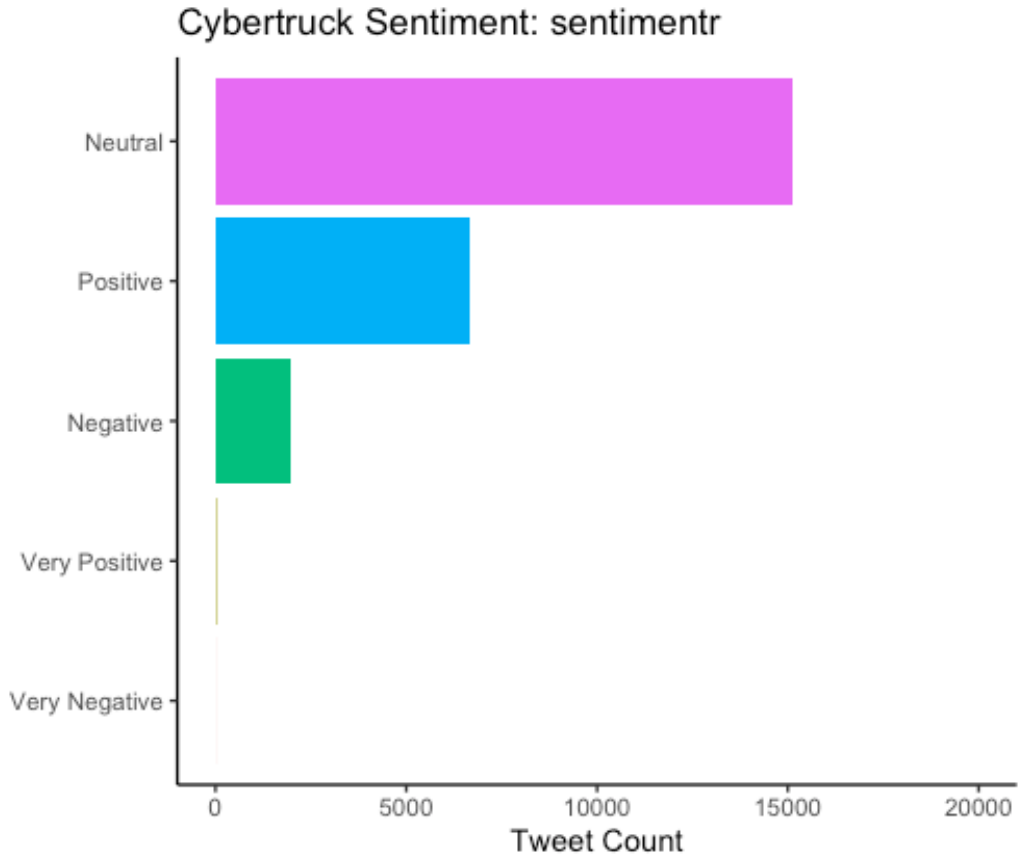
namely the one that is composed by values higher than 0. This can give an idea of the fact that the perception and sentiment associated to Cybertruck tweets is more positive than negative, even taking all the above-mentioned factors into account.

To have a clearer vision of that, another visualization tool was implemented by manually convert numeric values into ranges that went from "very negative" to "very positive" in the following order:

1. -2.5 to -1.51: "very negative"
2. -1.50 to -0.31: "negative"
3. -0.30 to 0.3: "neutral"
4. 0.31 to 1.5: "positive"
5. 1.51 to 2.5: "very positive"

The following is the graph obtained:

Figure 12: Cybertruck sentiment analysis results using the package "sentimentr" in value ranges



Retrieved from R Studio

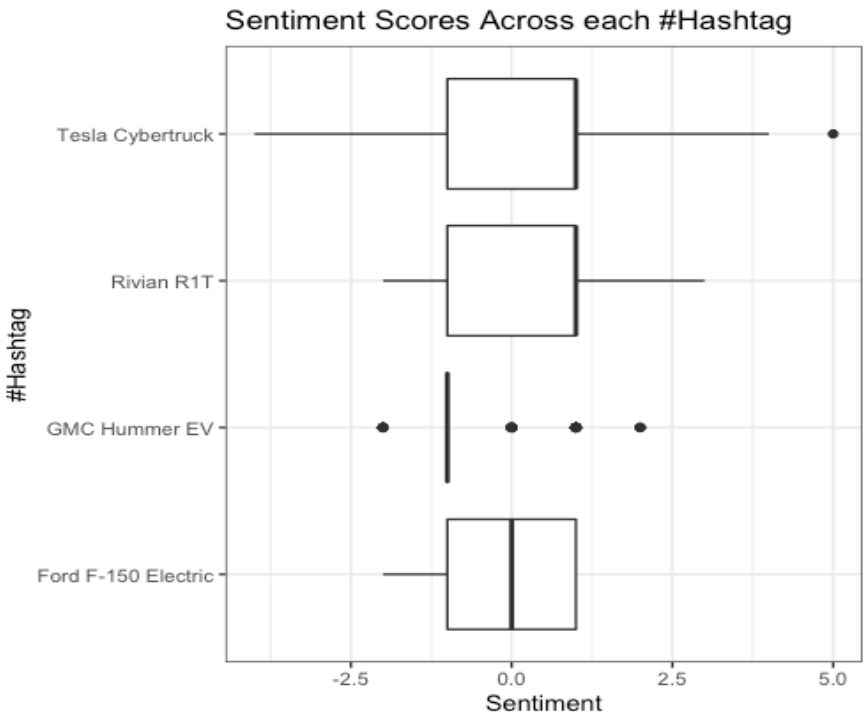
This division was made for the sake of a better visualization of the results already noticed from the previous plot. In addition, the ranges were enlarged due to the effect of the above-mentioned calculations made by the system. It is now clear that the number of positive tweets were higher than the negatives, despite neutrals having the highest frequency. These results confirmed what already found before, that is the feedbacks for Cybertruck and Tesla in general are more positive than negative. This is due to several factors like the uniqueness of the product, the organization and mission of Tesla and the excitement arising from the CEO

Elon Musk and his initiatives like the constructions of the huge Gigafactory. This model recalls to more than a pickup vehicle, making it a product that leaves space to imagination and wondering.
A final sentiment study for Tesla Cybertruck is the one that compares it to some of its main competitor in the sample's time range and it is presented in the following round.

Findings of round 4

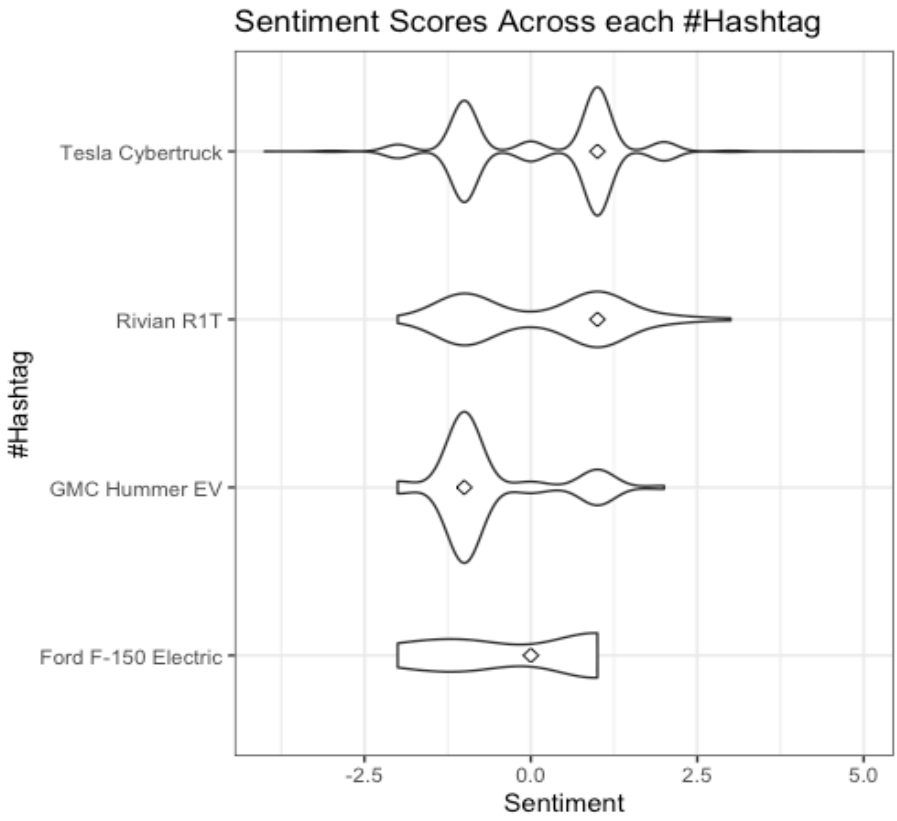
The final findings took into account Tesla Cybertruck with some of its competitors, namely Rivian R1T, Ford F-150 Electric and GMC Hummer EV. To process this phase of the analysis, another special R package was used and it is called "saotd" (Munson, et al., 2019), an acronym that stands for Sentiment Analysis of Twitter Data. Similar to other packages, "saotd" allows to compute a complete sentiment analysis, from the acquisition of data to visualization tools, by using the Bing dictionary. However, one of its main application is the comparison of tweets with different "hashtags", that in this case represent different brands. In fact, to work properly, a creation of a dedicated column called "hashtag" is necessary to identify the subject of each tweets. That is why, in a new data frame appositely created for this study, tweets of different electric pickups were divided by their brand in the column "hashtag".
After having cleaned the data, the following outcomes were retrieved (CRAN, 2019):

Figure 13: Box plot of sentiment score for different EVs brands using "saotd"



Retrieved from R Studio

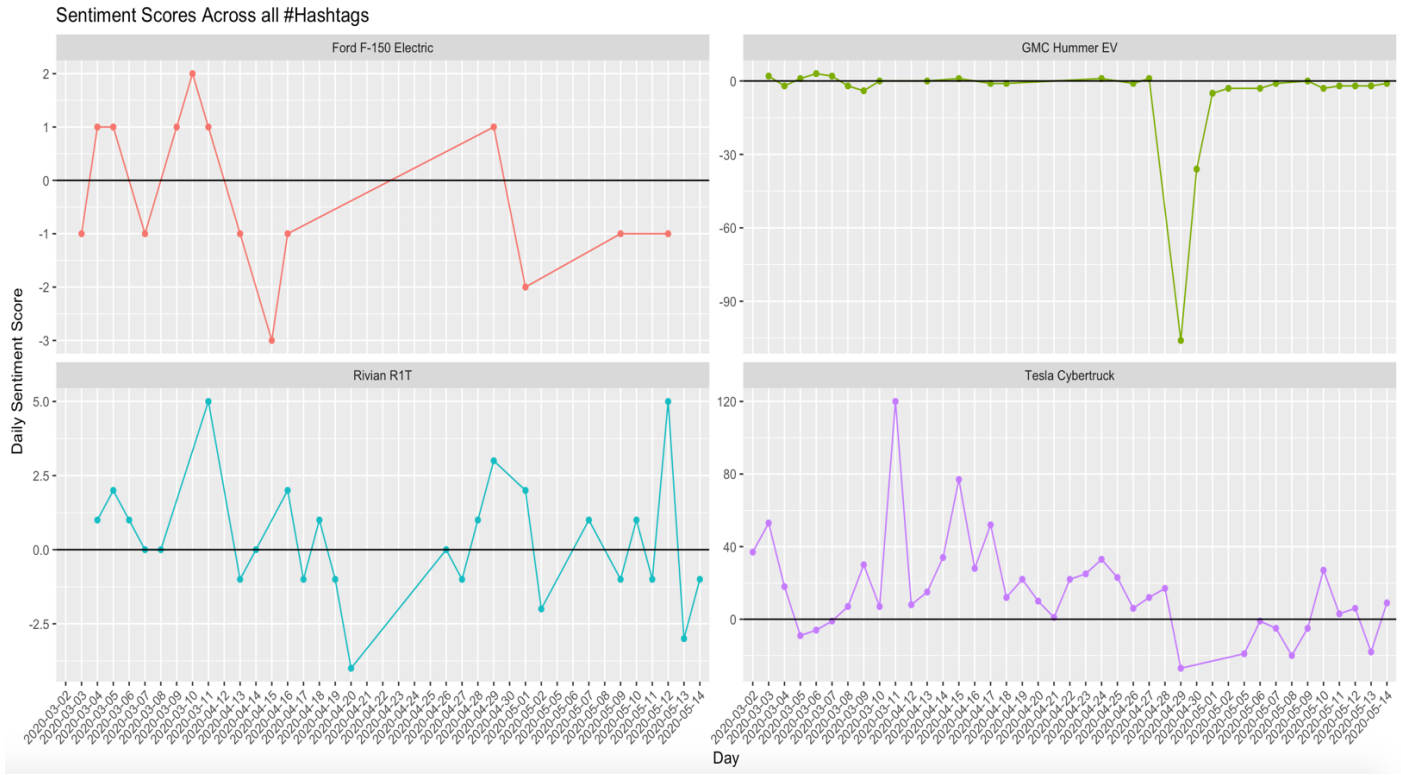
Figure 14: Violin plot of sentiment score for different EVs brands using "saotd"



Retrieved from R Studio

A first analysis of the data frame allowed to understand the general trend for each brand. For instance, the more positive tendency for Cybertruck was confirmed. Rivian R1T followed that path, with a mean that is similar to the Tesla Model. F-150 seemed to be generally neutral, thus the retrieved tweets were not polarized. Conversely, GMC Hummer EV shows the lowest average, in addition to the fact that an exact box plot could not be retrieved. The trends were confirmed by the following figure, that shows the violin plots for each brand. Having understood the differences across the models, an exact evaluation of the sentiment evolution allowed to inspect why the scores were different and to what extent. To find this out, the following plot showed the daily evolution of the samples' sentiment scores. The graph showed a sum of the lexicon sentiment results; thus, it did not take into account a sentiment range, but rather a daily sum of the scores:

Figure 15: Overtime sentiment score across each EVs brand using "saotd"



Retrieved from R Studio

This graph gave back meaningful insights even if the tweets collected for each brand differed in number. It showed the days of tweets samples in year-month-day order on the axis and the daily sentiment score sum in the ordinates. As expected, the trends showed value that were close to zero in most of the cases, except from the Cybertruck that could also count on higher sentiment sums. Moreover, going into the breakdown of the results across each electric pickup model, it is possible to point out that Rivian R1T had more positive peaks than negative, thus confirming what was seen before. Nevertheless, the results were somehow not volatile, with a sum range going from about -3 to 5 in score. The same applied for Ford, which had the smallest set of tweets and sentiment range, with the latter that went from -3 to 2. A narrower score range for this study, not taking into account the size of the dataset, can either mean that no big rumours or news were made about the car model or that the reaction towards were almost neutral overall. That can be particularly true for the Ford electric pickup, as not many information about the model are already in place. For what concerns Rivian R1T, the data changed more in the time range taken into account but still not with meaningful peaks and lows. What captured the attention the most were the scores for the models in the right side of the graph, namely GMC Hummer EV and Tesla Cybertruck. While the peaks of Tesla's electric pickups on March 11th 2020 and April 15th 2020 were already identified, since the former date is the one in which Elon Musk tweeted about scouting a new location in Central USA for the new Gigafactory and the latter coincides with the $1 billion offer made by Missouri for building the factory on their state (Crum, 2020) (HDMotori.it, 2020), the clear low for GMC Hummer EV registered on April 29th 2020 still needed to be inspected. On that date, in fact, GMC officially stated that the 20th May GMC Hummer EV's reveal date was postponed to an undetermined later date due to the Coronavirus pandemic (GMC Pressroom, 2020) (Capparella, 2020). That was also the reason why the

reveal of Rivian R1T was postponed too, thus eventually justifying the drop in April (Raia, 2020). Apart from that news, GMC maintained a pretty neutral score, so this can mean that the perception of product quality was not the reason why the GMC electric pickup had scored slightly more negative compared to the competitors overall.

Conclusion

The study has gone through the identification of the elements that characterizes the text mining field of study, describing some of its main form of implementation and techniques used to process large amount of textual information, highlighting also the relationship it has with machine learning and Natural Language Processing. Moreover, the subfields of sentiment and social media analysis were presented, highlighting how the tools arising from text mining can be adapted for this specific objective and domain, respectively. In particular, these two composed the structure of this study, as sentiment analysis was applied to the Twitter domain.

Furthermore, the object of this study was presented. Firstly, the company Tesla was introduced, with its totally electric transport vision and its ability to derive value out communication efficiently, especially thanks to its CEO Elon Musk. Therefore, the challenges in the change of the automotive ecosystem towards the total electrification of the vehicles were showed and discussed, evidencing the important role that Tesla had and still has in this conversation. Then, the actual product of interest for the study was discussed: this is the Tesla Cybertruck, a polarizing electric pickup truck that has made itself heard for its peculiar shape and its controversial unveiling event. As seen, these contributed to make this model really popular and talked about, thus making interesting to be analysed from a sentiment perspective. The last chapter was then focused on the sentiment analysis of this product, introducing programming and computational tools to conduct the study in the chosen domain Twitter. In particular, the findings were divided into four rounds: the first three were aimed at studying the Cybertruck with ever deeper levels of analysis, divided into different time ranges. The first was focused on a monthly analysis of single words, assessing their frequency and sentiment value through a lexicon-emotion association. The second was conducted on the general dataset focused on the identification of relevant topics of discussion in tweets through the use of a probabilistic method called Latent Dirichlet Allocation or LDA. After having individuated them, a lexicon-based sentiment analysis of the tweets containing some of these most relevant topics was conducted, to understand the opinion value for these specific arguments. The third was again not monthly and the aim was to adopt a lexicon that can retain as more textual information form tweets as possible. A special dictionary was then used that could give value to slangs, emoticon, emoji, valence shifters, amplifiers, adversatives and question weights, thus capturing more information contained in a sentence and deepening the level of analysis. Lastly, the tweets of Cybertruck were compared to the ones of three competitors, namely GMC Hummer EV, Rivian R1T, Ford F-150 Electric, in the fourth round. This analysis was conducted though special functions that allowed a sentiment analysis comparison for different set of tweets. In fact, the score for each brand and their sentiment evolution over time were then confronted. For all these rounds, specific visualization tools were implemented to better highlight the findings of these analysis.