Customer Satisfaction of the German Energy Supply Chain

Predicting star ratings and company responses on Trustpilot

Final Report

Data Science Graduation Project @DataScientest, Paris
Apr 10, 2024

Matthias Isele - Stefanie Arlt

Outline

Abbreviations

- i.e. = that is
- B2C = business to consumer
- e.g. = for example
- etc. = and others
- max. = maximum
- min. = minimum
- IQR = inter quartile range
- ML = machine learning
- PCA = principal component analysis
- MAE = Mean Absolute Error
- MSE = Mean Squared Error
- RMSE = Root Mean Squared Error
- R^2 = R-squared

Management Summary

Our subject is **energy suppliers in Germany on the Trustpilot web site**, from which we scraped content in September 2023 in German language, presented in two datasets:

- The **overall ranking** for energy suppliers in Germany based on the Trustpilot scores: There was rating information on 37 suppliers in Germany including number of votes, as well as supported energy supplier categories, e.g. eco power, solar energy etc. It could be established that neither number of votes nor supported categories have a direct impact on the overall ranking.
- The **customer reviews and company answers** for these suppliers dated back to 2011 and included more than 3000 pages and was investigated in terms user activity and mood, correlations between numerical variables and a first analysis of customer comments and supplier answers.

Our findings include that **user activity on Trustpilot** for German energy suppliers skyrocketed in the last four years, from 2,000 to over 25,000 reviews. 83% of those reported customer experiences were either negative (1 star) or excellent (5 stars), in other words: Reviews are only written if user experience is on the extreme.
The average customer star rating grew during the pandemic years 2020 – 2022 (2.6 - 4.2 stars), getting a hit due to the energy crisis 2023 (3.6 stars). Over all the years, customers tend to be content with German energy suppliers, as the average rating never dipped below 2.5 stars

The **number of words of a customer comment** was a good indicator for the star rating: The more elaborate the comment, the worse the rating – regardless which year the user experience was. Companies tend to write longer answers if a comment is more elaborate, but positive and negative reviews were answered in the same time frame. The distribution of **number of words of company answers** is discontinuous, i.e. there are company specific standard answers for certain star ratings. A segmentation in terms of companies is necessary.

For the **Machine Learning part of our project**, the goal was twofold:

- to predict the star rating of customer posts from customer comments and
- to predict the length of company answers to customer postings, respectively the full answers, if possible.

After applying simple ML models for **rating prediction based on number of words in customer feedback** and its headline, we could reach not over 70% in accuracy. The support vector machine classifier and Random Forest classifier were applied both untuned and tuned, with SVM classifier models performing better but both lacking severely in the middle classes of this imbalanced data set.

Not only were these features insufficient but also the data set was too big and reached the limit of available computing capacity. Leveraging business practice and everyday use cases, the data was reduced in a step-by-step approach: using data with comments, then with comments and headlines, lastly exploring data only from one supplier. In addition, the multiclass target was transformed into a binary classification, dividing ratings 4 and 5 into the dominant class 1 and ratings 1 to 3 into class 0.

A significant improvement could be reached after applying **sentiment analysis** on the customer comment field, and then training the same models again. While we saw overfitting for both models, the results were overall much improved since both classes were predicted over 75 % correctly, for the dominant class even up to 90%, with accuracy reaching up to 88% overall.

Closely related to sentiment analysis is the **key word analysis**, which leverages the same techniques to extract meaningful comparing positive and negative comments between suppliers for reporting and visualization of the main interest points of the customers such as billing, service and metering. As the same keywords are used both for negative and positive comments, these could be used as input for large language model prompting to create individual responses.

The second task was to **predict the length of company answers or even the full answers**, if possible. As company answers were highly dependent on the answer policy of the company, we selected two sub data sets:

- Predicting answers of E.ON Energy turned out to be a binary classification problem. There were only two distinct answers, up to trivial modifications like spaces and captions. Using a logistic regression on the numerical variables of comment length, headline length and star rating resulted in an astonishing accuracy of 99.82% on the total data set. This is due to a very strong correlation of the star rating to the company answer. Predicting the answer based on **sentiment analysis** on the comments could not improve the result, as we let go of the star rating variable, still reaching an accuracy of 85%.
- In contrast, the answers of Octopus Energy were highly personalized which made it necessary to target the answer length instead of the full answers. Training several models on the numeric variables of comment length, headline length and star rating we reached RMSE's ranging from 0.47 to 0.55, the best model being linear regression and the worst model being decision tree regressor. Predicting the answer lengths based on **sentiment analysis** on the comment variable ranged in between with an RMSE of 0.52. We expect sentiment analysis to be superior if more elaborate language models are used.

As next steps, we see optimization tasks to leverage sentiment analysis for better key word integration or to improve the modelling with advanced techniques like deep learning could be applied.

If we had had more time, we would integrate a pre-trained Language Model with an API to generate company answers to customer posts.

To conclude, in our project we have adopted key objectives in customer service management, i.e. to monitor customer feedback and regularly compare feedback between competitors.

In times of consumer markets and price comparison portals like Check24, customer expectations have evolved to personalized answers and timely feedback also on external platforms as the new normal. Therefore, the overall goal to turn happy customers into loyal customers can only be achieved by constantly checking competitors and applying AI supported customer service processes.

Introduction: Energy suppliers in Germany

Germany as part of the European energy market

The German energy market is a dynamic and complex sector that plays a crucial role in powering the nation's economy and meeting the energy needs of its citizens. To get an idea about the context of our data set, we will explore some key aspects of the German energy market, with special focus on consumers and important influential factors in the market.

Liberalization of the European Energy Market:

The liberalization of the European energy market in the late 1990s was facilitated by the European Union's directives, including the Third Energy Package[1]. It refers to the process of opening the energy sector to competition and creating a single market for energy across Europe. This initiative aimed to enhance competition, increase efficiency, and provide consumers with more choices. The liberalization has led to the establishment of a harmonized regulatory framework and the creation of a competitive market for energy products and services.

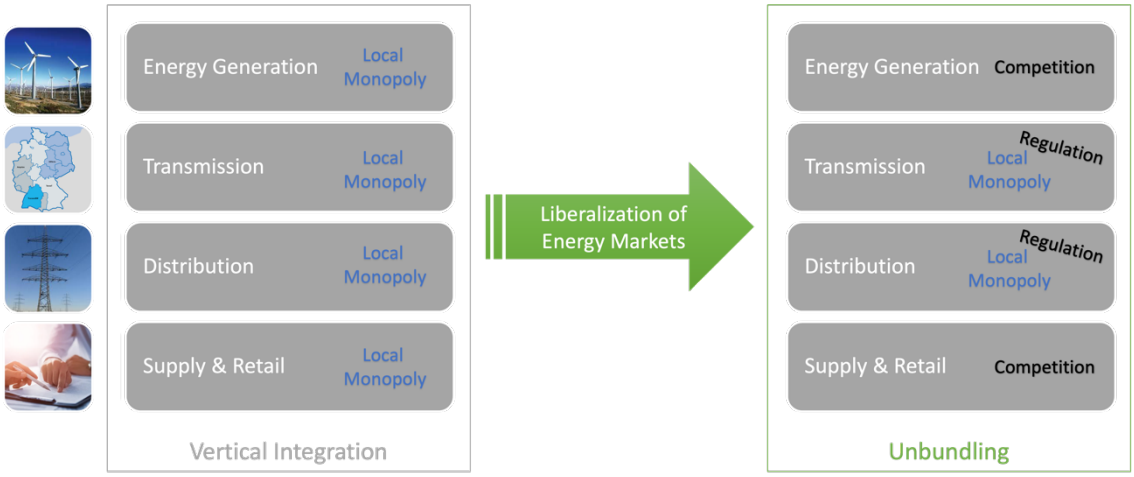


Figure 1: Energy supplier transformation after liberalization of energy markets

At the core of the regulation, we see unbundling as a key concept in the energy market, which involves the separation of energy production, transmission, and distribution activities to ensure fair competition and prevent monopolistic practices. Consequently , the vertically integrated energy companies in Germany have been separated into distinct entities, promoting transparency and fair market access for all players.

The Market

The German energy market is one of the largest in Europe, not only catering to its domestic demand but also interacting with other markets in Europe. In 2023, the net electricity generation in Germany amounted to 485 billion kWh, as opposed to a net power consumption of 466 billion kWh. Private households together with business and services consummated nearly the same as industry and traffic combined.

[1] Compare European energy market: https://www.europarl.europa.eu/factsheets/de/sheet/45/energiebinnenmarkt.

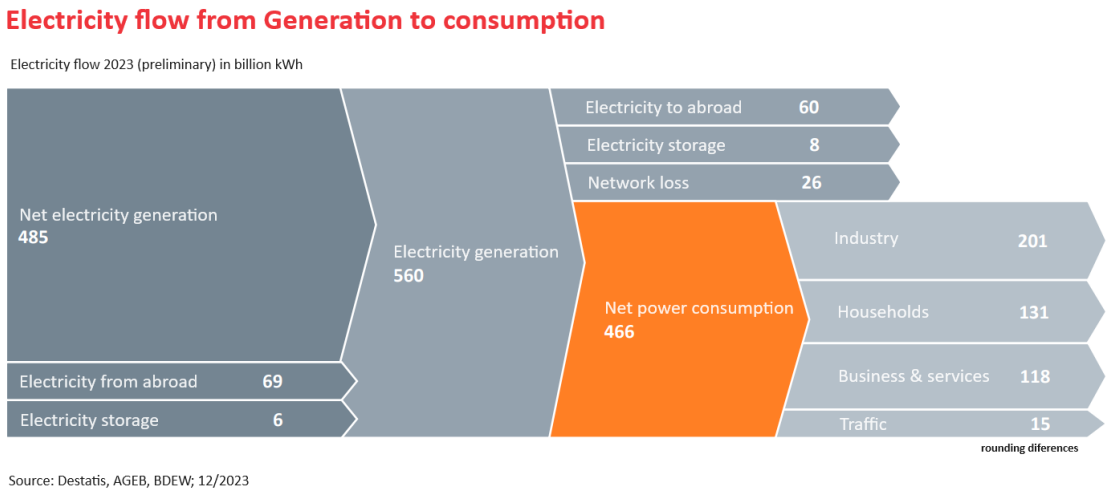


Figure 2: Electricity flow in Germany in 2023

Today the market, it is still dominated by four large energy supply companies: E.ON, RWE, EnBW and Vattenfall, since 2020 also joined by LEAG, formerly part of Vattenfall. The “big four” also include numerous subsidiaries and are involved in various segments of the energy value chain, including generation, distribution, and retail of electricity and gas.[2]

While the German energy market serves industry, business, and residential customers, we will focus in this study on the B2C segment: In 2022, end consumers could choose on average from 157 providers, without taking corporate connections into account. For the household customer segment, the nationwide average was 136 providers.[3]

Shift to renewable energies and new technologies

The German energy market has undergone significant changes and developments in the last decade, driven by various political programs and legal acts. The expansion of renewable energy, the phase-out of nuclear power, energy efficiency measures, market liberalization, digitalization, and the rise of decentralization have all shaped the market landscape. These developments reflect Germany's commitment to a sustainable and diversified energy sector, as well as its efforts to meet climate goals and ensure a reliable energy supply for its citizens

“Energiewende“ for renewable energy expansion

The “Energiewende”, which translates to "energy transition", is a key national policy, that has shaped the market by promoting the expansion of renewable energy sources and reducing greenhouse gas emissions.[4]

Since 2010, Germany has made substantial progress in expanding renewable energy capacity, particularly in wind and solar power. This has led to a significant increase in the share of renewable energy in the overall energy mix, with contributing more than 50% in 2023 for the first time.

[2] Compare https://de.wikipedia.org/wiki/Die_gro%C3%9Fen_Vier_(Energieversorgung).

[3] See Monitoring report Bundesnetzagentur, Nov 2023, p.28.

[4] This policy iss supported by legal texts such as the Renewable Energy Sources Act (EEG) and the Energy Industry Act (EnWG).

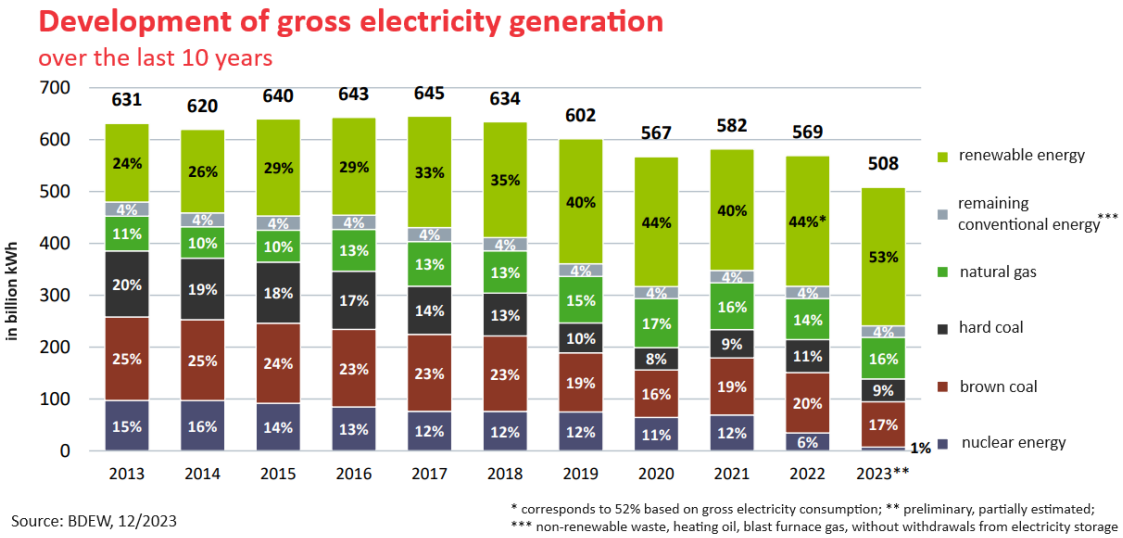


Figure 3: Development of gross electricity generation in Germany over the last 10 years

The goal is to transform the German electricity supply to almost climate-neutral, i.e. almost entirely provided by renewable energies and green hydrogen. The International Energy Agency IEA has highlighted that the electricity sectors of industrialized countries must be climate neutral by 2035 to achieve the 1.5° target.

To facilitate this development, not only industrial power generation but also the private sector has been addressed: Removing of legal barriers, tax relief measures and financial incentives have made rooftop photovoltaic systems a common occurrence not only for detached houses but also for tenement housing.[5]

Nuclear phase-out

In the aftermath of the Fukushima nuclear disaster in 2011, Germany made the decision to phase out nuclear power completely, as stated in the Atomic Energy Act (AtG). This decision led to the closure of several nuclear power plants -the last one in April 2023 - and supported the shift towards renewable energy sources, as is evident in the graph above.

Energy efficiency measures

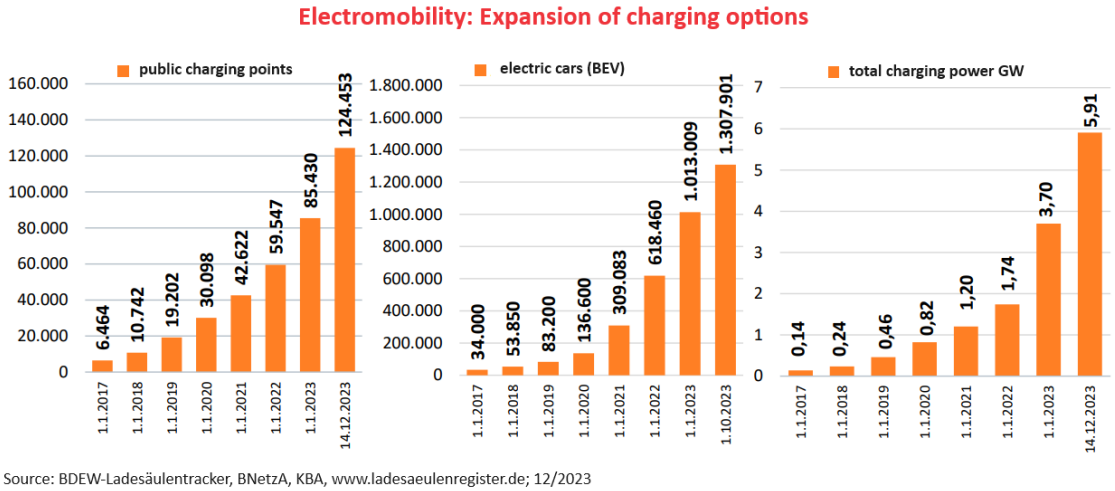
Energy efficiency has become a key focus in the German energy market. The government has implemented various measures to promote energy-saving practices in industries, buildings, and households. These include energy efficiency standards, financial incentives, and awareness campaigns. The aim is to reduce energy consumption, lower greenhouse gas emissions, and enhance the overall sustainability of the energy sector.

Electromobility

With the goal to stop the climate change, legal frameworks promoted the transformation and the increasing popularity of electromobility, specifically the growing adoption of electric cars. This shift in transportation has led to an accelerated demand for energy and new electricity distribution, as the implementation of a charging

[5] For details see the photovoltaics strategy of the German ministry of commerce and climate protection, May 2023: https://www.bmwk.de/Redaktion/DE/Publikationen/Energie/photovoltaik-stategie-2023.html

network across the country has become necessary as well.[6] The figure below shows the increase of public charging points going together with the rising number of battery electric vehicles in Germany, growing by an exponential rate.



Source: BDEW-Ladesäulentracker, BNetzA, KBA, www.ladesaeulenregister.de; 12/2023

Figure 4: Expansion of charging options in electromobility in Germany

Digitalization and smart grids

Advancements in digital technology were the prerequisite for the introduction of smart grids, which enable the integration of renewable energy sources and the efficient management of energy distribution. Smart meters and digital energy management systems have also become more prevalent, allowing consumers to monitor and optimize their energy consumption. With the new smart-meter law, a fixed timetable for the installation and widespread distribution of smart meters is set: From 2025 all consumers with an installed capacity of 6,000 to 100,000 kWh per year or more will be subject to mandatory installation, leading to equipment with an intelligent measuring system of least 95 percent by the end of 2030.[7]

Decentralization and prosumerism

Another trend in the German energy market has been the rise of decentralization and prosumerism. Increasing numbers of individuals and businesses are becoming energy producers through the installation of solar panels and other renewable energy systems. This shift towards decentralized energy production has led to the emergence of new business models, such as community energy projects and peer-to-peer energy trading.

[6] The Federal Emission Control Act (BImSchG) sets emission standards for vehicles, while the Energy Industry Act (EnWG) regulates the energy market and promotes the integration of renewable energy sources. Additionally, the Charging Infrastructure for Electric Vehicles Act (LISG) aims to expand the charging network across the country.

[7] Compare press notification on Smart Meter Law of the German ministry of commerce and climate protection, May 2023: https://www.bmwk.de/Redaktion/DE/Pressemitteilungen/2023/05/20230512-smart-meter-gesetz-final-beschlossen.html.

User experience of German households

German household customers have one of the following electricity supply contracts:

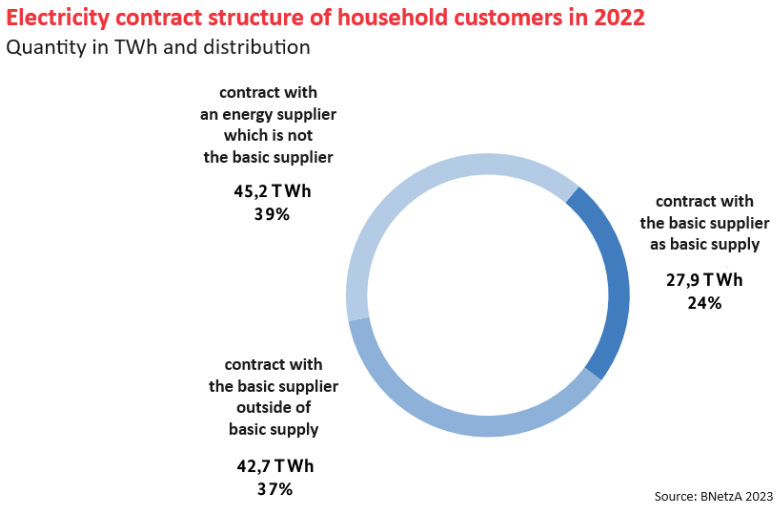


Figure 5: Distribution of electricity contracts of households in Germany

As shown in the graph above, most household customers still buy their electricity from their local providers[8]. That means, after 20 years of liberalization, only around 40% of the energy for private households is bought from energy suppliers outside of their local network area.

Contracts are differing in terms and obligations of the suppliers and the prices, but usually with monthly installments and annual statements, with an initial contract term of 12 months.

In addition to the overall price, contracts outside of the basic supply can have several other features that help suppliers compete for customers.

- There can be features that offer security either to the customer such as price stability guarantee or to the supplier, e.g. advance payment, minimum contract term.
- Also, many suppliers offer special bonuses at sign-up or for staying for another term.
- To promote efficient or renewable energy, green or eco tariffs are offered which redeem green electricity labels for the provided energy volume.
- Lately, dynamic tariffs which invoice spot market prices for energy, have gained increasing market shares. They require smart meters and aim to let the consumer participate directly from the volatile energy markets.

Contract changes and supplier changes are therefore an opportunity to influence the electricity price and the contract modalities.

[8] Energy supply is ensured by German law §36 EnWG, StromGVV, GasGVV. For volume shares of basic supplies, see Monitoring report Bundesnetzagentur, Nov 2023, p.28, see also graphs in p. 167.

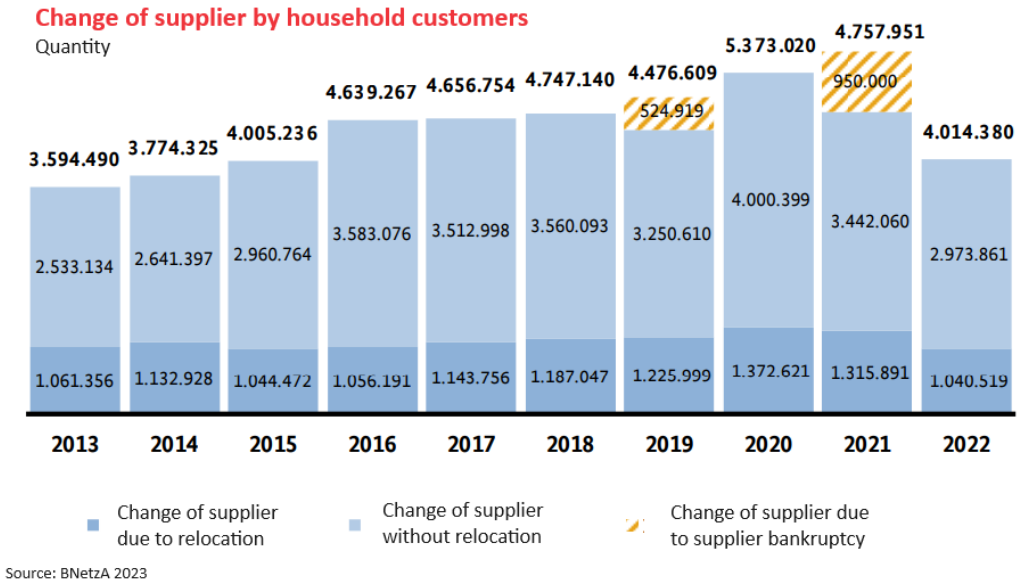


Figure 6: Change of supplier by household customers

With the liberalization of the energy markets, more and more customers have changed their energy provider for other reasons than relocating, going from 2.5 million in 2013 to 4 million in 2020.

Due to energy prices increasing by over 50%[9], the number of bankruptcies doubled in 2021 compared to the years before, resulting in nearly 15% less supplier changes in 2022.

Consumers became more careful and usually consult internet platforms like Check24 or Verivox[10], to compare prices and conditions for their contracts.

Exchanging opinions and commenting on suppliers' services on rating sites like Trustpilot was considered more important, helped along by increasing online activities in Germany during the lock-down periods in the Covid pandemic.

[9] Compare Groenveld, Josh: Energiekrise sorgt für Pleitewelle bei Strom- und Gasanbietern, Business Insider Jan 2022: https://www.businessinsider.de/politik/deutschland/fast-doppelt-so-viele-unternehmen-insolvent-wie-in-vergangenheit-energiekrise-sorgt-fuer-pleitewelle-von-strom-und-gasanbietern.

[10] See https://www.check24.de/strom/ or https://www.verivox.de/strom/.

Understanding and manipulation of data

Dataset from Trustpilot

Target of this analysis are ratings, customer votes and supplier feedback of energy suppliers in Germany on the Trustpilot web site[11]. The source for our investigation is **scraped content from September 2023 in German**, as the language settings on the portal define the respective market. We tried out English language as well but then we only got UK / American energy suppliers, or comments in English language only from English speaking expats currently living in Germany, which is only a small percentage of the local customer base.

We scraped content on two levels: File 1 contains general information of energy suppliers in Germany (supplier level), while File 2 collects for each energy supplier customer reviews and supplier answers (customer level). The total dataset may be obtained by joining both files on the supplier column.

File 1 - Energy suppliers in Germany

Web scraping

At the time of investigation, 37 distributors in Germany were listed on two pages on the Trustpilot web site for the category of 'power supply company' with 20 and 17 entries per page respectively.[12]

The standard sorting of the suppliers is "according to relevance". This filter showed all energy suppliers sorted by highest score and numbers of votes. In addition, the companies had to fulfill the following conditions, to ensure that the companies with the highest votes receive up-to-date customer feedback:

- The supplier needs to have received at least 25 evaluations in the last 12 months.
- The supplier must have the status "asking for evaluation".

We extracted the name, score, number of votes and business location of each energy distributor by using Beautiful Soup and URL Lib in addition to the classical pandas and NumPy packages.

On the page, the energy suppliers were all represented in a similar way, so we assumed logically that the structure of the respective web code was the same. The page was not set up as a table, but as a sequence of div containers, to whom specific key words in style or span tags helped identify the important information. Therefore, it was possible to iterate on the *ener_tp* variable of the first page, containing all the pertinent information on the energy suppliers. The information was then stored in lists, which could be transformed in a data frame with the zip function.

These steps could be repeated for the second page as well, stored into the *ener2_tp* variable.

Finally, both page data frames were concatenated to create one data frame, as shown in the next figure.

[11] See the following link: https://de.trustpilot.com/categories/electric_utility_company?

[12] See screenshots of the Trustpilot website in the chapter “Additional” in the appendix.

	supplier	location	score_votes	cat	comment
0	Octopus Energy Germany	München, Deutschland	4,8|8.392	Stromversorgungsunternehmen·Energieversorger·E...	https://de.trustpilot.com/review/octopusenergy.de
1	Ostrom	Berlin, Deutschland	4,8|1.607	Ökostromanbieter·Stromversorgungsunternehmen·E...	https://de.trustpilot.com/review/ostrom.de
2	Rabot Charge	Hamburg, Deutschland	4,3|176	Ökostromanbieter·Energieanbieter·Energieversor...	https://de.trustpilot.com/review/rabot-charge.de
3	MONTANA Group	Grünwald, Deutschland	4,0|3.153	Kraftstofflieferant·Energieanbieter·Stromverso...	https://de.trustpilot.com/review/montana-energ...
4	E.ON Energie Deutschland GmbH	München, Deutschland	3,7|13.467	Solartechnikanbieter·Energieanbieter·Stromvers...	https://de.trustpilot.com/review/eon.de

Figure 7: Newly scraped data set "Energy suppliers in Germany"

To make further processing easier, some basic cleaning and simple column transformations were applied:

- Separation of scores and votes in 2 separate columns
- Change of data types to float for ‘score’ column
- Overall change of decimal separator to English notation and removal of German thousand separator
- Change of data type to integer for ‘votes column.
- Split of location information into two columns *city* and *country*.

The final scraping result was stored with a new order of columns as a csv-file to be used for further investigation, as follows:

	supplier	city	country	cat	score	votes	comment
0	Octopus Energy Germany	München	Deutschland	Stromversorgungsunternehmen Energieversorger E...	4.8	8392	https://de.trustpilot.com/review/octopusenergy.de
1	Ostrom	Berlin	Deutschland	Ökostromanbieter Stromversorgungsunternehmen E...	4.8	1607	https://de.trustpilot.com/review/ostrom.de
2	Rabot Charge	Hamburg	Deutschland	Ökostromanbieter Energieanbieter Energieversor...	4.3	176	https://de.trustpilot.com/review/rabot-charge.de
3	MONTANA Group	Grünwald	Deutschland	Kraftstofflieferant Energieanbieter Stromverso...	4.0	3153	https://de.trustpilot.com/review/montana-energ...
4	E.ON Energie Deutschland GmbH	München	Deutschland	Solartechnikanbieter Energieanbieter Stromvers...	3.7	13467	https://de.trustpilot.com/review/eon.de

Figure 8: Final scraped data set “Energy suppliers in Germany”

Target and explanatory variables

Although data types were looking okay, it was clear that some values were missing or needed to be checked for clarification.

- If there were 0 votes, the data row was to be deleted after relevance clarification.
- For missing city information valid company information was researched and replaced accordingly.

After a reset of index and uneventful check for special chars, the data was deemed clean.

Features and Limitations

The target of this exploration is the rating and comments of energy suppliers on the German market, where some companies use specifications which are used as synonyms in the German language, e.g. electricity company and power supplier.

The Trustpilot search was for “energy supplier”, i.e. “Stromversorgungsunternehmen” in German, but with the result it became clear that some companies also render additional services and are listed in several categories.

This information was consolidated in the *cat* column: To access its content, all information was exported to text for a comprehensive list of unique categories, which were then added as separate columns into the data frame according to the following strategy:

Finding	Action	Key words in category and translations
Synonyms for energy suppliers	To be consolidate into one common category	"Energieversorger", "Energieanbieter", "Energieversorgungsunternehmen", "Stromversorgungsunternehmen", "Stadtwerke“
Specialized power labels	Filter in separate columns	• 'eco energy' for 'Ökostromanbieter' • 'solar energy' for 'Solarenergieunternehmen' • 'heat flow' for 'Wärmeenergie-Unternehmen'
Diversification labels	Filter in separate columns	• 'gas supplier' for 'Gasversorgungsunternehmen' • 'fuel supplier' for 'Mineralölunternehmen', 'Kraftstofflieferant', • 'water supplier' for 'Wasserversorgungsunternehmen' • 'telecommunications provider' for 'Telekommunikationsanbieter', 'Internetanbieter', 'Telefon- und Internetdienst' • 'energy solutions' for 'Energieanlagen und -lösungen', 'Solartechnikanbieter', 'Heizungsanlagenanbieter', 'Anbieter von Elektronikbauteilen', 'Technischer Kundendienst', 'Elektronikunternehmen' • virtual' for 'Reiseanbieter', 'Online-Marktplatz'

Table 1: Categories for power labels and diversification

After critical review, *energy* as identical common entry for all suppliers was deleted. Also *heat* and *solar* were removed due to representing a technical subcategory of *eco* energy. Other labels were summarized, and single entries removed. The data set was now ready for exploration and visualization:

	supplier	city	eco	gas	telco	energy_solutions	num_votes	score
0	Octopus Energy Germany	München	1	1	0	0	8042	4.8
1	Ostrom	Berlin	1	0	0	0	1598	4.8
2	Rabot Charge	Hamburg	1	0	0	0	174	4.3
3	MONTANA Group	Grünwald	1	1	0	0	3146	4.0
4	E.ON Energie Deutschland GmbH	München	1	1	0	1	13223	3.7

Figure 9: Data Set “Energy suppliers in Germany” after Clean-up and Feature Engineering

File 2 – Customer reviews and supplier answers

Web scraping

The reader may have a look at Figure 34 and Figure 35 in the appendix to see a typical customer feedback and an optional supplier response on Trustpilot. The number of reviews for each energy supplier varies widely, from less than 10 to over 10.000. For each review, there is either none or one supplier response. For a fixed energy supplier, the reviews are distributed over several pages such that each page contains exactly 20 reviews except the last page, where review numbers from 1 to 20 are possible. In the scraping phase we tried to get as much data as possible from customer reviews and supplier answers, i.e. we wanted to gather for each German energy supplier all reviews and answers there were (September 2023). This could be accomplished using Beautiful Soup and the following methodology.

From the supplier column in File 1 we generated for each supplier their respective start page URL in Trustpilot. The list of start page URL’s is used to iterate over in a for-loop. On the respective start pages one can read out the total number of review pages (per supplier). A second sub-for-loop now iterates over all pages and reads out all review information and answers and stores them in respective lists. These lists are concatenated to create the data set File 2. In practice, Trustpilot noted that we are scrapers and denied access to the page each

time after scraping around 300 pages. We had to take measures to disguise ourselves as casual users, including random time delays during iterations and rotating User-Agents. This improved the situation to scrape up to 700 pages without getting blocked. Still, this is not enough for big suppliers like Octopus Energy with over 1000 pages, so we had to generalize the algorithm to allow for selection of iteration intervals manually. The obtained sections could be concatenated to File 2 containing all possible customer reviews and supplier answers of German energy suppliers there were September 2023. The anti-blocking measures could have been improved by including e.g. rotating Proxies, but the pragmatic manual section-wise approach got us to the goal faster.

Target and explanatory variables

The following features were directly scraped from customer reviews and supplier answers. See Figure 34 and Figure 35 in the appendix for screenshots of typical reviews and answers on Trustpilot.

Label	Description of variable	Appearance	Data type
Nickname	Customer nickname	*optional*	*object*
Location	Location of customer	*mandatory*	*object*
Stars	Star rating of customer	*mandatory*	*Int64*
Headline	Headline of post	*mandatory*	*object*
DoP	Date of post	*mandatory*	*datetime64[ns, UTC]*
DoE	Date of experience	*mandatory*	*datetime64[n]*
Comment	Comment of customer	*optional*	*object*
Answer	Answer of energy supplier	*optional*	*object*
DoA	Date of answer	*if, and only if answer exists*	*datetime64[ns, UTC]*
Company	Company name, i.e. energy supplier)	*mandatory*	*object*

Table 2: Feature description "Customer feedback and supplier answers"

After scraping we computed and included the following variables.

Label(s)	Description(s) of variable(s)	Data type
DoP.day, DoP.month, …	Splits of DoP, DoE, DoA into day, month, year	*int64*
Comment_TF	Is there a comment? (value: 1 or 0)	*int64*
Answer_TF	Is there an answer? (value: 1 or 0)	*int64*
Words_Headline	Number of words of headline	*int64*
Words_Comment	Number of words of comment	*int64*
Words_Answer	Number of words of answer	*int64*
Response_time	Response time of energy supplier to review in days (DoA-DoP)	*float64*

Table 3: Newly engineered features for data set 2

The two target variables are the star ratings *Stars* and the supplier answers *Answer*, which will be predicted by comments and related variables. In a first approach, instead of predicting answers, we will predict answer length *Words_Answer*.

Features and Limitations

Some variables of the data set may be negligible, e.g. location, which carries the value “DE” in 98% of cases or user nicknames, which are hard to classify, e.g. for gender. Other explanatory variables, e.g. key words of comments to predict star ratings, are still left to be derived.

Visualizations and data exploration

File 1 - Energy suppliers in Germany

Relations between explanatory variables and target

To investigate the relationship between the variables of the data set, both Spearman and Pearson tests have been conducted and visualized by a heat map: Red colors indicate a high positive correlation between the variables, whereas blue colors show a poor and even negative correlation, as shown below.

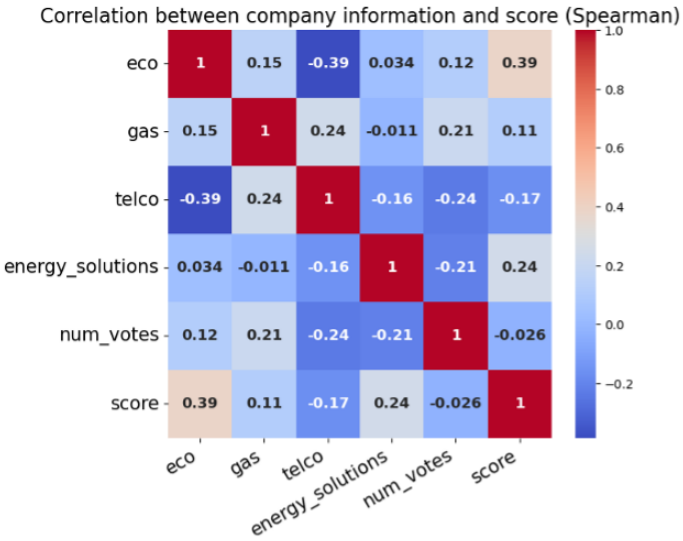


Figure 10: Heat map (Spearman r) for data set 1

It has been observed that eco-friendly energy seems positively correlated to the rating score. Diversification in energy solutions and gas supply is also represented as positive, but number of votes slightly negative.

So, we can postulate that that the supported categories have no direct impact on the overall ranking. Checking with a Spearman-r test, it could also be concluded by a p-value of 0.887 that the number of votes and the score are not correlated as well.

Indeed, if we look at the distribution of scores and the number of votes vs. the rating score, we can see that for every rating there are suppliers with low and high number of votes. Even if there are some extreme values for the middle scores, most supplier ratings are based on less than 500 votes (Figure 11).

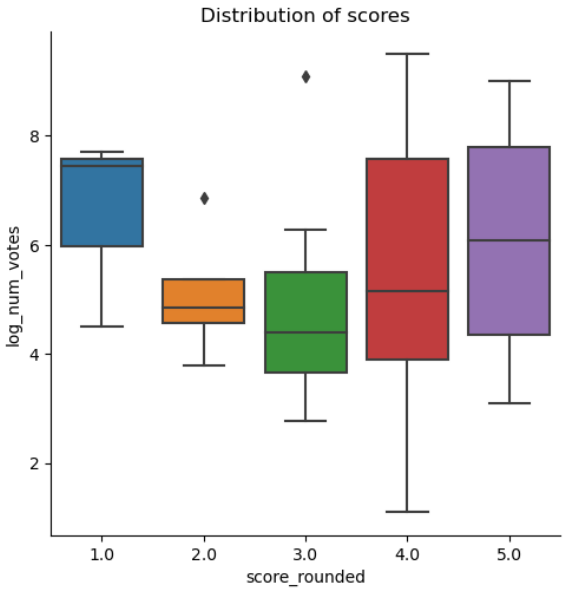


Figure 11: Distribution of number of votes per score for German Energy suppliers on Trustpilot

Although we see limited data points, we can observe a tendency of higher number of votes for low ranking and higher ranking. An explanation might be that customers tend to express more feedback when they are exceptionally happy or exceptionally unhappy.

To consider is also the history as some suppliers, like e.g. Vattenfall, have been in business very long and their average score is changing over time.

Overall ratings presented as the following energy provider as the top and bottom five in the Trustpilot ranking:

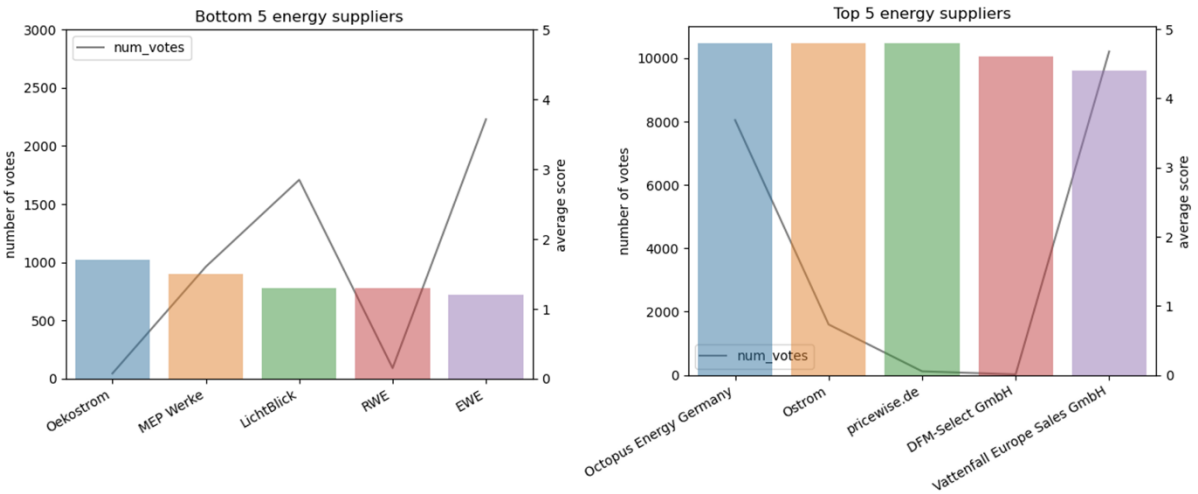


Figure 12: Bottom 5 and Top 5 energy suppliers in Germany

As we can see, the number of votes is spread all over the spectrum: Small suppliers with only a few comments, like Ostrom or pricewise, are neck-and-neck as top performers with big international corporations like Octopus Energy or Vattenfall with many votes. While new eco-oriented suppliers like LichtBlick have nearly as many votes as traditional suppliers such as EWE with the same low average score, we can also see the opposite: new supplier Oekostrom and traditional RWE with only a few votes and equally low ranking.

Relationships between variables

Having a look at the diversified offering mentioned before[13] we can see the following situation:

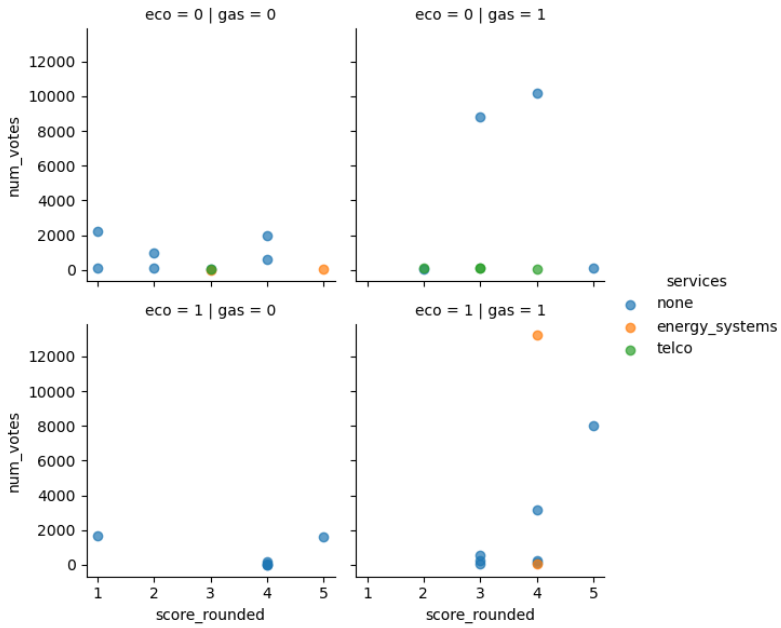


Figure 13: Impact of diversification on energy suppliers in Germany

The offering of eco power is no guarantee for many (positive) votes or a high rating. However, suppliers who are also delivering gas, tend to have more votes but are distribute over scores 2 to 5.

In the overall ranking, telecommunication has no impact, but suppliers also offering energy systems like photovoltaic technologies tend to have higher ranking.

Could this be an indication, that functioning business processes and customer orientation are the most important factors for a good score? We should analyze the comments for feedback.

[13] See classification in chapter “Features and Limitations” for file 1.

File 2 – Customer reviews and supplier answers

Overview of user activity and mood

The number of comments on German energy suppliers in Trustpilot saw a strong increase in the recent years, see Figure 14. The threshold of more than 1500 comments per year was first reached 2019, reaching 25,000 comments by 2023. The website saw a strong boost in the pandemic years 2020-2022, as more people were forced to go digital. One can expect that the war in Ukraine increased the number of comments in 2023 additionally, as a lot of people were affected by the resulting energy crisis. The true number of comments at the end of 2023 is expected to be higher, as the figure on the right is to date September 2023. The users are mostly from Germany (98.1%), followed by Austria (0.3%), Netherlands (0.2%), US (0.2%) and Spain (0.1%). In total, users state to be from 78 different countries. One can expect that a portion of country labels stem from accidental miss clicks or are wrong by choice to protect customer privacy.

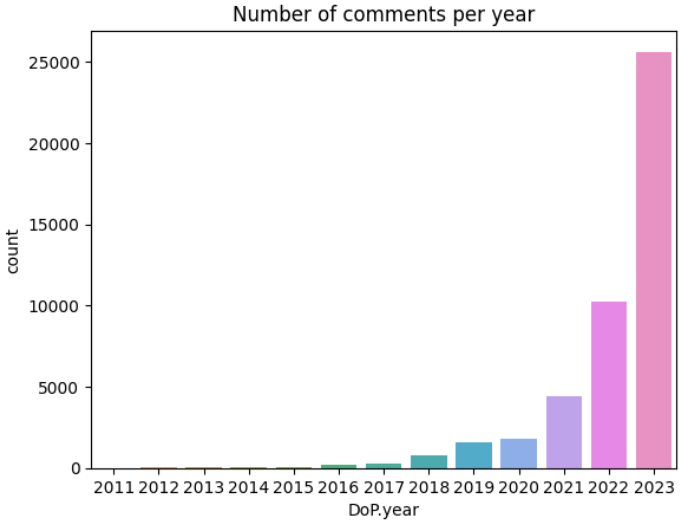


Figure 14: Number of comments on German energy suppliers to date September 2023

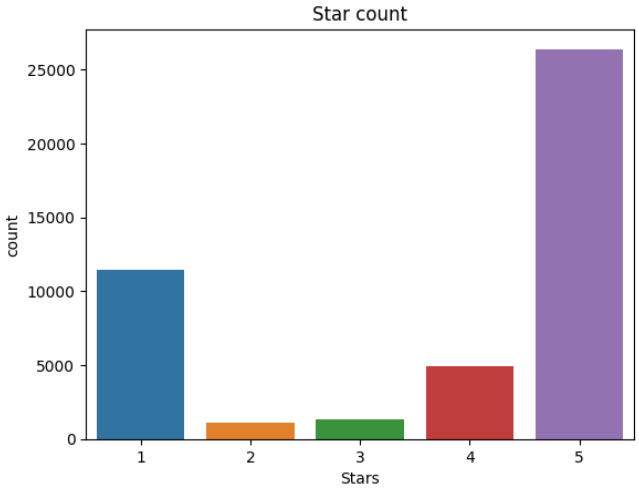


Figure 15: Count of star ratings

Figure 15 shows counts of the five possible star ratings on German energy suppliers. Customers tend to review only if their experience is on the extremes, i.e. bad (1 star) or great or excellent (4 or 5 stars).[14] In particular, 69% of reviews are great (11%) or excellent (58%), 25% of reviews are bad, and 6% of reviews are poor or average.

[14] For an explanation oft he star rating see: https://support.trustpilot.com/hc/en-us/articles/201748946-TrustScore-and-star-rating-explained#:~:text=A%20TrustScore%20is%20the%20overall,how%20they're%20calculated%20here.

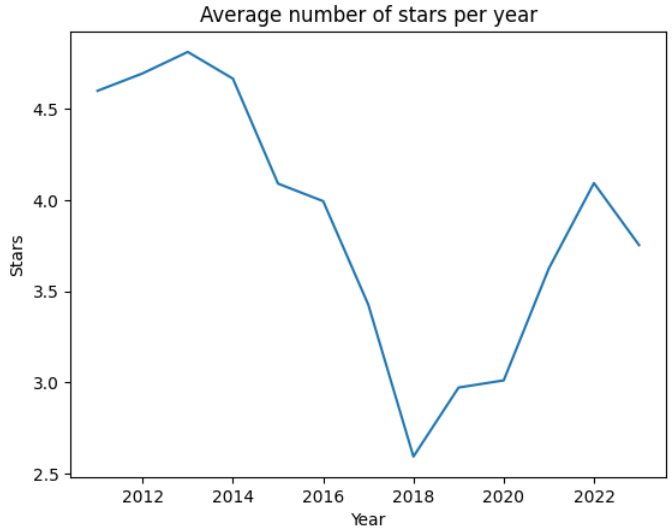


Figure 16: Average star rating per year

The average number of stars for German energy suppliers varies drastically over the years from 2011 to 2023, as shown in Figure 16. Starting from the global maximum of 4.8 stars in 2013 we see a steep downward trend to the global minimum of 2.6 stars in 2018. From here a steady upward trend to 2 stars in 2022, followed by a dip to 3.7 stars onto 2023. The dip from 2022 to 2023 is in line with the energy crisis. The pandemic years 2020-2022 seem to not bother customers negatively. In contrary, a lot of customers seem to have found time to write positive reviews. The liberalization of the German energy market 2019[15] led to a reduction of monopoles which is in favor of customers. A comparison with the number of comments per year, i.e. Figure 14, shows that the average number of stars before 2019 is rather insignificant, due to low number of reviews. Special effects should be considered to explain the trend from 2013 to 2018. When the website was new and unknown, workers of registered companies could dominate the reviews leading to high ratings. As more external customers entered the platform, more realistic reviews were established. Another effect is high volatility, i.e. early companies could potentially easily dominate the data, as so few data was amenable.

Relationship between variables

Pearson correlations of selected numerical variables are visualized in Figure 17. The year of experience is almost completely uncorrelated to any of the variables ($|r_y| < 0.11$). The number of words of a comment and the number of words of the answer are good indicators for the number of stars ($r_{cs}=-0.49$, $r_{as}=-0.55$). Lesser words mean more stars. The number of words of the headline is slightly correlated to the number of stars ($r_{hs}=-0.17$). The number of words of comments and of answers is correlated by $r_{ca}=0.31$. If customers write more text then companies tend to answer more elaborately. The response time of the company is almost completely uncorrelated to any of the variables

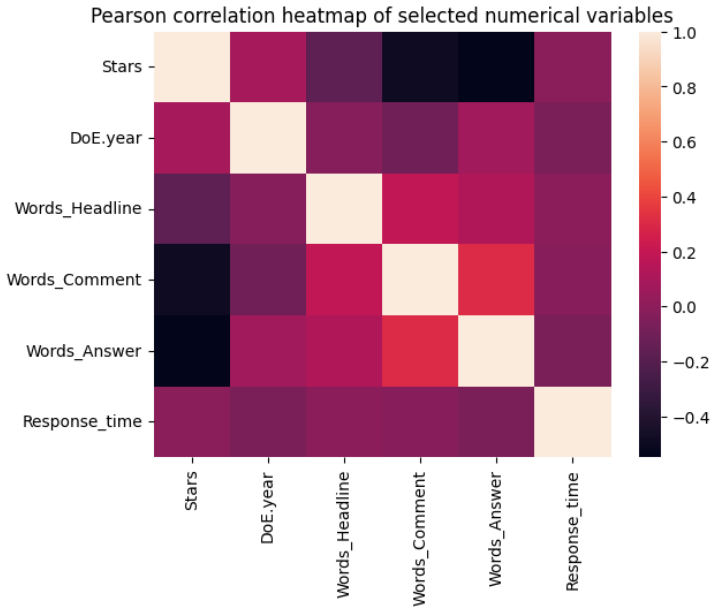


Figure 17: Correlation heatmap

[15] See: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://static.agora-energiewende.de/fileadmin/Projekte/2019/Liberalisation_Power_Market/Liberalisation_Electricity_Markets_Germany_V1-0.pdf

(|r_rt| < 0.06). This means that a company will answer good and bad reviews in around the same time frame. The year of experience is almost completely uncorrelated to any of the variables (|r_y| < 0.11).

Analyzing comments and answers

In average, the number of words of a comment is a very strong indicator for the number of stars (r_cs=-0.98) as well as the number of words of an answer (r_as=-0.90), see Figure 21. However, the distributions of comment-word count per star rating overlap. The variance of word count per comment is higher for lower ratings, the

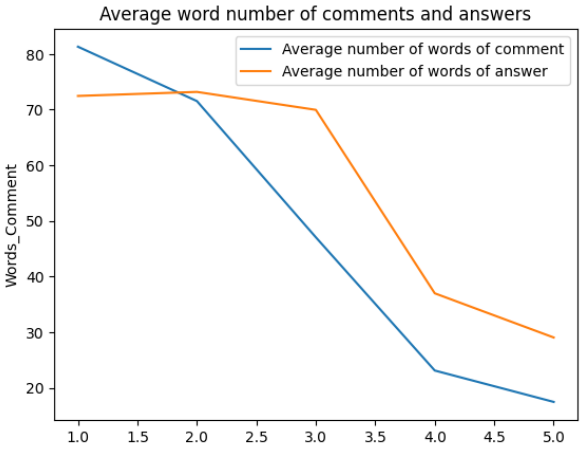


Figure 21: Average word count of comments and answers

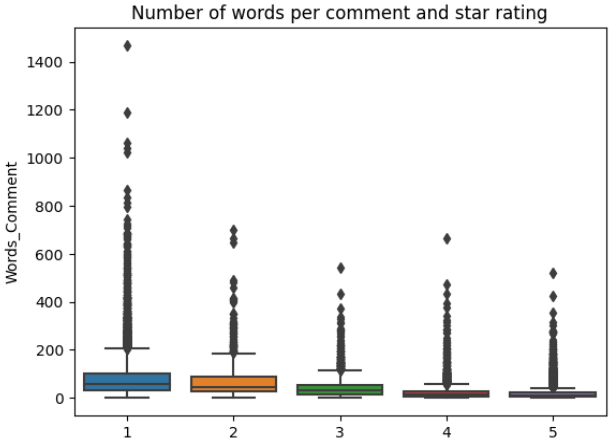


Figure 20: Boxplots: Word counts of comments per star rating

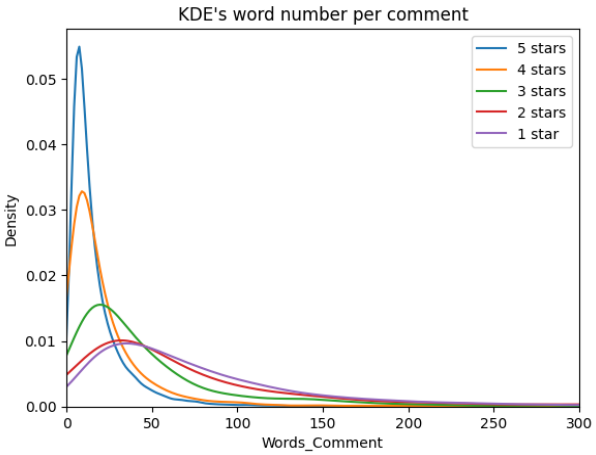


Figure 18: KDE: Word count of comments per star rating

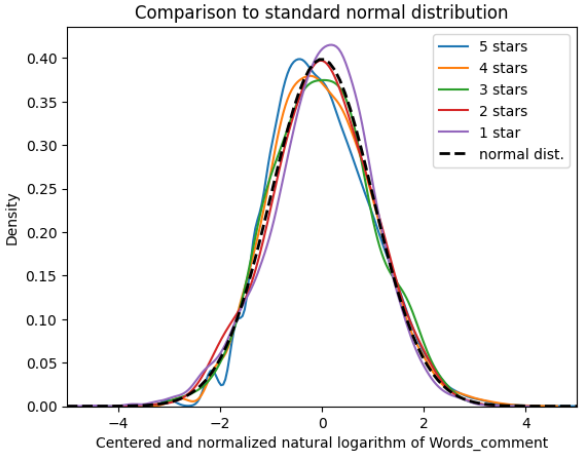


Figure 19: Word count of comments is lognormal distributed

most extreme outliers being for a star rating of 1. The median and quantiles tend to shrink for higher star ratings (Figure 20). The Kernel Density Estimations (KDE's) have higher variance and skew to the right for lower star ratings (Figure 18). This is a strong hint that word counts of comments are lognormal distributed. Indeed, after taking the natural logarithm, centering and normalizing, the data follows a standard normal distribution, as seen in Figure 19. Regarding word counts of answers the situation is not that straightforward. Two peaks for each rating in Figure 22 indicate that companies tend to have standard answers for each rating and that the statistic is dominated by the two companies with the most comments. Indeed, separate answer counts for the

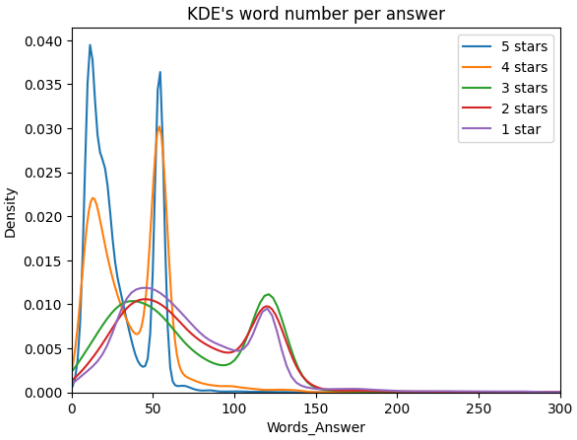


Figure 22: Word count of company answers per star rating

two biggest players “E.ON Energy” and “Octopus Energy”, Figure 23 and Figure 24, support the claim. A second peak in the 5-star rating of “Octopus Energy” may be explained by a change in the policy of standard answer length.

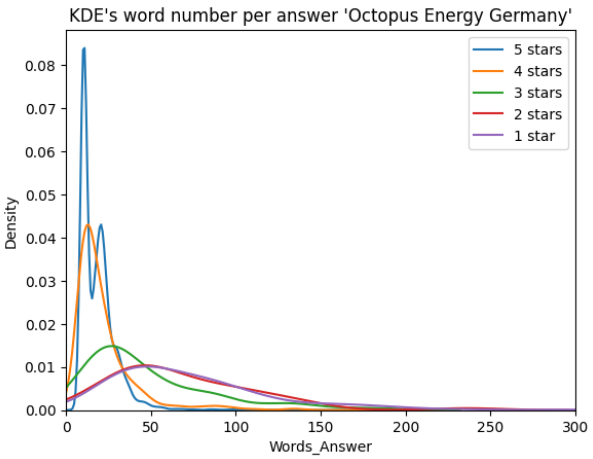


Figure 24: Word count of Octopus Energy's answers

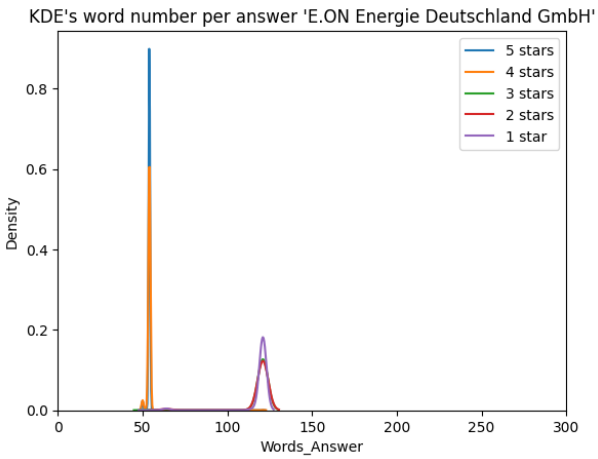


Figure 23: Word count of E.ON Energy's answers

Data selection and preparation for Machine Learning

Feature selection

For Machine Learning, the focus was on File 2, containing customer reviews and supplier answers. We started from the data set as explored and explained above[16]. The goal was

1. to predict the star rating of customer posts from customer comments and
2. to predict the length of company answers to customer posts, respectively the full answers.

For the first case it was necessary to restrict the data set to rows, where a comment existed, i.e. *Comment_TF* = 1, for the second case, in addition also answers were required, i.e. *Comment_TF* = 1 & *Answer_TF* = 1.
Due to unique answer policies applied by some companies, we select specific companies by the variable *Company*.

Feature preparation

We applied the natural logarithm on the features word counts of comments, headlines and answers, building the engineered variables *log_Words_Comment*, *log_Words_Headline*, *log_Words_Answer*.

Furthermore, the min-max-scaler was applied to the target variable *Stars* creating the variable *Stars_min_max_scaled*.

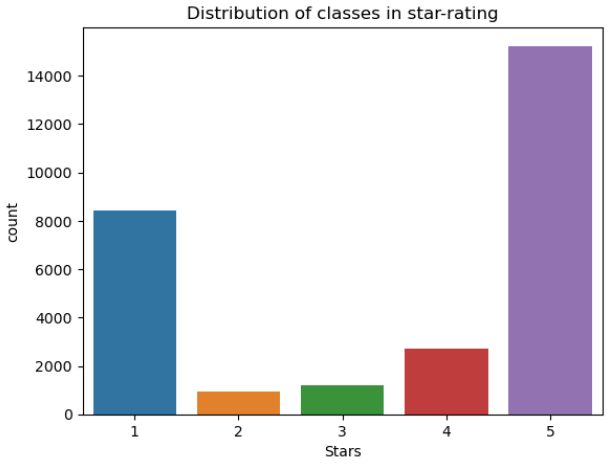


Figure 25: Class distribution in target variable

As seen in the chart above, the target variable 'Stars' is not evenly distributed in the dataset: Over 50% of the customer comments had 5-star ratings, followed by around 30% of 1-star ratings and around 10% of 4-star labels. Less than 5% each went to 3 and 2-star ratings.

Following the classical use case in business reports and analysis papers, where the focus is usually on the very best and the very worst, to define controls and measures for improvement, we grouped the star ratings into 'good' (*Stars* >= 4) and 'bad' (*Stars* < 4) ratings[17], consolidated in the Boolean variable *Stars_geq4_TF.*

[16] For details see p. 14f.

[17] Alternative solution would be to remove ratings 2-4 stars for a perfect training set, but the platform would deliver continuously new ratings with 2 to 4 stars, which makes this approach too theoretical.

Dataset for machine learning

The following features were used for testing our ML models, a subset only for keyword analysis.

Label	Description	Appearance	Modelling Case 1	Modelling Case 2	Keyword Analysis
Headline	Headline of post	*object*	explanatory	explanatory	explanatory
Comment	Comment of post	*object*	explanatory	explanatory	explanatory
Answer	Answer to post	*object*	not used	target	explanatory
Company	Name of Company	*object*	selector	selector	selector
Comment_TF	Checks, if there is a comment, Boolean	*Int64*	= 1	= 1	selector
Answer_TF	Checks, if there is an answer, Boolean	*Int64*	not used	= 1	selector
log_Words_Headline	Natural logarithm of word count of headline	*float64*	explanatory	explanatory	not used
log_Words_Comment	Natural logarithm of word count of comment	*float64*	explanatory	explanatory	not used
log_Words_Answer	Natural logarithm of word count of Answer	*float64*	not used	target	not used
Stars	Rating number of stars 1 - 5	*Int64*	target	not used	selector
Stars_min_max_scaled	Min-max-scaled star rating	*float64*	not used	explanatory	not used
Stars_geq4_TF	Checks for Stars > 3, Boolean.	*Int64*	target	explanatory	not used

Table 4: Feature description ML dataset

Outlier handling

In the distribution of the main variables, there were some data points with extreme values, which were investigated with KDE and box plots, looking into distribution and interquartile range. The most extreme values were capped.[18]

Dataset size and selection

To develop the most practical and appropriate approach, simple models were applied first on the complete data set, but due to long run times and differences in suppliers' answer policies, the following subsets were built, dividing the dataset down to a more manageable size while preserving logical cohesion.

Figure 26: Refinement of dataset size

This method corresponds to business practices where the target is usually limited to the company's own customer feedback, or a limited number of competitors in a benchmark analysis.

Other methods to reduce the dataset into more manageable portions could have been:

- Selection by date, e.g. only data younger than 2 years
- Limitation to a fixed line count supported by star ratings to maintain target class distribution.

[18] In most cases we had skewed distributions, so we looked at the inter-quartile range (IQR): Data points below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR can be considered as outliers. See also Dey, Akash: How to handle outliers, Feb 2022: https://www.kaggle.com/code/aimack/how-to-handle-outliers/notebook.

Rating prediction based on customer feedback

Modeling problem and performance metric

As a target of this ML problem, we will try to predict the star rating based on customer feedback, as available and engineered in the steps described earlier.

The rating prediction based on customer feedback was a **classification problem**, with **accuracy** and **F1-score** as the **main performance metric**.

"Precision measures the extent of error caused by False Positives (FPs) whereas recall measures the extent of error caused by False Negatives (FNs)."[19] Recall measures the model's ability to detect positive samples, which does not consider other classes. Even though precision includes all classes for measuring how reliable the model is classifying one class, both metrics are focusing on one class of interest, and we must consider here more than one class.

Accuracy describes how the model performs over all classes as it calculates the ratio between the number of correct predictions to the total number of predictions. In other words, the focus is here on penalizing the FPs and FNs.[20] In our case, we need to concentrate on True Positives (TPs) and True Negatives (TNs), so we decided on accuracy as main performance metric but also not discarding F1-score for comparing different models. "F1-score is balancing precision and recall on the positive class while accuracy looks at correctly classified observations of both positive and negative."[21]

Model choice and optimization

Support vector machine (SVM) from scikit-learn ensemble was chosen first because we have limited features, and we were looking for a simple but effective model which would also support multi-class classification. Preparation consisted only of standardization and the data set must only consist of numerical features.[22]

As second model we tried RandomForestClassifier also from scikit-learn ensemble which is a meta estimator that fits several decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. With this model we kept all numerical features and did not need scaling, as it is a tree-based model.[23]

---

[19] See Zeya LT: Essential things you need to know about F1-score, Towards Data Science, Nov 2021: https://towardsdatascience.com/essential-things-you-need-to-know-about-f1-score-dbd973bf1a3

[20] Compare Huilgol, Purva: Accuracy vs. F1-Score, Medium Aug 2019: https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2.

[21] See Czakob, Jakub: F1 Score vs ROC AUC vs Accuracy vs PR AUCH: Which Evaluation Metric Should I Choose?, Neptune.ai Blog Sep 2023: https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc#:~:text=F1%20score%20vs%20Accuracy,observations%20both%20positive%20and%20negative

[22] Compare user guide support vector machines: https://scikit-learn.org/stable/modules/svm.html; cited as Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[23] Compare Sklearn documentation on RandomForestClassifier: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

Model training and observations

Applying train_test_split from sklearn.model_selection we specified a test set of 20%.

The first simple ML models applied for rating predictions for all 5-star ratings showed poor results since we faced a multiclass qualification problem with an imbalanced data set. Neither application of hyperparameter tuning and cross validation did improve the results, nor adding more features for model training.
In addition, run times were very long and handling of the code impractical, due to the initially very large size of the data set.

Model	Dataset	Performance metric before tuning		Performance metric after tuning[24]	
		Accuracy	F1-score	Accuracy	F1-score
SVM Classifier	All suppliers, only data with answers; *Words_Comment*	0.69 No overfitting	Class 1: 0.69 Class 2 to 4: 0.00 Class 5: 0.79	0.59 No overfitting	Class 1: 0.65 Class 2: 0.10 Class 3: 0.09 Class 4: 0.03 Class 5: 0.76
SVM Classifier	All suppliers, only data with answers; *Words_Comment, Words_Headline*	0.70 No overfitting	Class 1: 0.69 Class 2 to 4: 0.00 Class 5: 0.79	0.54 No overfitting	Class 1: 0.66 Class 2: 0.09 Class 3: 0.10 Class 4: 0.13 Class 5: 0.70
Random Forest Classifier	All suppliers, but only data with answers; *Words_Comment, Words_Headline*	0.67 Slight overfitting	Class 1: 0.67 Class 2: 0.00 Class 3: 0.01 Class 4: 0.00 Class 5: 0.78	0.67 Notable overfitting	Class 1: 0.73 Class 2: 0.00 Class 3: 0.01 Class 4: 0.00 Class 5: 0.78

Table 5:Performance metrics for multi-class prediction of star ratings (simple models)

Hyperparameter tuning, cross validation and boot strapping were applied with the tuning step, in regard of coping with the multiclass classification and the imbalanced data set.

As we can see, the 1- and 5-star ratings were predicted best. With hyperparameter tuning we saw a small improvement on the middle ratings while dominant classes performed a little less. RandomForestClassifier did better while we encountered an overfitting issue with the SVM classifier models. However, overall performance was not much over 50% for the dominant ratings 1 and 5 which is very close to random reliability.

The greater the imbalance between classes, the less successful the classical models will be in predicting the minority class. Therefore, applying undersampling was considered to increase the number of observations of the minority classes, as the data set was sizable enough.
However, the focus for customer service and customer satisfaction KPIs lies in the extreme opinions: “Is the customer really happy or really dissatisfied?” With this guideline for improvement, the middle star ratings, which were quantitatively low, could be neglected.

[24] Tuning for imbalanced data set and multi-class decision for SVM. Tuning with GridSearchCV with 3 fold crossvalidation for RandomForestClassifier.

To improve performance, the following changes were implemented:

- The size of the dataset was decreased as explained above[25]. Following everyday use cases, the data set for prediction would most likely contain only customer feedback from one supplier.
- As star ratings were almost binary distributed, we simplified the modelling to a binary classification problem targeting the *Stars_geq4_TF* variable.
- Because we expect the star rating of the customer to be closely tied to the sentiment of the customer's comment, we apply a sentiment analysis to the explanatory variable *Comment*.

In a first step, model performance was checked with a reduced data set limited to E.ON customer feedback, also applying natural logarithm to the word count features before capping the extreme values. The distribution of classes between positive (1) and negative (0) classes was still imbalanced, with nearly 70 % of the customer comments deriving from 4 or 5-star ratings and 30% going to 1- to 3-star ratings.

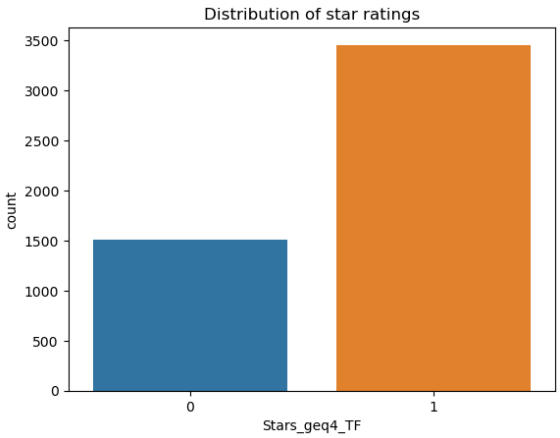


Figure 27: Class distribution in E.ON customer feedback

The limited and for one supplier specialized data was much easier to predict, reaching an accuracy around 80%.

Even when adding the information from word count of headline to the comment feature, we could see only small improvement for the non-dominant class 0 to the detriment of class 1, as the imbalance could not be overcome. Again, the SVM classifier showed the best result, while we saw some overfitting with the Random Forest model.

Model	Dataset	Performance metric before tuning		Performance metric after tuning[26]	
		accuracy	F1-score	accuracy	F1-score
SVM Classifier	E.ON all data; *log_Words_Comment*	0.79 No overfitting	Class 0: 0.65 Class 1: 0.88	0.79 No overfitting	Class 1: 0.69 Class 5: 0.84
SVM Classifier	E.ON all data; *log_Words_Comment, Log_Words_Headline*	0.82 No overfitting	Class 0: 0.63 Class 1: 0.88	0.79 No overfitting	Class 1: 0.70 Class 5: 0.84
Random Forest Classifier	E.ON all data; *log_Words_Comment, Log_Words_Headline*	0.76 Slight overfitting	Class 0: 0.59 Class 1: 0.84	0.77 Slight overfitting	Class 1: 0.60 Class 5: 0.84

Table 6:Performance metrics for binary prediction of star ratings (simple models)

[25] See description and figure 2 in chapter Dataset size and selection.

[26] Tuning for imbalanced data set and multi-class decision for SVM. Tuning with GridSearchCV with 3 fold crossvalidation for RandomForestClassifier.

Sentiment analysis

Most customers included in *Headline* only a few words or a summary sentence but wrote extensively about problems and good experiences in the *Comment* field. So, we chose the feature *Comment* to undergo sentiment analysis because it contained more detailed and therefore more meaningful content than the field *Headline* with its character limitation on the Trustpilot platform.

All texts in *Comment* were prepared with the following steps:

- With regex, all non-letters in the comments were replaced by spaces, consecutively removing words of length 2 or less.
- Each comment was converted to lowercase.
- Most common fill words were removed by stop_words function with German settings

Comment	Comment_alpha	Comment_no_stopwords
Korrekte Auflistung des Zählerstandes und Verb...	korrekte auflistung des zählerstandes und verb...	korrekte auflistung zählerstandes verbrauchs g...
Leichte Eingabe der Daten und schneller Wechse...	leichte eingabe der daten und schneller wechse...	leichte eingabe daten schneller wechsel abschl...
Hallo,Leider muss man sich mehr Fach mit dem S...	hallo leider muss man sich mehr fach mit dem s...	hallo leider mehr fach service wenden geschätz...
Ich bin rundum zufrieden mit e-on. Umzug mit Ü...	ich bin rundum zufrieden mit umzug mit übersch...	rundum zufrieden umzug überschneidung geklappt...
Alles korrekt. Allerdings erscheint bei uns n...	alles korrekt allerdings erscheint bei uns nur...	korrekt allerdings erscheint froschkönigweg pc...

Figure 28: Text transformation during sentiment analysis

After creating a test and training set, the filtered customer comments were converted to numerical columns with CountVectorizer from sklearn.feature_extraction. This new feature set was then applied to train SVM classifier and Random Forest classifier, again in two steps with and without tuning.

While we saw overfitting for both models, the results were overall much improved and far away from random guesses. Both classes were predicted over 75 % correctly, for the dominant class even up to 90%, with accuracy reaching up to 88% overall.

Model	Dataset	Performance metric before tuning		Performance metric after tuning[27]		
		accuracy	F1-score	accuracy	F1-score	
SVM Classifier	E.ON all data; *Comment_no_stopwords*	0.88 Some overfitting	Class 1: 0.78 Class 5: 0.92	0.88 Some overfitting	Class 1: 0.81 Class 5: 0.92	★
Random Forest Classifier	E.ON all data; *Comment_no_stopwords*	0.88 Abundant overfitting	Class 1: 0.77 Class 5: 0.91	0.88 Abundant overfitting	Class 1: 0.79 Class 5: 0.92	

Table 7:Performance metrics for Prediction of star ratings (sentiment analysis)

Interpretation of results

The reduction of the data set to one supplier and setting the target feature up as a binary problem helped significantly in improving performance of the simple machine learning models.

As expected, sentiment analysis on the *Comment* feature was the correct approach to predict star ratings. In case of not having entries in this column, the mandatory field *Headline* would be a good substitute, as we found here in most cases the content being a summary of the *Comment* content.

[27] Tuning for imbalanced data set and multi-class decision for SVM. Tuning with GridSearchCV with 3 fold crossvalidation for RandomForestClassifier.

Key word analysis

Closely related to sentiment analysis is the key word analysis, which leverages the same techniques to extract meaningful content which can support customer satisfaction reportings and deliver input for improvement initiatives.

To identify the most common topics in customer feedback and ratings, we were looking for key words in the headlines, comments, and supplier responses. In this case it made sense to focus on one supplier only, not only to keep the data volume manageable but also due to individual approaches to replying to customer feedback: We looked at E.ON Energie Deutschland GmbH first with nearly 5,000 lines, followed by Octopus Energy Germany with 6,000 lines of feedback.

The process started always with checking duplicates for each feature and creating a string containing the concatenation of all entries in the text column of the data set, inserting thereby a space between each line. Applying a function to streamline the process, the following steps were performed:

- Converting the text to lowercase characters only with the casefold method
- Cleaning a text of special characters, numbers etc. by regex and character mapping
- Tokenizing the content of the string by the TweetTokenizer from nltk.tokeninze library
- Filtering the word token list with the stop_words file specified for each supplier and German language setting.

Separating the comments by star rating allowed to compare 10586 positive and 2480 negative words of feedback, which showed similar key words for both features *Headline* and *Comment*.

Figure 29: Positive and negative feedback for E.ON (feature Headline)

For more advanced visualizations, the function was expanded to create a data frame listing every word with its number of occurrences, sorted by word count in descending order. With help of a color dictionary leveraging matplotlib color palettes, this allowed to compare directly key words and cluster information.[28]

Customers look for easy (einfach, einfacher, übersichtlich) and fast (schnell, schnelle) transactions in change of supplier (wechsel, vertrag, zählerstand, service, eingabe, abschlag).

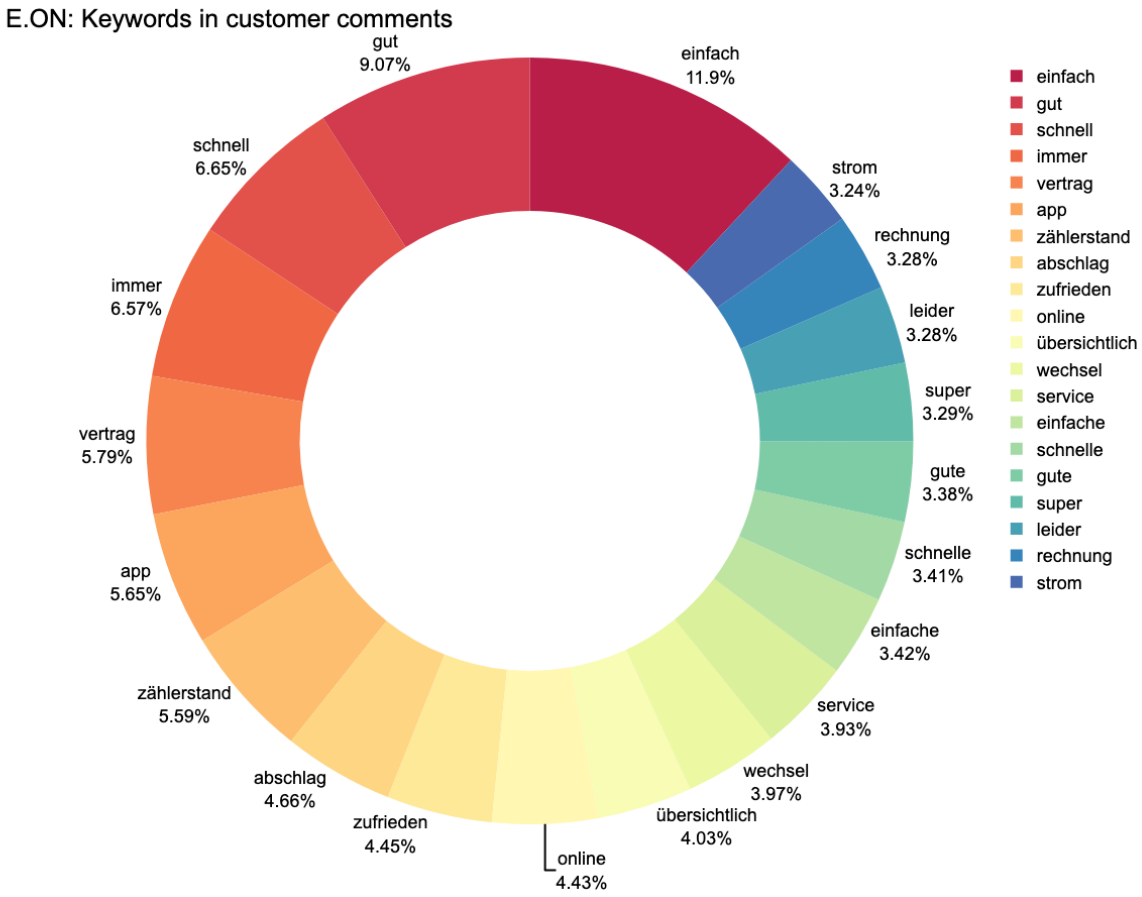


Figure 30: 20 most mentioned words in E.ON customer comments

Compared to classic word clouds the content is more quantified and systematic. This is especially helpful when comparing positive and negative comments between suppliers, as in the chart below, where 63178 key words in customer comments have been separated into 45867 positive and 4719 negative tokens.

Customers praise the easy (einfach, einfacher, übersichtlich, reibungslos) and fast (schnell, schnelle) transactions like change of supplier (anbieterwechsel, abwicklung, ablauf), also applicable for contact with Customer service (kundenservice, service, kommunikation, bearbeitung, app). Adjectives like good (gut, gute, guter, top), friendly (freundlich) and to be recommended (empfehlenswert) describe very positive feelings.

[28] For color dictionary and visualization ideas, see Boriharn, K.: Beyond the Cloud: 4 Visualizations with Python to use instead of Word Cloud, Towards Data Science Jul 2022: https://towardsdatascience.com/beyond-the-cloud-4-visualizations-to-use-instead-of-word-cloud-960dd516f215

Whereas on the complaint side, there have been issues with longer waiting times (wochen, monat, monate), concerning contract (vertrag, grundversorgung), change of supplier (gekündigt, wechsel) and invoice (abrechnung, endabrechnung, jahresabrechnung). The impression from the headline section is confirmed; in addition to the headlines, we see here also an issue with the email communication(mails, mail, email).

Figure 31: Word count of keywords in customer comments for Octopus Energy

Prediction of number of words of response

Classification of the modeling problem

The goal is to predict the length of company answers to customer posts, respectively the full answers, if possible. Hence, we restrict the data set to posts (rows) where a comment and an answer exists, i.e. *Comment_TF* = 1 & *Answer_TF* = 1. Recall that the answer policy is company specific (Rendering 1, p.13). Therefore, we choose the companies E.ON Energy and Octopus Energy Germany for further investigation, creating a data set for each using the *Company* variable. These are the companies with the most entries, 4965 and 8059, respectively. It turns out that these companies require rather contrary modelling approaches.

E.ON Energy

The E.ON data set turns out to be a binary classification problem. There are only two pairwise unique answers in the *Answer* variable, up to trivial modifications like spaces and captions. (In total there are six pairwise unique answers.) This makes it possible to predict the full *Answer* variable, i.e. there is no need to simplify the target by considering *log_Words_Answer.*

For E.ON energy, the prediction of answers is a **binary classification problem**, with **accuracy** as the **main performance metric**.

We want to maximize the True Positives (TPs) and True Negatives (TNs), so we decided on accuracy as main performance metric. For a deeper look into classification performance metrics see p. 24.

Octopus Energy

The answers of Octopus Energy Germany are highly personalized. Company answers reference user names directly. In the 5- and 4-star regime there seem to be standard answers (up to user names), but below, answers are personalized to a very high degree. To simplify the modelling, we choose the target variable *log_Words_Answer.*

For Octopus Energy, the prediction of *log_Words_Answer* is a **regression problem**, with **root mean squared error (RMSE)** as the **main performance metric**.

The RMSE is sensitive to outliers. However, this is not a problem as outliers are tamed by the natural logarithm. The square root accounts for the errors being in the same order of magnitude as the data.

Model choice and optimization

E.ON Energy

First, we replace the two possible answers in the *Answer* variable by 0 and 1, converting the Dtype of *Answer* to int64. The now binary target *Answer* is closely tied to the star rating. The variable *Stars_geq4_TF* is Pearson-correlated to *Answer* by 0.9957. In fact, there are only 9 cases out of 4965 where *Stars_geq4_TF* and *Answer* do not match. These cases occur when people confuse the star rating (5 is the best) with the German grading system (1 is the best), leading to comments which are contrary to the ratings. E.ON Energy gave the correct answers to the sentiment of the comment, not to the star rating. This means they either use a strong sentiment analysis model on the comments for automatized answers, or a human assigns the two standard answers manually.

In a first step we try to capture the strong relationship of the target variable to the star rating using logistic regression on the numeric columns *log_Words_Comment*, *log_Words_Headline*, *Stars_geq4_TF.* The test set is 20% of the total population. The model with default hyper parameters suppresses the first two numeric columns. It reduces to be a copy of *Stars_geq4_TF*. It is exactly the 9 cases discussed that it cannot predict correctly, neither on the training, nor on the test set. Still, we reach an astonishing accuracy on the total data set of 1 - 9/4965= 99.82%.

We tried to improve the model by projecting the three explanatory variables on a 2- and 1-dimensional subspace using principal component analysis, which gave no better results.

Finally, we perform sentiment analysis on the *Comment* variable. Using regex, we replace all non-letters by spaces, consecutively removing words of length 2 or less. Each comment is converted to lowercase. We filter for german stop words and replace the special german characters ä, ö, ü, ß, by ae, oe, ue, ss. We convert the column to numerical columns with CountVectorizer from sklearn.feature_extraction.text. These numerical columns derived from *Comment* are used to train a GradientBoostingClassifier with respect to a test set size of 20% and hyperparameters n_estimators=100, learning_rate=1, max_depth=1. On the test set we obtain an accuracy of (188+643)/(51+111+188+643)= 85%.

Octopus Energy

The first approach is to predict answer lengths *log_Words_Answer* on the numerical variables *log_Words_Comment*, *log_Words_Headline*, *Stars_min_max_scaled.* A standard scaler is applied on the two logarithmic varables. We will check the performance of several models (default hyperparameters) on a test set of test size 20%. This includes a custom model defined as follows: On the training set, compute the averages of *log_Words_Answer* grouped by *Stars_min_max_scaled.* On the test set, the predictions are defined as the computed averages (learned from the test set) rise to *Stars_min_max_scaled.* The performance metric is the square root of the mean squared error, i.e. the cartesian distance of the prediction- to the test-vector. The results are collected in the following table.

Model	Performance metric: RMSE
XGBRegressor	0.5006
RandomForestRegressor	0.5036
DecisionTreeRegressor	0.5516
LinearRegression	0.4732
Custom Model	0.4972

Table 8: Performance metrics for Prediction of number of words of response

We try to improve the situation with sentiment analysis. Following the same procedure on the *Comment* column as in the last section, but with a GradientBoostingRegressor, we get a performance of 0.5211.

Interpretation of results

E.ON Energy

The star rating variable *Stars_geq4_TF* is already a really strong indicator for the company answer with an accuracy of 99.82% on the total data set. A logistic regression model is able to mimic the variable *Stars_geq4_TF,* reducing to the identity after training. Using sentiment analysis with CountVectorizer and GradientBoostingClassifier, the accuracy 85% is worse, but still good. Language comprehension models that are more sophisticated should improve the accuracy. It seems that E.ON Energy either uses an advanced language model or a human decides which of the two standard answers are replied.

Octopus Energy

The model that performed the best to predict *log_Words_Answer* from numeric columns is LinearRegression with an RMSE of 0.47. Sentiment analysis with CountVectorizer and GradientBoostingRegressor leads to an RMSE of 0.52. This is slightly worse, though we assume that sentiment analysis should beat LinearRegression if a more sophisticated model than CountVectorizer is used.

Outlook

Practical Application

Monitoring customer feedback and regularly assessing customer satisfaction are key objectives in customer service management. With the analysis and predictive modeling in this scraped data set of Trustpilot ratings and comments for Energy Suppliers in Germany, we have simulated two typical use cases:

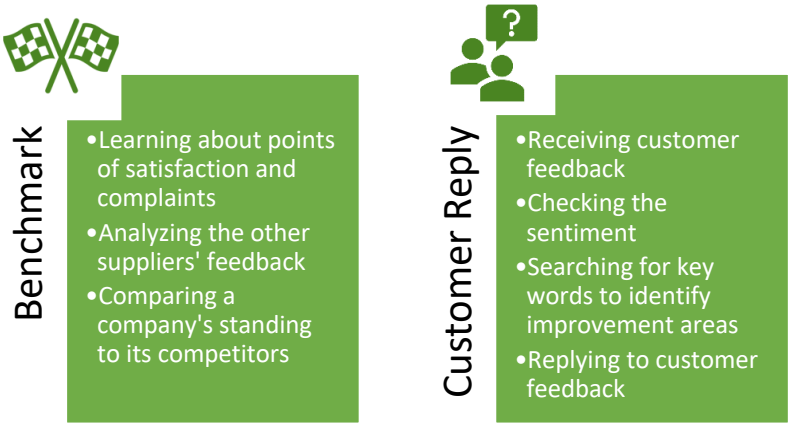


Figure 32: Use Cases from Customer Service Management

For both uses cases named above, double-checking the rating with the sentiment of the written comments is important, as some customers might have misunderstood the numbers for grades instead as rating points, and the written feedback should match rather the comments than the rating itself.

Looking for expected or important key words is the basis for the next shared step: Identifying points of interest to compare keywords or take them as input for a response to the customer.

We have learned that it makes sense to limit the data to one supplier, to accommodate for a supplier's specific product portfolio but also to compare within the same reference system. Consequently, the ML models needed to be trained for one supplier very efficiently and achieve high and reliable performance in rating prediction, e.g. as part of a sentiment analysis.

The next steps for Customer Reply could be to collect the previous answers as a library to design new and individualized answers, which could be prompted to a large language model, together with expected length, customer name, focus words and company values such as #gerneperdu, i.e. personal address, or eco-friendly marketing messages.

For Benchmarking, we have shown already that some keywords are shared for both positive and negative feedback. Nevertheless, in case of specific complaints accumulating, warning systems could be triggered to rectify what was crooked as soon as possible. Distribution of information and quantified basis for decision making can be achieved more easily with pre-set dashboards and visualizations in reports.

To summarize: In times of consumer markets and price comparison portals like Check24, customer expectations have evolved to personalized answers and timely feedback also on external platforms as the new normal. Therefore, the overall goal to turn happy customers into loyal customers can only be achieved by constantly checking competitors and applying AI supported customer service processes.

If we had more time

As an area of **optimization**, dedicating additional time would allow for the implementation of advanced machine learning models like Keras convolutional neural networks, thus improving classification accuracy.

Furthermore, employing key word analysis to categorize answers into relevant groups would streamline data processing, leading to more efficient resource utilization and enhancing decision-making processes.

For a **follow-up project**, a seamless integration of a pre-trained Language Model with an API could be pursued in order to automate the generation of company responses to customer inquiries.

- Utilizing keywords identified through our keyword analysis as prompts, the model could generate responses, drawing from a library of authentic company answers.
- Additional prompts such as answer length, customer name, focus words, and core company values could further refine the response generation process.
- To ensure accuracy and alignment with company procedures, a comparison sentiment analysis could be employed to evaluate both the generated and authentic responses.

Our graduation project has been a captivating journey, showcasing the potential of machine learning in tailoring seemingly personalized customer interactions, supporting typical use cases from customer service management like customer response and benchmarking.

By leveraging the large data resources of customer comments, ratings and supplier feedback, not only big enterprises but also smaller companies can gain valuable insights into customer sentiment, serving as a foundation for continuous improvement, competitive market analysis and enhanced customer satisfaction.

This effort highlights the transformative power of machine learning to foster better customer connections and driving organizational growth, regardless of company size or industry stature.

Appendix

Bibliography

This is an annotated list of literature and links supporting the research.

General information energy industry

- Stromversorgung und Netzbetrieb in Deutschland: https://www.verbraucherzentrale.de/wissen/energie/preise-tarife-anbieterwechsel/wer-macht-was-stromanbieter-netzbetreiber-messstellenbetreiber-38444#:~:text=Das%20Wichtigste%20in%20K%C3%BCrze%3A,k%C3%B6nnen%20Sie%20diesen%20also%20nicht.
- Regulation for the energy sector: https://www.ewerk.com/digitalhappen/energie/regulierung-gesetze-energiebranche
- European energy market: https://www.europarl.europa.eu/factsheets/de/sheet/45/energiebinnenmarkt
- Wikipedia for “the big four” energy companies in Germany: https://de.wikipedia.org/wiki/Die_gro%C3%9Fen_Vier_(Energieversorgung)
- Photovoltaics strategy of the German ministry of commerce and climate protection, May 2023: https://www.bmwk.de/Redaktion/DE/Publikationen/Energie/photovoltaik-stategie-2023.html
- Press notification on Smart Meter Law of the German ministry of commerce and climate protection, May 2023: https://www.bmwk.de/Redaktion/DE/Pressemitteilungen/2023/05/20230512-smart-meter-gesetz-final-beschlossen.html
- Groenveld, Josh: Energiekrise sorgt für Pleitewelle bei Strom- und Gasanbietern, Business Insider Jan 2022: https://www.businessinsider.de/politik/deutschland/fast-doppelt-so-viele-unternehmen-insolvent-wie-in-vergangenheit-energiekrise-sorgt-fuer-pleitewelle-von-strom-und-gasanbietern
- Monitoring report BNetzA, Nov 2023: https://data.bundesnetzagentur.de/Bundesnetzagentur/SharedDocs/Mediathek/Monitoringberichte/MonitoringberichtEnergie2023.pdf
- Annual report energy supply 2023 by BDEW (Federal Association of the Energy and Water Industry), Jan 2024: https://www.bdew.de/media/documents/Jahresbericht_2023_final_18Dez2023_V2.pdf
- Slide set for annual report energy supply 2023 by BDEW, Jan 2024: https://www.bdew.de/media/documents/Jahresbericht_2023_Foliensatz_final_18Dez2023_V2.pdf

Webscraping Target Site

- Trustpilot Website Search categories: https://support.trustpilot.com/hc/de/articles/360022026634
- Trustpilot transparency report: https://assets.ctfassets.net/b7g9mrbfayuu/tHyJSsKiNJxZvAuGPr6hz/5c6a42f3719debd02ca25989a722

5222/Trustpilot_Transparency_Reporting__Active_Recipients__under_the_Digital_Services_Act_-_21_March_2023.pdf

- Trustpilot energy suppliers: https://de.trustpilot.com/categories/electric_utility_company
- Official overview energy market actors (Bundesnetzagentur): https://www.marktstammdatenregister.de/MaStR/Akteur/Marktakteur/IndexOeffentlich

Visualization

- GeoPandas User Guide: https://geopandas.org/en/stable/docs/user_guide.html
- Hex code colors and palettes: https://www.color-hex.com/
- Seaborn color palettes: https://seaborn.pydata.org/tutorial/color_palettes.html
- Controlling figure aesthetics: https://seaborn.pydata.org/tutorial/aesthetics.html
- Pyplot Tutorial in Matplotlib: https://matplotlib.org/stable/tutorials/introductory/pyplot.html
- Plotly User Guide for Python: https://plotly.com/python/
- Matplotlib Examples: https://matplotlib.org/stable/gallery/index.html#subplots-axes-and-figures

Data Cleaning

- Datacamp.com: Data Cleaning Tutorial, Feb 2022: https://www.datacamp.com/tutorial/tutorial-data-cleaning-tutorial
- Ngo, Huong: How to Clean Your Data in Python, Jul 30, 2022: https://towardsdatascience.com/how-to-clean-your-data-in-python-8f178638b98d

Outliers

- Grace-Martin, Karen: Outliers: to drop or not to drop, 2018: https://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/
- Sharma, Natasha: Ways to Detect and Remove the Outliers, May 22, 2022: https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba
- Dey, Akash: How to handle outliers, Feb 2022: https://www.kaggle.com/code/aimack/how-to-handle-outliers/notebook
- IQR score: https://en.wikipedia.org/wiki/Interquartile_range
- Z-Score: https://en.wikipedia.org/wiki/Standard_score

Missing Values:

- Narang, Mohita: Handling Missing Values. Beginners Tutorial, Apr 4, 2022: https://www.naukri.com/learning/articles/handling-missing-values-beginners-tutorial/
- Kumar, Satyam: 7 Ways to Handle Missing Values in Machine Learning, Jul 24,2020: https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e

Statistics

- Bevans, Rebecca: Choosing the Right Statistical Test. Types & Examples, Jan 28, 2020: https://www.scribbr.com/statistics/statistical-tests/

Performance metrics

- Javatpoint tutorial on precision and recall: https://www.javatpoint.com/precision-and-recall-in-machine-learning
- Zeya LT: Essential things you need to know about F1-score, Towards Data Science, Nov 2021: https://towardsdatascience.com/essential-things-you-need-to-know-about-f1-score-dbd973bf1a3
- Huilgol, Purva: Accuracy vs. F1-Score, Medium Aug 2019: https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2
- Czakob, Jakub: F1 Score vs ROC AUC vs Accuracy vs PR AUCH: Which Evaluation Metric Should I Choose?, Neptune.ai Blog Sep 2023: https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc#:~:text=F1%20score%20vs%20Accuracy,observations%20both%20positive%20and%20negative
- Brownlee, Jason: Regression Metrics for Machine Learning, Jan 20, 2021: https://machinelearningmastery.com/regression-metrics-for-machine-learning/

Model choice and optimization

- User guide support vector machines: https://scikit-learn.org/stable/modules/svm.html; cited as Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
- Deepthi, A.R.: Support Vector Machines & Imbalanced Data. How does SVM work in the case of an imbalanced dataset?, Medium 2019: https://towardsdatascience.com/support-vector-machines-imbalanced-data-feb3ecffbb0e
- Sklearn documentation on C support vector classification: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
- Sklearn documentation on RandomForestClassifier: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
- Filho, Mario: Does Random Forest Need Feature Scaling or Normalization?, Forcastegy Jun 2023: https://forecastegy.com/posts/does-random-forest-need-feature-scaling-or-normalization/#:~:text=Random%20Forest%20is%20a%20tree,can%20be%20skipped%20during%20preprocessing.

Parameter optimization

- Sklearn documentation on cross validation: https://scikit-learn.org/stable/modules/cross_validation.html#stratified-k-fold
- Imbalanced data sets: https://medium.com/sfu-cspmp/surviving-in-a-random-forest-with-imbalanced-datasets-b98b963d52eb
- Sklearn documentation on Stratified K-Fold: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html
- https://en.wikipedia.org/wiki/Stratified_sampling
- HalvingGridSearchCV: https://towardsdatascience.com/stop-using-grid-search-cross-validation-for-hyperparameter-tuning-b962160dd6ae

- Sklearn documentation on HalvingGridSearchCV: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.HalvingGridSearchCV.html#sklearn.model_selection.HalvingGridSearchCV
- Kumar, Satyam: 20x times faster Grid Search Cross-Validation. Speed-up your cross-validation workflow with Halving Grid Search, Towards Data Science May 2021: https://towardsdatascience.com/20x-times-faster-grid-search-cross-validation-19ef01409b7c

Keyword analysis

- Brownlee, Jason: A Gentle Introduction to the Bag-of-Words Model, Machine Learning Mastery Aug 2019: https://machinelearningmastery.com/gentle-introduction-bag-words-model/
- TZJY: Natural Language Processing: Bag-Of-Words, Medium Oct 2021: https://medium.com/@tzjy/natural-language-processing-bag-of-words-python-code-included-ede3cfe979d2
- Boriharn, K.: Beyond the Cloud: 4 Visualizations with Python to use instead of Word Cloud, Towards Data Science Jul 2022: https://towardsdatascience.com/beyond-the-cloud-4-visualizations-to-use-instead-of-word-cloud-960dd516f215
- Luvsandorj, Zolzaya: Simple word cloud in Python, Towards Data Science, Jun 2020: https://towardsdatascience.com/simple-wordcloud-in-python-2ae54a9f58e5
- Choosing colormaps in Matplotlib: https://matplotlib.org/3.2.1/tutorials/colors/colormaps.html

List of figures

All figures have been prepared and created by the authors. Exceptions are listed below:

- Figures 2, 3 and 4 are translated versions from slides 29, 31 and 51 of the slide set for annual report energy supply 2023 by BDEW, Jan 2024: https://www.bdew.de/media/documents/Jahresbericht_2023_Foliensatz_final_18Dez2023_V2.pdf
- Figures 5 and 6 are translated versions of figure 73, page 167 and figure 75, page 169 in the 2023 monitoring report by BNetzA, Nov 2023: https://data.bundesnetzagentur.de/Bundesnetzagentur/SharedDocs/Mediathek/Monitoringberichte/MonitoringberichtEnergie2023.pdf

List of tables

All tables have been prepared and created by the authors.

Data Files and Contributions

All data files are available for download on GitHub:

https://github.com/DataScientest-Studio/jun23_cds_supply_chain

Additional Figures

TrustPilot Website

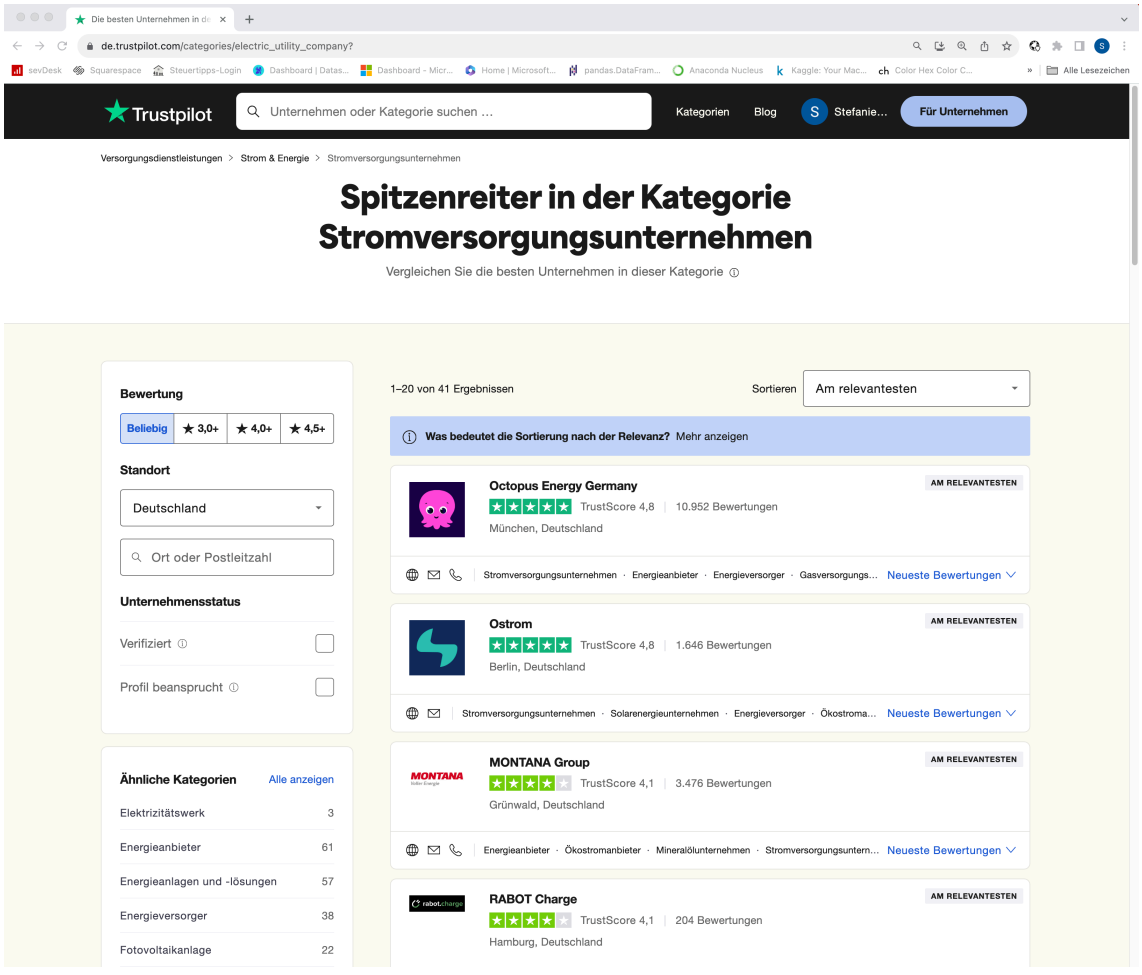


Figure 33: Energy suppliers on TrustPilot website

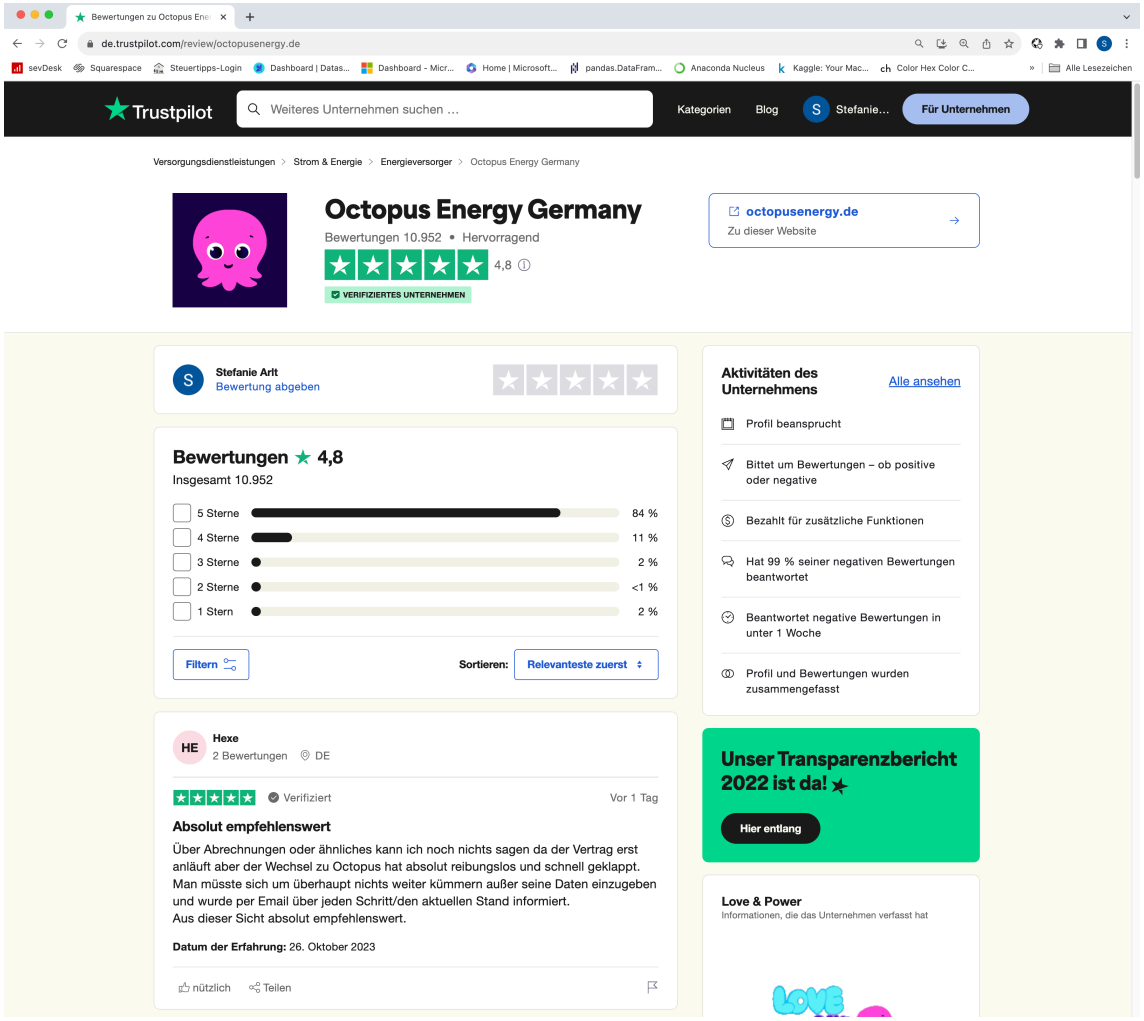


Figure 34: TrustPilot rating site "Octopus Energy"

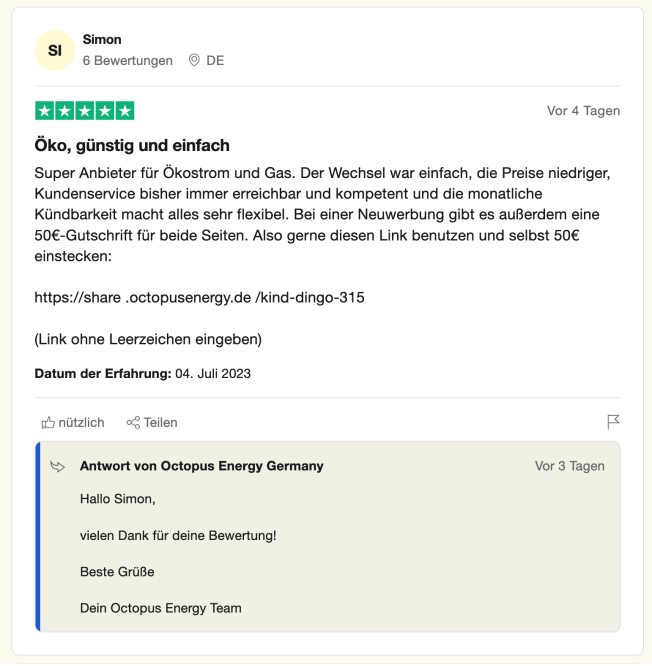


Figure 35: Customer vote with supplier feedback on TrustPilot