Unsupervised Learning

Elon Musk:
An Exemplary Twitter User Analysis

Lukas Hermans

Università degli Studi di Milano
lukas.hermans@studenti.unimi.it

June 29, 2021

Abstract

The analysis of social media profiles plays an essential role in a variety of fields, e.g., for companies that seek to increase their returns via targeted advertising. In the present work, an exemplary user analysis of the Twitter profile of the tech entrepreneur Elon Musk is provided. The dataset consists of every Tweet posted by Elon Musk until the end of 2020. His Twitter activity has increased significantly since his first Tweet in 2010. An initial exploratory data examination reveals the dominance of his companies Tesla and SpaceX in the Tweets. The main part of the present work consists in the clustering of the Tweets by applying unsupervised machine learning techniques. The application of K-means clustering offers informative insights for $K = 8$ clusters. Hierarchical clustering leads to 15 interpretable clusters. A comparison of the obtained insights with the general knowledge about Elon Musk reveals a precise description of his career, interests, and personality by the applied clustering techniques. The exemplary analysis of Elon Musk's Twitter profile proofs the applicability of text mining and unsupervised clustering techniques for social media profiling.

Contents

1 Introduction

Text mining and unsupervised clustering techniques are important tools for the profiling of social media users. The insights from the application of these techniques are essential in a variety of fields. Among these are the maximization of returns via the personalization of advertisements, the analysis of the home locations of users, and the social sciences, where - for instance - the popularity of politicians is infered from social media posts [1][2][3].

The scope of the present work is to prove the applicability of text mining and unsupervised clustering techniques for the profiling of Twitter users. Twitter is a social media platform that was introduced in 2006 by Jack Dorsey [4]. The analysis is focused on Elon Musk - the CEO and founder of Tesla, SpaceX, Neuralink, and several other companies [5]. The applied unsupervised clustering techniques are K-means clustering and hierarchical clustering. In general, these unsupervised approaches are difficult to evaluate as there are no labels like for supervised machine learning tasks. The idea of the present work is that the focus on a public person like Elon Musk, however, allows to compare the obtained insights about the Twitter user with an expectation as Elon Musk is not an unkown person.

The structure of the present work is as follows. Section 2 describes the dataset that contains the Tweets of Elon Musk and related information. First insights are gained by an exploratory data analysis. The applied clustering techniques are described in Section 3, while the details of their implementation are covered by Section 4. Section 5 presents the results of the application of the unsupervised clustering techniques on the Tweet dataset. In Section 6, the insights from the exploratory data analysis and the application of the unsupervised clustering techniques are combined, and compared with the general knowledge about Elon Musk. Finally, a summary of the present work is given in Section 7.

2 Dataset & First Insights

The dataset of Elon Musk's Tweets is taken from Kaggle, see [6]. The data on Kaggle was retrieved from the Twitter API, see [7]. It contains all Tweets from Elon Musk from his first Tweet on 4 June 2010 to the last Tweet of 2020 on 28 December 2020. In total, the dataset consists of 11717 Tweets that were posted by Elon Musk in this time period. For each Tweet, additional information such as the number of likes and replies is provided.

Figure 1: Screenshot of Elon Musk's Twitter profile on 26 June 2021, see https://twitter.com/elonmusk?lang=en.

2.1 Twitter Activity

Fig. 1 contains a screenshot of Elon Musk's Twitter profile from 26 June 2021. At that date, the Twitter profile has 57.5 million followers, and follows 108 other Twitter users. The typical format of a Tweet is depicted in Fig. 2. A Tweet consists of a text of up to 280 characters, and may contain URLs, referals to other Twitter users, photos, or videos. Other Twitter users can share their reaction to the Tweet by liking, Retweeting or replying to it. A Retweet shows the Tweet on the profile of the Retweeting user, and is somewhat comparable to forwarding e-mails.

Figure 2: Screenshot of an examplary Elon Musk Tweet on 21 December 2020, see `https://twitter.com/elonmusk/status/1341006575650140161`. Each Tweet consists of a text of up to 280 characters, and may contain URLs, referals to other Twitter users, photos, or videos. Other Twitter users can like the Tweet, reply to it, and Retweet it.

10.00 % of Elon Musk's Tweets contain URLs to other websites (that may be simply another part of Twitter). 5.66 % of his Tweets contain photos, and 6.81 % embed videos. A Tweet is not necessarily an independent text, but may refer to another Tweet. This type of Tweet is called reply. 64.27 % of Elon Musk's Tweets are replies to other Twitter users, showing a strong interaction with the Twitter community.

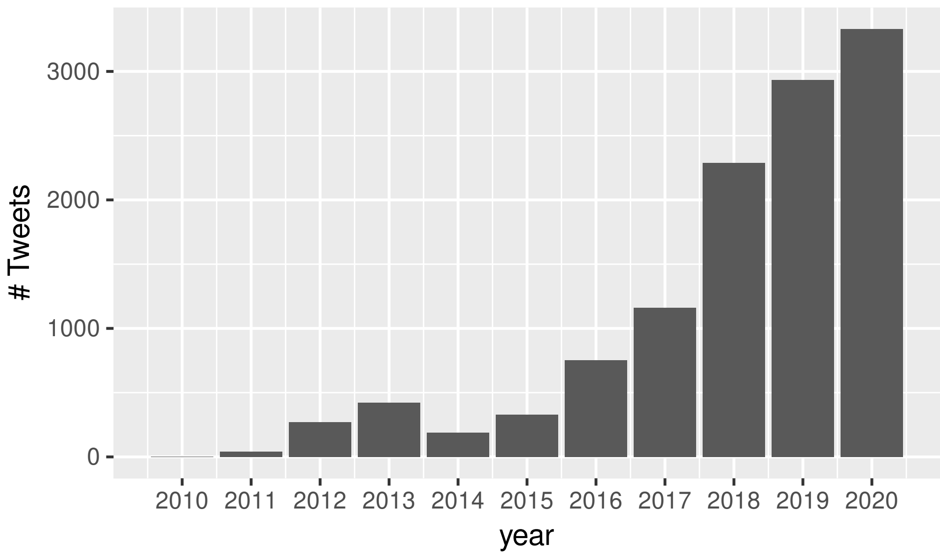

Figure 3: Total number of Tweets published by Elon Musk per year from 2010 to 2020. The barplot reveals a continuous increase of posted Tweets per year. In total, Elon Musk published 11717 Tweets in the reported time period.

Fig. 3 reports the total number of Tweets posted by Elon Musk per year. A continuous increase from the year 2010 to 2020 is observable. In the same time period the reactions of other Twitter users on his Tweets has increased, as is shown in Fig. 4. The Figure displays the average number of likes, replies, and Retweets on Elon Musk's Tweets for each year. The reported uncertainties are the standard error on the average. As there is only one Tweet in 2010, no uncertainty is displayed for that year.

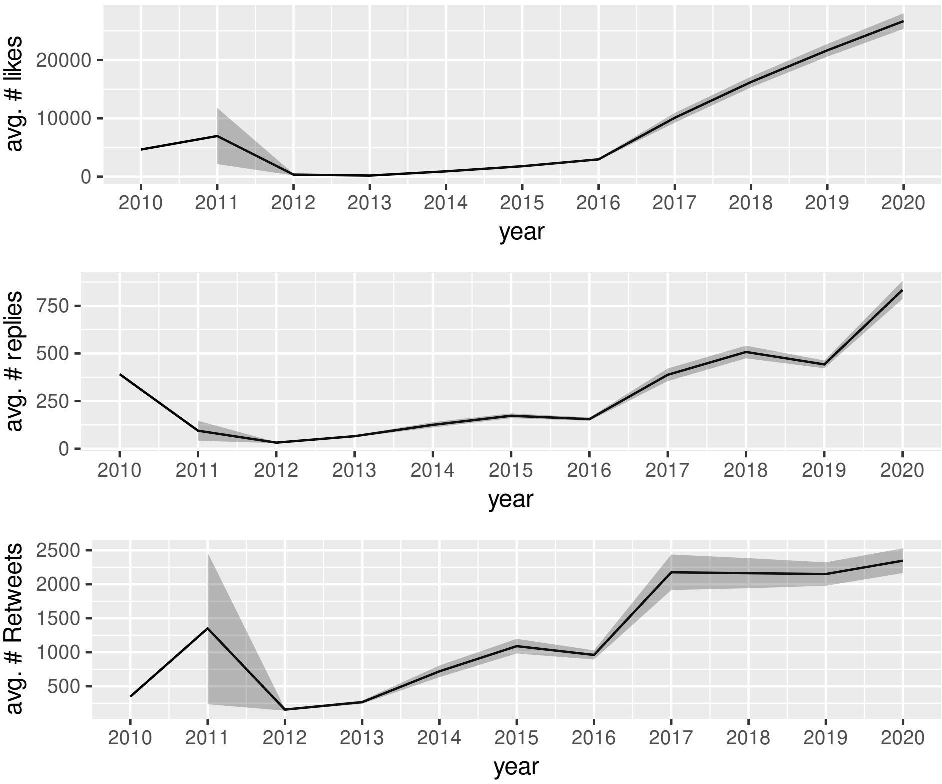

Figure 4: Average number of likes, replies, and Retweets on Elon Musk's Tweets per year. The displayed uncertainties are the standard error of the average. For 2010, no standard error is computed as Elon Musk published only one Tweet in this year. Parallel to Elon Musk's increasing Twitter activity over the years observed in Fig. 3, also the average number of likes, replies, and Retweets grows.

Elon Musk's activity on Twitter is roughly equally distributed over the 24 hours of a day apart from a drop in the morning hours between 2 AM and 7 AM, see Fig. 5. The time refers to the pacific day time (PDT) which is the time zone of Elon Musk's home state California.

2.2 Interaction with Other Users & Websites

As already explained above, 64.27 % of Elon Musk's Tweets are replies to other Twitter users. First insights on Elon Musk's interests may be gained by analyzing the Twitter users that he mentions most often. An ordered barplot of these users can be found in Fig. 6. His main interest seem to be @Tesla and related accounts like @Teslarati, and @teslaownersSV. Secondly, @SpaceX and space-related users such as @Erdayastronaut, @NASASpaceflight, and the US space agency @NASA appear commonly.

Another source of information are the URLs in Elon Musk's Tweets. In total, 10.00 % of his Tweets until the end of 2020 contain URLs. The corresponding barplot is reported in Fig. 7. Note that only the domain name is reported as the focus lies on the detection of general interests. For this purpose, the analysis of specific URLs is not suitable, and would not be visualizable as a barplot. First of all, he commonly refers to social networks like Twitter itself, YouTube, and Instagram. This illustrates that he is also active (at least passively) on these platforms. Again, several URLs refer to Tesla. URLs containinig SpaceX are less common. Additionally, Elon Musk appears to be informed on economic topics as his referals to Bloomberg, Buiseness Insider, and the Wall Street Journal show.

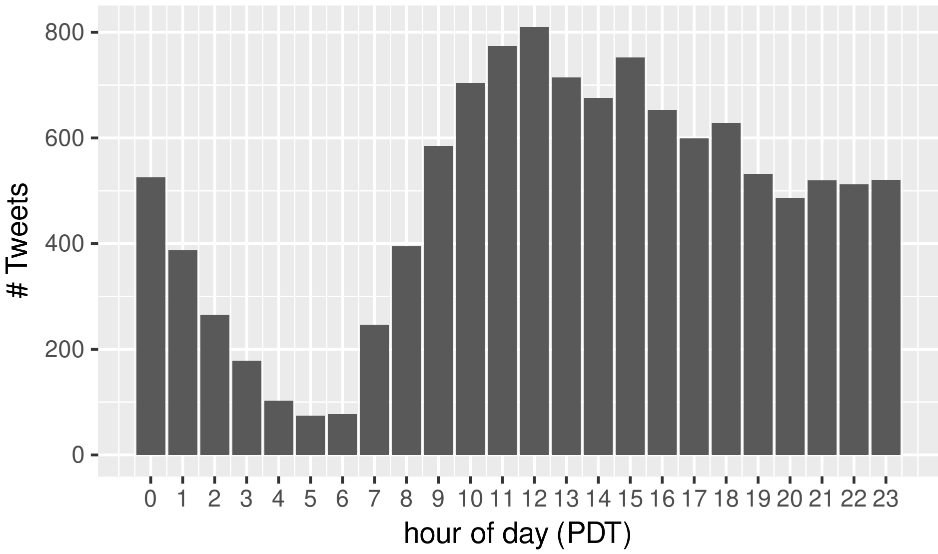

Figure 5: Total number of Tweets published by Elon Musk per hour of the day, which refers to pacific day time (PDT) - the time zone of Elon Musk's home state California. The Tweets are almost equally distributed over all hours of the day apart from a drop during the morning hours from 2 AM to 7 AM.

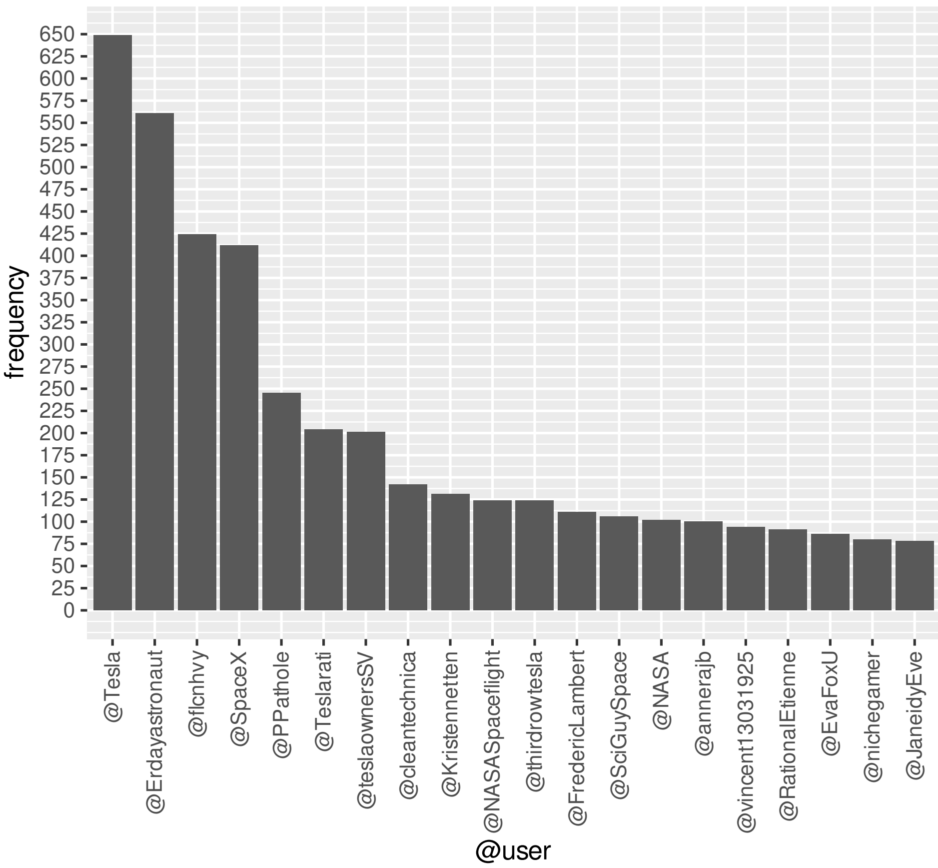

Figure 6: Frequency of referals to other Twitter users that are mentioned in at least 75 Tweets. In total, 64.27 % of Elon Musk's Tweets until the end of 2020 are replies to other Twitter users. Note, however, that referals to other Twitter users (@user) are also possible in Tweets that are not a direct reply. The interaction with Twitter accounts related to @Tesla is striking. In addtion, an interest in @SpaceX as well as space related users like @Erdayastronaut, @NASASpaceflight, and the US space agency @NASA is observed.

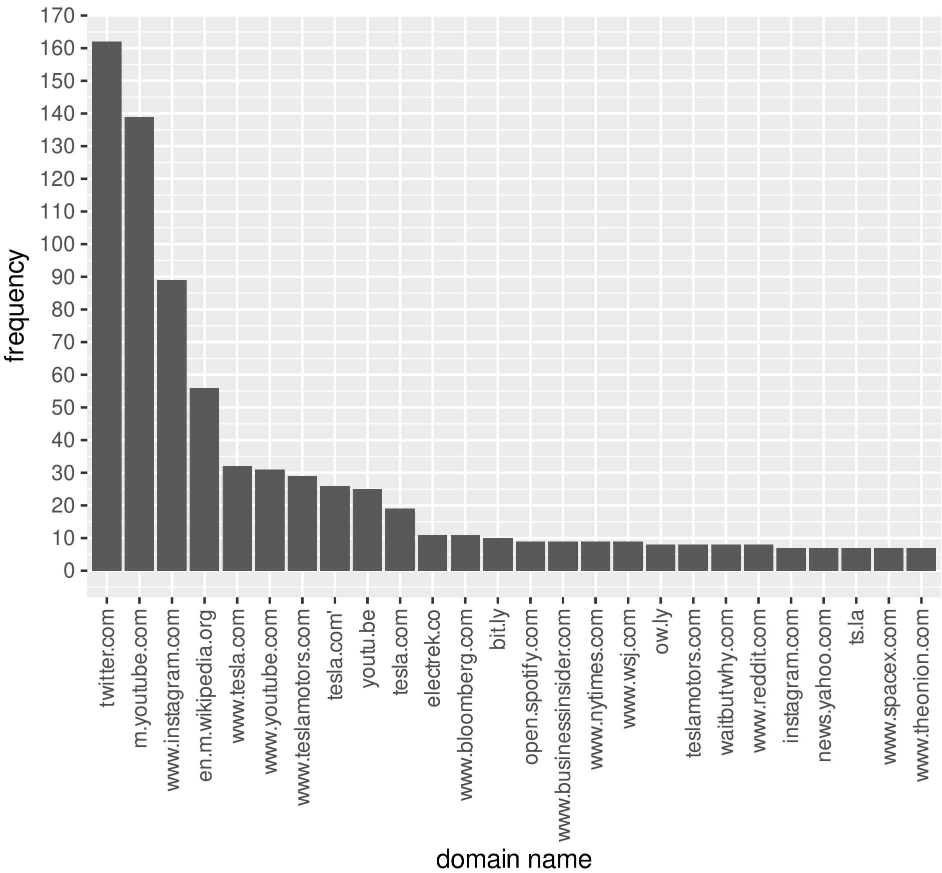

Figure 7: Frequency of referals to other websites that are mentioned in at least 6 Tweets. In total, 10.00 % of Elon Musk's Tweets until the end of 2020 contain URLs. Apart from the domains of social networks like Twitter itself, YouTube, and Instagram, again the interest in Tesla is striking. Elon Musk also seems to be up to date on economic related topics as his referals to Bloomberg, Buiseness Insider, and Wall Street Journal show.

2.3 Tweet Preprocessing & Word Usage

For the application of unsupervised clustering techniques in the following part of the present work, the text of the Tweets in the Kaggle dataset is preprocessed. The preprocessing consists in the following steps:

- The removal of all URLs and Twitter user names as they were already inspected above, and the following analysis is based on actual words in the Tweets.
- The transformation of all words to lower case because the analysis should not make a difference between upper and lower case words like “Tesla” and “tesla”.
- The removal of common stopwords, as reported in Fig. 13 and Fig 14 in the Appendix. Stopwords are words that appear commonly in (english) texts, and are thus not helpful in clustering the Tweets as they do not add meaning but have mostly a grammatical function.
- The removal of punctuation like dots, commas, and exclamation points.
- The removal of single standing numerical digits like “12” and “7”.

After the preprocessing, the remaining text of each Tweet is a sequence of words (that are not stopwords). All of the Tweets together contain 11572 different words (that may occur multiple times). In Fig. 8, the frequency of the most common words in all preprocessed Tweets is reported. Again, the word “tesla” occurs frequently. Other common words are “car(s)” and “model”. Besides general terms, referals to rockets and space like “rocket”, “spacex”, “Mars”, and “launch” appear in several Tweets.

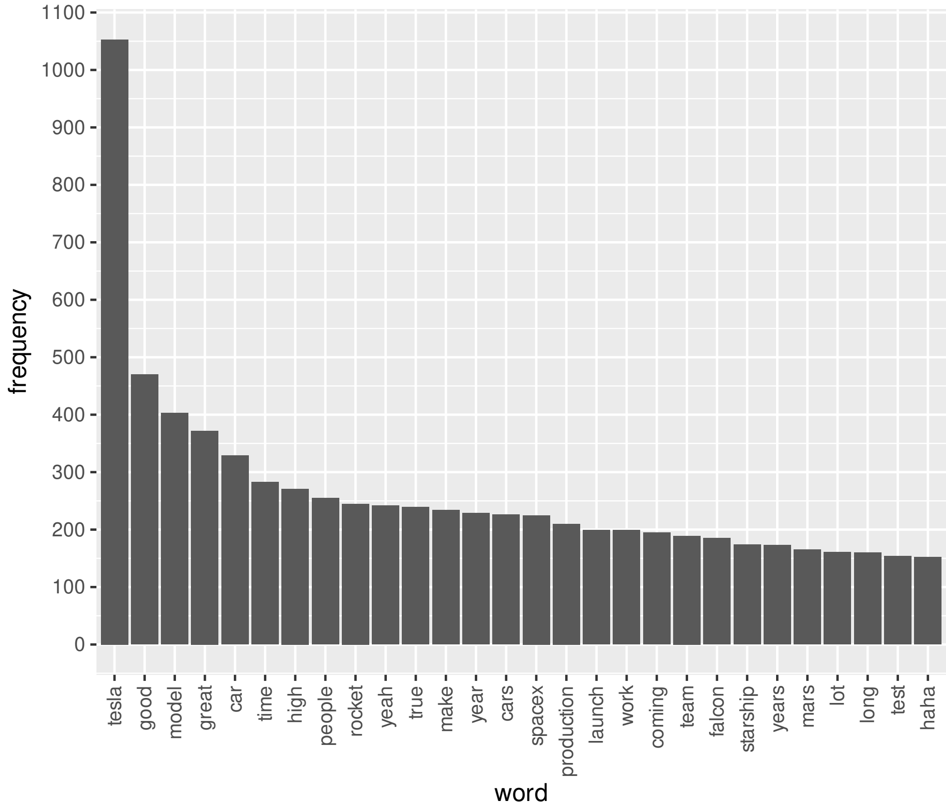

Figure 8: Frequency of words in Elon Musk's Tweets that occur at least 150 times. In total, all Tweets contain 11572 different words after the application of the stopword lists in Fig. 13 and Fig 14 in the Appendix. Striking is again the heavy usage of the word "tesla". In addition, the words "car(s)", and "model" occur frequently. Besides more general terms such as "good", and "true", several words refering to rockets and space occur, e.g. "rocket", "spacex", "Mars", and "launch".

3 Theory

The application of unsupervised clustering techniques requires that the preprocessed Tweets from the last Section are transformed into a document-term matrix, and a term-document matrix. The transformation is explained in Subsection 3.1. Based on these matrices, the K-means clustering and the hierarchical clustering approach are described in Subsection 3.2 and Subsection 3.3 [8].

3.1 Document-Term Matrix & Term-Document Matrix

The preprocessed Tweets are essentially a list of word sequences. The list has the length of the total number of Tweets, in the present case 11717. This Tweet list has to be translated into a mathematical expression. In the present work, the document-term matrix and the term-document matrix are used.

The rows of the *document-term matrix* X are the (preprocessed) Tweets, while the columns represent the words in each Tweet. Thus, the entry $x_{ij} = (X)_{ij}$ of the document-term matrix is 1 if the Tweet i contains the word j once. If a word occurs twice in a Tweet, the entry is set to 2, and so on. From this point of view, the Tweets are the observations, and the words are the features.

After the preprocessing, there are still 11324 different words. As each Tweet contains only a small subset of these words, the document-term matrix has many entries 0. In mathematical terms, the document-term matrix has a high sparsity. In order to reduce the sparsity, and to keep the memory footprint reasonable, a reduced document-term matrix is obtained by keeping only words that occur at least in 0.5 % of all Tweets - a negligible loss of information with respect to the original matrix for clustering purposes. The reduced document-term matrix reports only the occurence of the remaining 140 words, and has the dimension 11717×140.

The *term-document matrix* - as the name may suggest - is simply the transposed of the document-term matrix, denoted by X^T. In the present case, X^T (in its reduced form) has the dimension 140×11717. The term-document matrix encodes the words as observations, and the Tweets as features.

In the remaining part of the present work, whenever the document-term matrix or the term-document martix is mentioned, the reduced forms are meant.

3.2 K-Means Clustering

In the present work, the K-means clustering is applied on the document-term matrix, i.e. Tweets are clustered based on the occurence of words. The scope of K-means clustering is to cluster the Tweets into K categories. Each Tweet is contained in exactly one cluster. The hyperparameter K has to be specified prior to the application of K-means clustering. The idea is to minimize the so-called total within-cluster variation

$$G(C_1, \ldots, C_K) = \sum_{k=1}^{K} W(C_k),$$

where $C_1, ..., C_K$ are the K clusters. The within-cluster variation of cluster C_k is defined as

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2,$$

where x_{ij} and $x_{i'j}$ are the entries of the document-term matrix X, and $p = 140$ the number of words. In mathematical terms, K-means clustering seeks to compute

$$\underset{C_1,\ldots,C_K}{\operatorname{argmin}} \left\{ G(C_1, \ldots, C_K) \right\}.$$

An efficient algorithm to find a local optimum of this minimization problem starts with a random allocation of the Tweets to K clusters. Then, the clusters are improved iteratively until a sufficient convergence is obtained. It can be shown that each iteration of the algorithm always leads to a smaller within-cluster variation G.

3.3 Hierarchical Clustering

In contrast to K-means clustering, the hierarchical clustering makes use of the term-document matrix. This means that words are clustered based on their occurence in Tweets. Hierarchical clustering is an agglomorative, i.e. bottom-up, approach where every word is initially considered as a cluster. So, at the beginning, there are 140 clusters. Then, the two words with the lowest dissimilarity (based on some distance measure) between them are fused to a larger cluster, and there are 139 remaining clusters. This process is repeated until there is only 1 remaining cluster. In doing so, it is necessary to define a generalized dissimilarity measure that is not only applicable between words but also between clusters of words. This dissimilarity measure is called *linkage*.

4 Implementation & Software

The unsupervised clusering algorithms in the present work are implemented in the programming language R (version 3.6.3). As IDE RStudio (version 1.4.1717) is used. All R code, the dataset, and the LaTeX-code of the present paper can be found in the following GitHub repository: `https://github.com/lukher98/stat-learn`. For the present work, the folder “unsupervised” is relevant. The subfolder “data” contains the original dataset from Kaggle (“tweets_musk.csv”) as well as the preprocessed Tweets (“tweets_clean.csv”).

All of the code was executed on a Linux machine, but it should also work on a Mac machine. For Windows, the user might have to change some path names inside the code.

5 Results

5.1 K-Means Clustering

For the application of K-means clustering, a suitable choice for the hyperparameter K - the desired number of clusters - has to be found. In order to do so, Fig. 9 reports the total within-cluster variation G as a function of the number of clusters K. A decrease of the total within-cluster variation G means that the words in each cluster become more similar. However, the within-cluster variation G is a strictly decreasing function of K. Thus, a smaller G does not necessarily mean that the clustering is more accurate as a larger K leads *always* to a smaller G. A reasonable tradeoff is to chose K such that the decrease of G as a function of the cluster number K has already become small. In other words, K is taken in an interval where the total within-cluster variation decreases only slightly in comparison to its initial decrease for small values of K. Usually, this method is called ellbow method as the most suitable value of K lies in an ellbow-shaped part of the curve. In the present case, the value $K = 8$ - indicated by the dashed, red line in the Figure - marks a point in such an ellbow, and is used for further analyses.

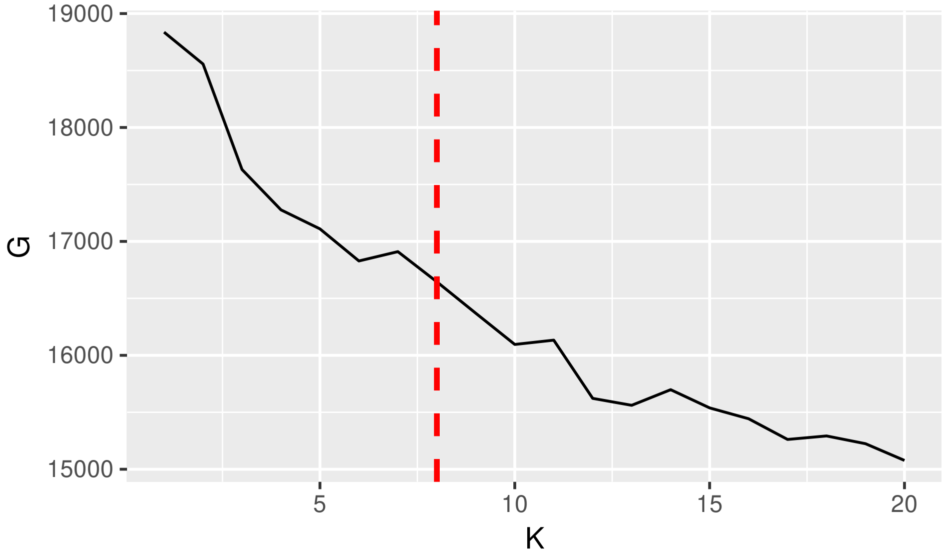

Figure 9: Total within-cluster variation G as a function of the number of clusters K using K-means clustering. The dashed, red line indicates the value $K = 8$ which is used for the further analysis based on the ellbow method, see the main text for more details. The most common Tweeted words in the $K = 8$ clusters are reported in Fig. 10, while the number of Tweets in each cluster is summarized in Tab. 1.

If the desired number of clusters is set to $K = 8$, the Tweets are distributed among the clusters as reported in Tab. 1. Most remarkably is cluster 3 with 9693 associated Tweets. Note that the numbering of the clusters is arbitrary, and does not indicate any hierarchy among the clusters.

cluster	1	2	3	4	5	6	7	8
# Tweets	154	831	9693	128	113	413	102	283

Table 1: Number of Tweets per cluster for the application of K-means clustering with $K = 8$ clusters. The total number of tweets is 11717.

Fig. 10 shows the 10 most common words in the Tweets in each of the $K = 8$ clusters. The numbering of the clusters is the same as in Tab. 1.

Cluster 3 mainly contains positive emotions like “good”, “great”, and “yeah”. These are connected to the term “model”. There is also the word “coming”. A summary of cluster 3 could be Elon Musk’s euphoria towards a “coming model”.
Cluster 2 and cluster 8 give insight in what is meant with the term “model”. Both clusters show a connection with the word “tesla”. Especially cluster 8 reveals that the term “model” refers to a new “car”. Additionally, the “electric” car might have an “autopilot”.

Cluster 7 includes mostly energy-related terms such as “solar”, “power”, and “battery”. Here, also the term “tesla” - that is related to cars - occurs.
The clusters 1, 4, and 6 are related to rockets and space. In cluster 1, especially the planet “mars” appears to be of interest to Elon Musk. Cluster 4 reveals that “raptor” might be the name of a rocket “engine”. In general, the terms “falcon” and “starship” seem to be related to rockets though it is not clear what they specifically describe.

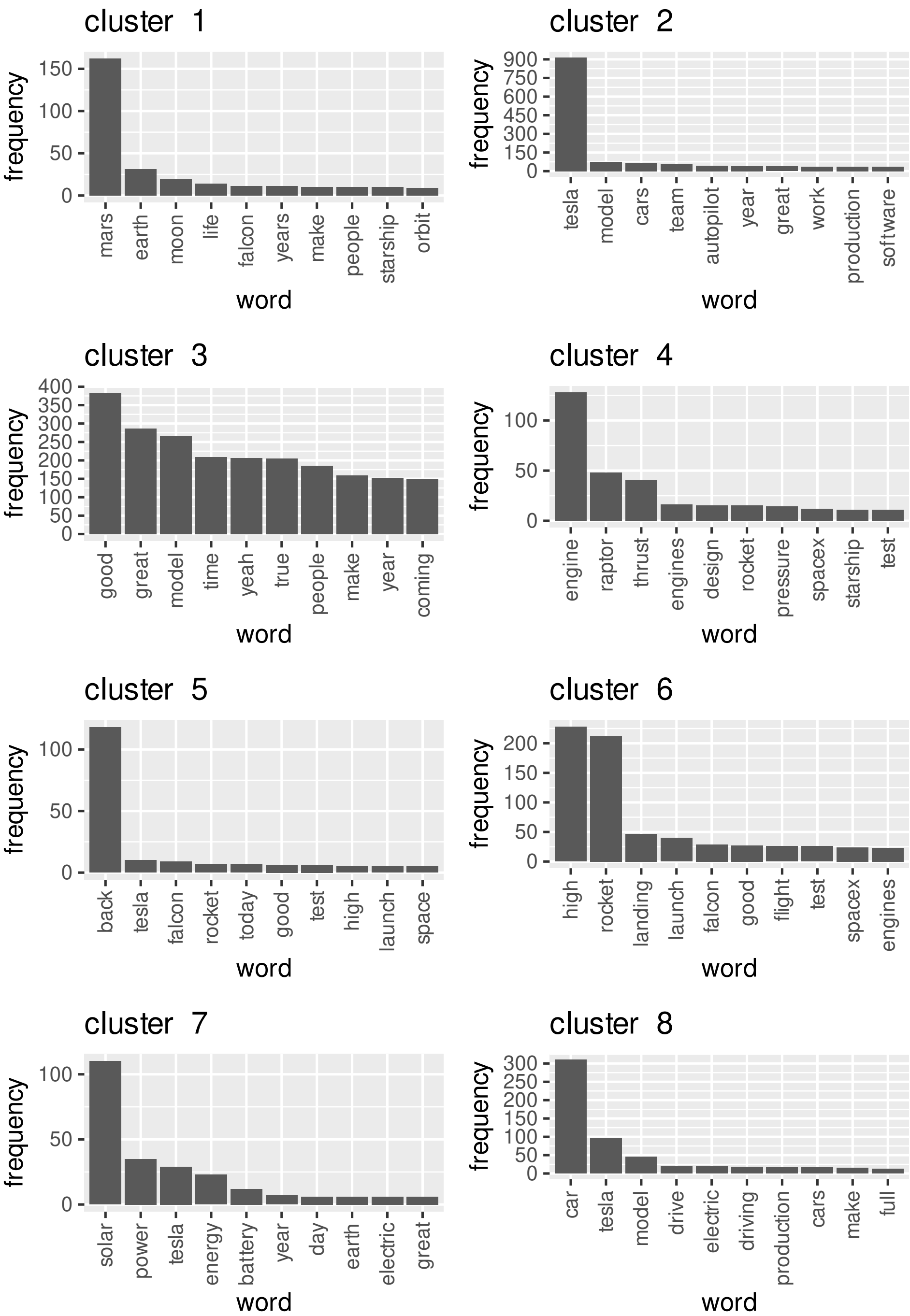

Figure 10: Barplot of the 10 most common Tweeted words in $K = 8$ clusters using K-means clustering. The meaning of the clusters in detail is discussed in the main text.

5.2 Hierarchical Clustering

In contrast to the K-means clustering from the last Subsection, the hierarchical clustering does not require an a priori selection of the desired number of groups. However, as explained in the theory Section, a distance measure between the words as well as the linkage as a measure of dissimilarity between clusters has to be defined. In the present work, the binary distance (also called Jaccard distance) in combination with Ward's minimum variance method as the linkage are found to lead to the most interpretable clustering results. Note that Ward's minimum variance method is originally designed to be combined with the Euclidean distance. Nonetheless, given the excellent results of the combination with the binary distance, this combination underlies the following discussion of the hierarchical clustering.

Fig. 11 shows the so-called dendogram of the application of the hierarchical clustering to the 140 words of the term-document matrix. Each leaf of the dendogram is one word. The dendogram reports how the single words are merged to clusters, and how these clusters are fused to larger clusters. The root of the dendogram is a cluster that contains all words. The vertical axis at the lefthand side of the dendogram shows the value of the dissimilarity measure (in this case Ward's minimum variance method in combination with the binary distance) at the respective height of the dendogram. Clusters that are merged at a lower height are more similar than clusters that are merged at a higher point of the dendogram. The red boxes in the Figure indicate 15 clusters. The words in each cluster are visualized as wordclouds in Fig. 12. Note that the number of clusters does not have to be the same as for the K-means clustering because the clusters are heuristic categorizations of the words. Necessarily, this categorization depends on the applied clustering technique. In addition, here, words are clustered, whereas in the previous n Tweets were clustered.

The clusters 1, 2, 3, 10, and 11 are in relation to Elon Musk's interest into rockets and space. In particular, new information in comparison to the results from the K-means clustering is obtained from cluster 3. Here, it appears that Elon Musk announces test flights and space missions that take place "tomorrow" or "today". Thus, his interest into space seems to be related to particular events, i.e. rocket launches.
The clusters 5, 6, and 8 are related to tesla, and his interest in cars. Most importantly, cluster 8 shows that Elon Musk also announces the release of cars (indicated by the word "drive") in the next "days" or "weeks".
Cluster 7 regards announcments of "coming software updates". From the words in the cluster, it is not clear to what software these announcments refer to.
While cluster 15 is focused on the topic "energy", cluster 4 contains completely new information that was not visible in the results of the K-means clustering above. Apparently, Elon Musk is also interested in another company that is related to "tunnels". He seems to be particularly fascinated by the "boring" process of these tunnels.
Finally, the remaining clusters 9, 12, 13, and 14 do not allow the retrieval of further information about Elon Musk. They seem to be simply accumulations of losely related words. A reason for the observation of these clusters is that the hierarchical clustering puts *all* words into a cluster. However, there might be words and Tweets that have a single meaning, and are not really part of a certain cluster.

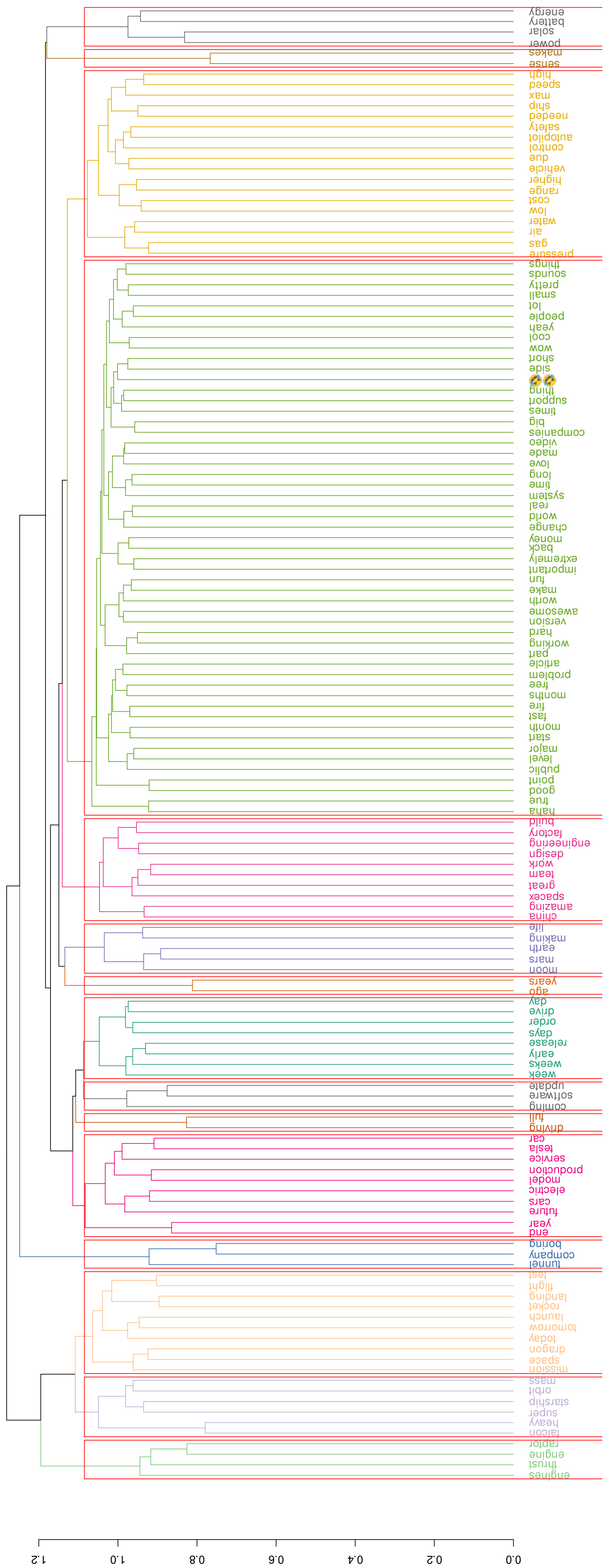

Figure 11: Dendogram for the application of hierarchical clustering on the 140 words of the term-document matrix of Elon Musk's Tweets. The vertical axis shows the value of the dissimilarity measure (in this case Ward's minimum variance method in combination with the binary distance) at the respective height. The red boxes and colorization of the different parts of the dendrogram show the word categorization for 15 clusters. A visualization of the words in the 15 clusters is given in Fig. 12.

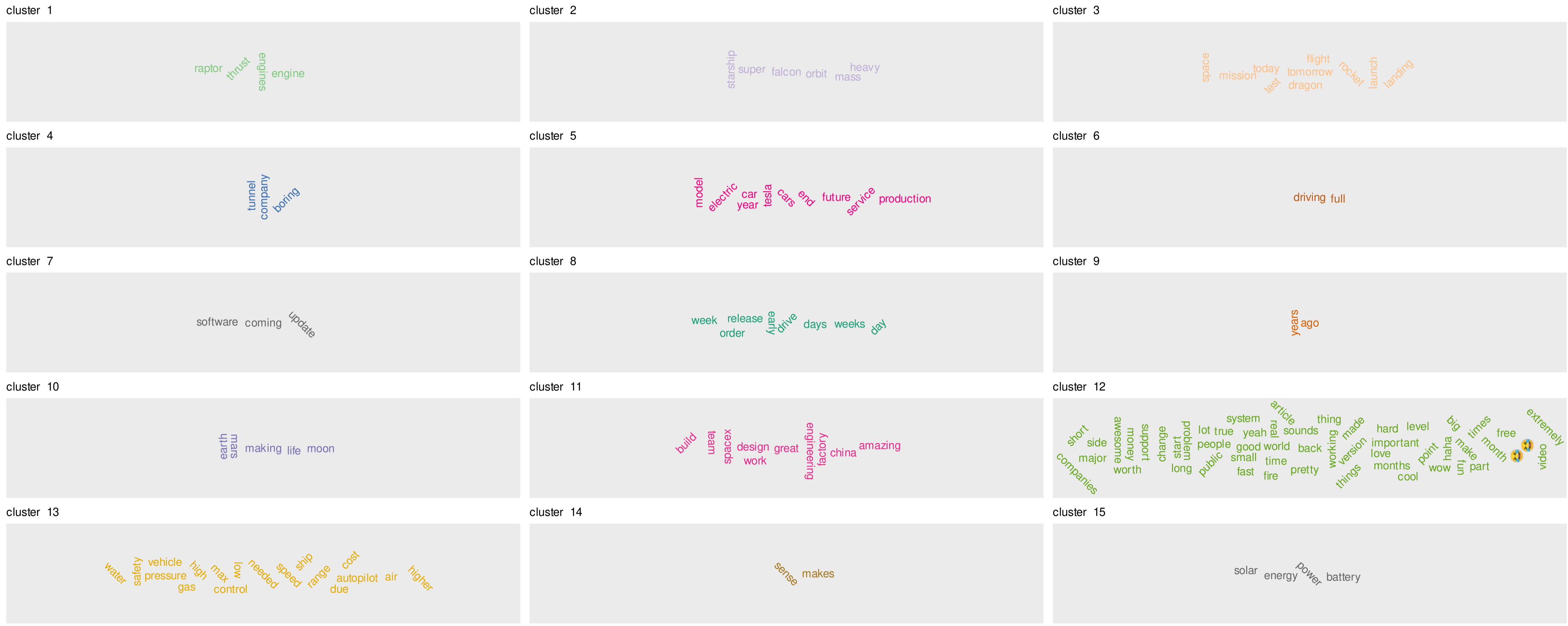

Figure 12: Wordclouds of the 15 clusters obtained from the application of hierarchical clustering in Fig. 11. An interpretation of the clusters can be found in the main text.

6 Discussion

In the present work, information on Elon Musk's personality, interests, and career are obtained from an initial exploratory data analysis, followed by an application of the unsupervised machine learning techniques K-means clustering and hierarchical clustering. All these information lead to a detailed panorama about the person Elon Musk. Tab. 2 contains the profile of Elon Musk that is the combination of the results from the analyses in the present work. The obtained profile could for example be used to show Elon Musk personlized advertisements based on his interests.

name	Elon Musk
Twitter followers	57.5 million
# Tweets	11717
replies	64.27 %
acitivity	all over the day, less from 2 AM to 7 AM
tesla	electric car company, autonomeous driving
spacex	rockets, engine called “raptor”, starship, falcon, Mars, Moon, Earth
energy	solar power, batteries
tunnels	company, boring tunnels
personality	euphoria about upcoming software updates and new models
information	up to date on economic-related topics

Table 2: Summary of the results from the intitial exploratory data analysis, the K-means clustering, and the hierarchical clustering. The combination of the different approaches lead to a detailed profile of the person Elon Musk. In practice, this profile could for example be used to show Elon Musk personalized advertisements.

The remaining part of this Section regards the question of how accurate this profile of Elon Musk is. First of all, his main interests in the automobile manufacturer Tesla, and the space company SpaceX are obvious given that he is the founder and CEO of these companies. The clustering reveal even that the rocket engines of SpaceX are called “Raptor”, but do not show that “Starship” and “Falcon” are the names of rockets. It is also true that Tesla is an electric car company that produces cars also with an autopilot. As new models of Tesla cars, and updates of the autopilot are developed frequently, it makes sense that he shows euphoria about these aspects and advertises them in some degree. The clustering techniques show that he is interested in these topics, but do not reveal his function as the CEO in these companies. However, for personalized advertisements this information is not necessary.
Then, the interest in energy and tunnels is explained by his companies SolarCity and the Boring Company. The analysis does not reveal the name of these companies but it does show their product spectrum.
The analyses in the present work reveal most of Elon Musk's interests based on his companies. Only the company Neuralink - that develops an implementable brain-machine interface - is not abstracted from any cluster. A reason is that this company was founded in 2016, and is thus not be included in many Tweets.

All in all, it is concluded the the profile of Elon Musk based on his Twitter activity is highly accurate compared to what is known in general about his companies. This proves the applicability of text mining and unsupervised clustering techniques for the profiling of social media users. The analyses can be repeated for less known persons, and the profiling results can then for example be used for personalized advertisements.

7 Conclusion

The exemplary social media analysis of Elon Musk's Twitter account proves that text mining techniques and unsupervised clustering techniques are applicable for the profiling of social media users. The retrieved profiles can for example be used to show the users personalized advertisments. The combination of the results of an initial exploratory data examination, and the application of K-means clustering and hierarchical clustering is summarized in Tab. 2. The analyses in the present work lead to a detailed profile of Elon Musk. A comparison with the general knowledge about Elon Musk and his companies reveals the accuracy of the obtained profile. For the analyses in the present work, the K-means clustering is applied with $K = 8$ desired clusters, while the hierarchical clustering leads to 15 interpretable clusters. These differences lie in the different natures of the clustering approaches.

References

[1] Gabriella Kazai, Iskander Yusof, and Daoud Clarke. "Personalised News and Blog Recommendations Based on User Location, Facebook and Twitter User Profiling." In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval.* SIGIR '16. Pisa, Italy: Association for Computing Machinery, 2016, 1129–1132. ISBN: 9781450340694. DOI: 10.1145/2911451.2911464. URL: https://doi.org/10.1145/2911451.2911464.

[2] "Towards social user profiling: Unified and discriminative influence model for inferring home locations." English (US). In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2012, pp. 1023–1031. ISBN: 9781450314626. DOI: 10.1145/2339530.2339692.

[3] Fabricio De França, Denise Goya, and Claudio Penteado. "User profiling of the Twitter Social Network during the impeachment of Brazilian President." In: *Social Network Analysis and Mining* 8 (Dec. 2018). DOI: 10.1007/s13278-017-0482-x.

[4] Twitter. *Our Company.* 2021. URL: https://about.twitter.com/en/who-we-are/our-company (visited on 06/28/2021).

[5] Ashlee Vance. *Elon Musk : Tesla, SpaceX, and the quest for a fantastic future.* eng. First edition. New York, NY: Ecco, an imprint of HarperCollinsPublishers, 2015. ISBN: 9780062301239.

[6] Ayhm Dalila (Kaggle User). *Elon Musk Tweets 2010-2021.* 2021. URL: https://www.kaggle.com/ayhmrba/elon-musk-tweets-2010-2021?select=2020.csv (visited on 06/25/2021).

[7] Twitter. *Twitter API.* 2021. URL: https://developer.twitter.com/en/docs/twitter-api (visited on 06/28/2021).

[8] Gareth James et al. *An Introduction to Statistical Learning: With Applications in R.* Aug. 2013. ISBN: 9781461471370.

Appendix

"i"	"me"	"my"	"myself"	"we"	"our"	"ours"	"ourselves"	"you"	"your"	"yours"
"yourself"	"yourselves"	"he"	"him"	"his"	"himself"	"she"	"her"	"hers"	"herself"	"it"
"its"	"itself"	"they"	"them"	"their"	"theirs"	"themselves"	"what"	"which"	"who"	"whom"
"this"	"that"	"these"	"those"	"am"	"is"	"are"	"was"	"were"	"be"	"been"
"being"	"have"	"has"	"had"	"having"	"do"	"does"	"did"	"doing"	"would"	"should"
"could"	"ought"	"i'm"	"you're"	"he's"	"she's"	"it's"	"we're"	"they're"	"i've"	"you've"
"we've"	"they've"	"i'd"	"you'd"	"he'd"	"she'd"	"we'd"	"they'd"	"i'll"	"you'll"	"he'll"
"she'll"	"we'll"	"they'll"	"isn't"	"aren't"	"wasn't"	"weren't"	"hasn't"	"haven't"	"hadn't"	"doesn't"
"don't"	"didn't"	"won't"	"wouldn't"	"shan't"	"shouldn't"	"can't"	"cannot"	"couldn't"	"mustn't"	"let's"
"that's"	"who's"	"what's"	"here's"	"there's"	"when's"	"where's"	"why's"	"how's"	"a"	"an"
"the"	"and"	"but"	"if"	"or"	"because"	"as"	"until"	"while"	"of"	"at"
"by"	"for"	"with"	"about"	"against"	"between"	"into"	"through"	"during"	"before"	"after"
"above"	"below"	"to"	"from"	"up"	"down"	"in"	"out"	"on"	"off"	"over"
"under"	"again"	"further"	"then"	"once"	"here"	"there"	"when"	"where"	"why"	"how"
"all"	"any"	"both"	"each"	"few"	"more"	"most"	"other"	"some"	"such"	"no"
"nor"	"not"	"only"	"own"	"same"	"so"	"than"	"too"	"very"		

Figure 13: English stopword list applied to Elon Musk's Tweets prior to further analyses. The english stopword list is part of the R text mining package "tm".

"a"	"a's"	"able"	"about"	"above"	"according"	"accordingly"	"across"	"actually"
"after"	"afterwards"	"again"	"against"	"ain't"	"all"	"allow"	"allows"	"almost"
"alone"	"along"	"already"	"also"	"although"	"always"	"am"	"among"	"amongst"
"an"	"and"	"another"	"any"	"anybody"	"anyhow"	"anyone"	"anything"	"anyway"
"anyways"	"anywhere"	"apart"	"appear"	"appreciate"	"appropriate"	"are"	"aren't"	"around"
"as"	"aside"	"ask"	"asking"	"associated"	"at"	"available"	"away"	"awfully"
"b"	"be"	"became"	"because"	"become"	"becomes"	"becoming"	"been"	"before"
"beforehand"	"behind"	"being"	"believe"	"below"	"beside"	"besides"	"best"	"better"
"between"	"beyond"	"both"	"brief"	"but"	"by"	"c"	"c'mon"	"c's"
"came"	"can"	"can't"	"cannot"	"cant"	"cause"	"causes"	"certain"	"certainly"
"changes"	"clearly"	"co"	"com"	"come"	"comes"	"concerning"	"consequently"	"consider"
"considering"	"contain"	"containing"	"contains"	"corresponding"	"could"	"couldn't"	"course"	"currently"
"d"	"definitely"	"described"	"despite"	"did"	"didn't"	"different"	"do"	"does"
"doesn't"	"doing"	"don't"	"done"	"down"	"downwards"	"during"	"e"	"each"
"edu"	"eg"	"eight"	"either"	"else"	"elsewhere"	"enough"	"entirely"	"especially"
"et"	"etc"	"even"	"ever"	"every"	"everybody"	"everyone"	"everything"	"everywhere"
"ex"	"exactly"	"example"	"except"	"f"	"far"	"few"	"fifth"	"first"
"five"	"followed"	"following"	"follows"	"for"	"former"	"formerly"	"forth"	"four"
"from"	"further"	"furthermore"	"g"	"get"	"gets"	"getting"	"given"	"gives"
"go"	"goes"	"going"	"gone"	"got"	"gotten"	"greetings"	"h"	"had"
"hadn't"	"happens"	"hardly"	"has"	"hasn't"	"have"	"haven't"	"having"	"he"
"he's"	"hello"	"help"	"hence"	"her"	"here"	"here's"	"hereafter"	"hereby"
"herein"	"hereupon"	"hers"	"herself"	"hi"	"him"	"himself"	"his"	"hither"
"hopefully"	"how"	"howbeit"	"however"	"i"	"i'd"	"i'll"	"i'm"	"i've"
"ie"	"if"	"ignored"	"immediate"	"in"	"inasmuch"	"inc"	"indeed"	"indicate"
"indicated"	"indicates"	"inner"	"insofar"	"instead"	"into"	"inward"	"is"	"isn't"
"it"	"it'd"	"it'll"	"it's"	"its"	"itself"	"j"	"just"	"k"
"keep"	"keeps"	"kept"	"know"	"knows"	"known"	"l"	"last"	"lately"
"later"	"latter"	"latterly"	"least"	"less"	"lest"	"let"	"let's"	"like"
"liked"	"likely"	"little"	"look"	"looking"	"looks"	"ltd"	"m"	"mainly"
"many"	"may"	"maybe"	"me"	"mean"	"meanwhile"	"merely"	"might"	"more"
"moreover"	"most"	"mostly"	"much"	"must"	"my"	"myself"	"n"	"name"
"namely"	"nd"	"near"	"nearly"	"necessary"	"need"	"needs"	"neither"	"never"
"nevertheless"	"new"	"next"	"nine"	"no"	"nobody"	"non"	"none"	"noone"
"nor"	"normally"	"not"	"nothing"	"novel"	"now"	"nowhere"	"o"	"obviously"
"of"	"off"	"often"	"oh"	"ok"	"okay"	"old"	"on"	"once"
"one"	"ones"	"only"	"onto"	"or"	"other"	"others"	"otherwise"	"ought"
"our"	"ours"	"ourselves"	"out"	"outside"	"over"	"overall"	"own"	"p"
"particular"	"particularly"	"per"	"perhaps"	"placed"	"please"	"plus"	"possible"	"presumably"
"probably"	"provides"	"q"	"que"	"quite"	"qv"	"r"	"rather"	"rd"
"re"	"really"	"reasonably"	"regarding"	"regardless"	"regards"	"relatively"	"respectively"	"right"
"s"	"said"	"same"	"saw"	"say"	"saying"	"says"	"second"	"secondly"
"see"	"seeing"	"seem"	"seemed"	"seeming"	"seems"	"seen"	"self"	"selves"
"sensible"	"sent"	"serious"	"seriously"	"seven"	"several"	"shall"	"she"	"should"
"shouldn't"	"since"	"six"	"so"	"some"	"somebody"	"somehow"	"someone"	"something"
"sometime"	"sometimes"	"somewhat"	"somewhere"	"soon"	"sorry"	"specified"	"specify"	"specifying"
"still"	"sub"	"such"	"sup"	"sure"	"t"	"t's"	"take"	"taken"
"tell"	"tends"	"th"	"than"	"thank"	"thanks"	"thanx"	"that"	"that's"
"thats"	"the"	"their"	"theirs"	"them"	"themselves"	"then"	"thence"	"there"
"there's"	"thereafter"	"thereby"	"therefore"	"therein"	"theres"	"thereupon"	"these"	"they"
"they'd"	"they'll"	"they're"	"they've"	"think"	"third"	"this"	"thorough"	"thoroughly"
"those"	"though"	"three"	"through"	"throughout"	"thru"	"thus"	"to"	"together"
"too"	"took"	"toward"	"towards"	"tried"	"tries"	"truly"	"try"	"trying"
"twice"	"two"	"u"	"un"	"under"	"unfortunately"	"unless"	"unlikely"	"until"
"unto"	"up"	"upon"	"us"	"use"	"used"	"useful"	"uses"	"using"
"usually"	"uucp"	"v"	"value"	"various"	"very"	"via"	"viz"	"vs"
"w"	"want"	"wants"	"was"	"wasn't"	"way"	"we"	"we'd"	"we'll"
"we're"	"we've"	"welcome"	"well"	"went"	"were"	"weren't"	"what"	"what's"
"whatever"	"when"	"whence"	"whenever"	"where"	"where's"	"whereafter"	"whereas"	"whereby"
"wherein"	"whereupon"	"wherever"	"whether"	"which"	"while"	"whither"	"who"	"who's"
"whoever"	"whole"	"whom"	"whose"	"why"	"will"	"willing"	"wish"	"with"
"within"	"without"	"won't"	"wonder"	"would"	"would"	"wouldn't"	"x"	"y"
"yes"	"yet"	"you"	"you'd"	"you'll"	"you're"	"you've"	"your"	"yours"
"yourself"	"yourselves"	"z"	"zero"					

Figure 14: SMART stopword list applied to Elon Musk's Tweets prior to further analyses. The SMART stopword list is part of the R text mining package "tm". The list includes some words that are also part of the english stopword list from Fig. 13.