Importing the libraries

To start off the analysis, the first step was to load the necessary libraries. For this case, we, first of all, need Pandas to handle the raw data, which is contained in a CSV file. Second, for cleaning the data, we use the NLTK library and the TextBlob library, which is an API to the NLTK library. Additionally, we download NLTK's stopwords. Then for vectorizing the cleaned data, we load the TF-IDF and BoW classes of the Sklearn library. Finally, we need NumPy to prepare the data for visualization and Matplotlib for plotting.

Loading the dataset

After importing the necessary libraries, the next step was to load the customer complaints dataset. Additionally, we drop the NaN columns. A look at the shape of the dataset then reveals a format of 5629 rows and four columns. We also look at a histogram of the ratings to see the general sentiment in the dataset, which reveals that, indeed this dataset contains mostly complaints.

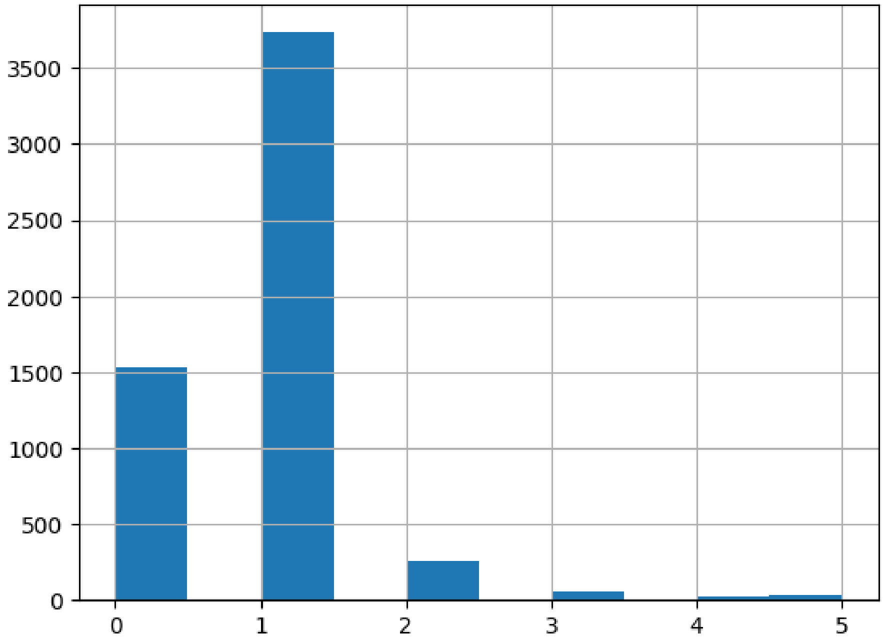


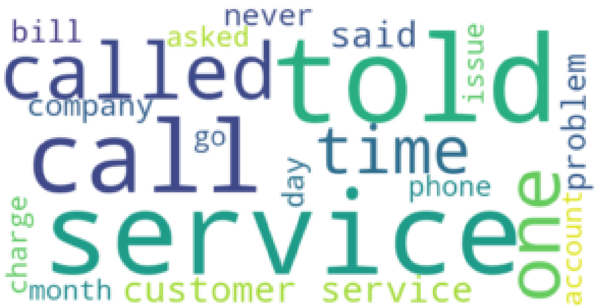
Figure 1: Histogram of the ratings column

Preprocessing the complaints

For preprocessing the complaints, there are four different options, which will be examined individually. The options are lemmatization, stemming, forming bigrams, and extracting noun phrases. For all options, the English stopwords, as defined by the NLTK library, are removed. Additionally, variations of the company name are added to the stop words as they occur fairly often and obscure the words more important to the analysis. To get an overview of the preprocessed complaints we also create word clouds for each option.

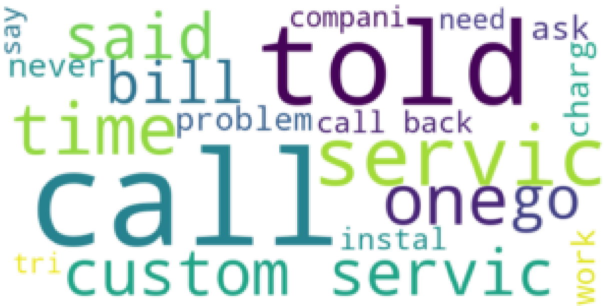
Preprocessing Option 1: Removing stopwords and lemmatization

By looking at the word cloud, we can already get a picture of what some of the major underlying issues of the company could be. For example, we could derive problems such as the customer service in general or that customers never got called back.



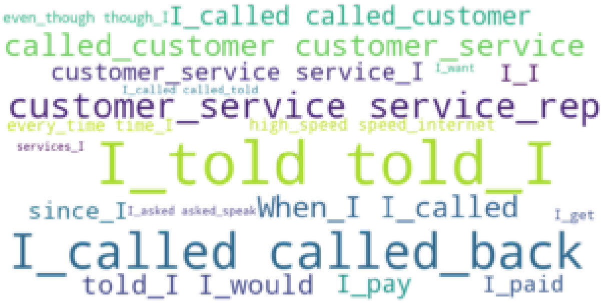
Preprocessing Option 2: Removing stopwords and stemming

For stemming, we get similar results as for lemmatization. Although lemmatization arguably is the better option as complete words are derived.



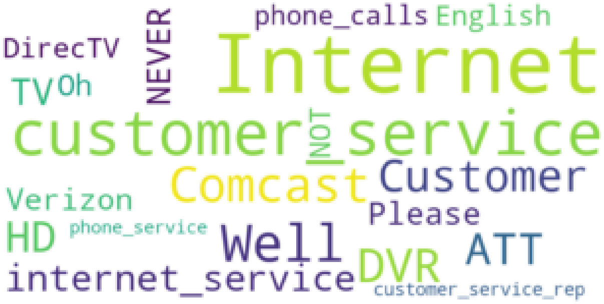
Preprocessing Option 3: Removing stopwords and forming bigrams

Like the previous two options, this word cloud also gives us a clue about possible issues with the customer service. Although in contrast to the previous word clouds, it is not so clear anymore if there is a problem with the customer service not calling back. This is most likely because of the formation of the bigrams.



Preprocessing Option 4: Extracting noun phrases

The fourth and last processing option reveals another possible set of problems in contrast to the customer service issues. According to this word cloud, there could also be problems with the internet service in general and also with the DVR, which is Comcasts' digital video recorder.



Vectorizing the processed complaints

For vectorizing, we will look into the BoW and TF-IDF methods. The process here is very similar to both options and consists of initializing the respective class, converting the complaints into a list, and passing the list to the vectorizers' fit_transform method. Additionally, we save the feature names which we need further on for topic modeling.

Topic Modeling

As for vectorizing, we will look into two methods for topic modeling, namely LSA and LDA. Here the process is also very similar for both options. We first initialize the respective class from the Sklearn library and pass the vectorized corpus to the fit method. The n_components parameter, which defines the number of topics, is set to four. After fitting the model, we then plot the top five words for each of the four topics. Following on the next page is the presentation of the 16 derived variations that resulted from combining the individual preprocessing, vectorization, and topic modeling options.

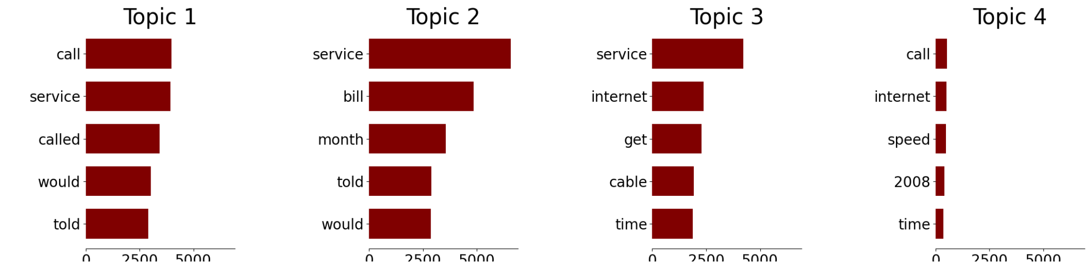


Figure 2: Lemmatization - BoW - LDA

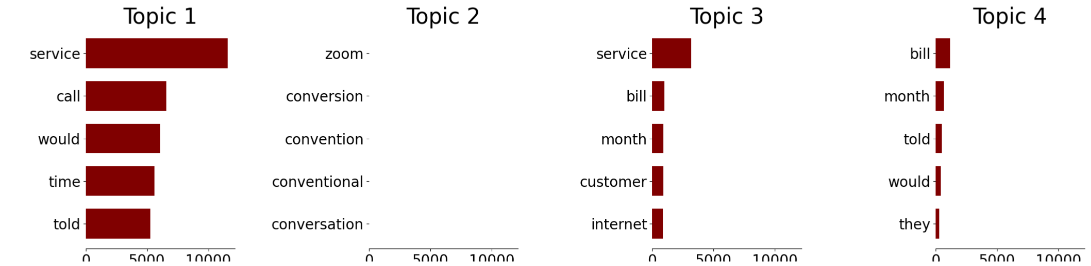


Figure 3: Lemmatization - BoW - LSA

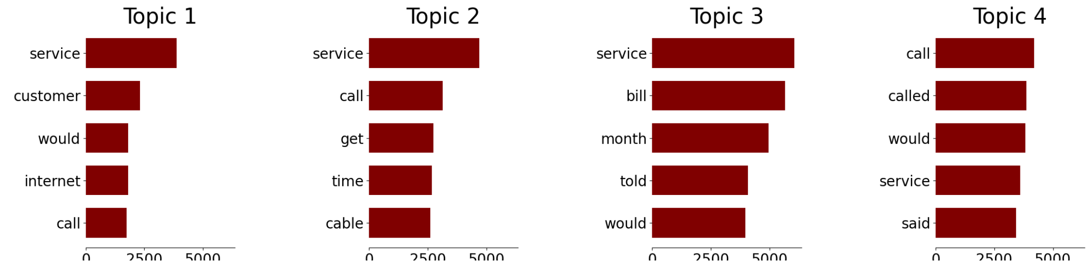


Figure 4: Lemmatization - TFIDF - LDA

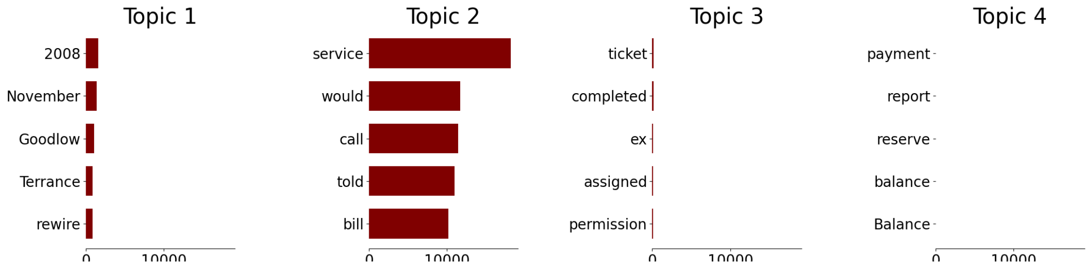


Figure 5: Lemmatization - TFIDF - LSA

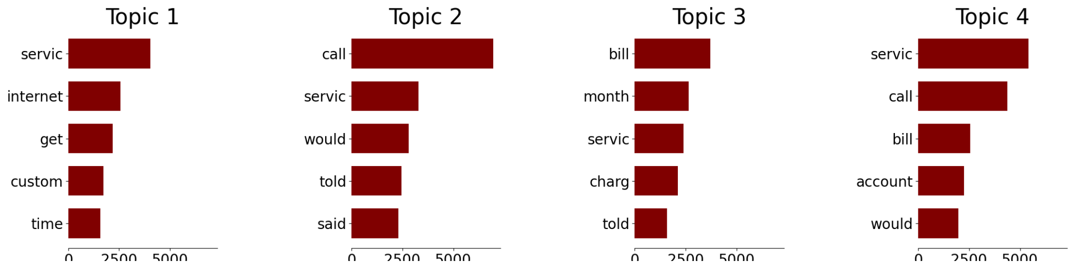


Figure 6: Stemming - BoW - LDA

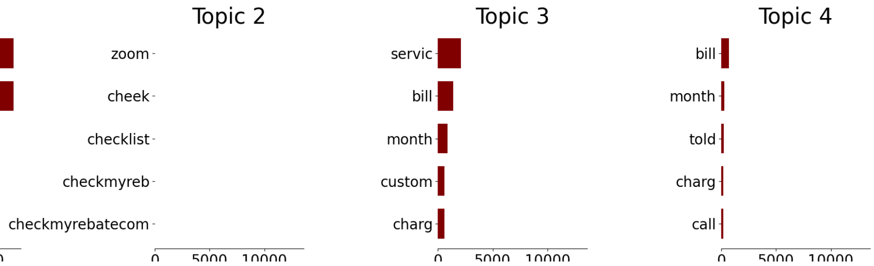


Figure 7: Stemming - BoW - LSA

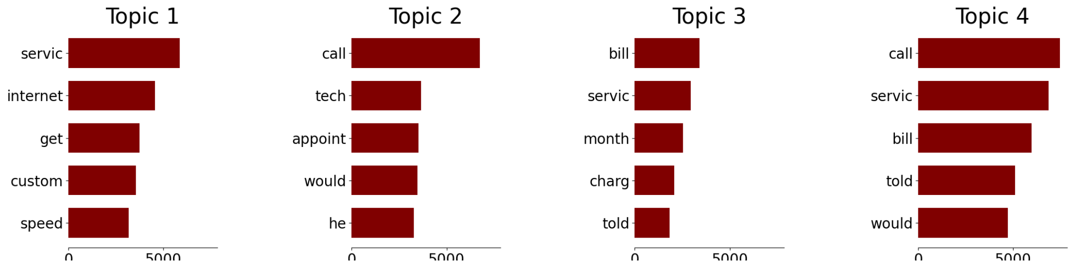


Figure 8: Stemming - TFIDF - LDA

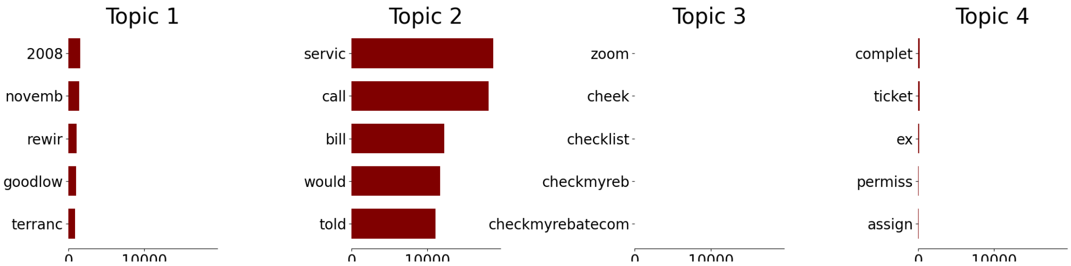


Figure 9: Stemming - TFIDF - LSA

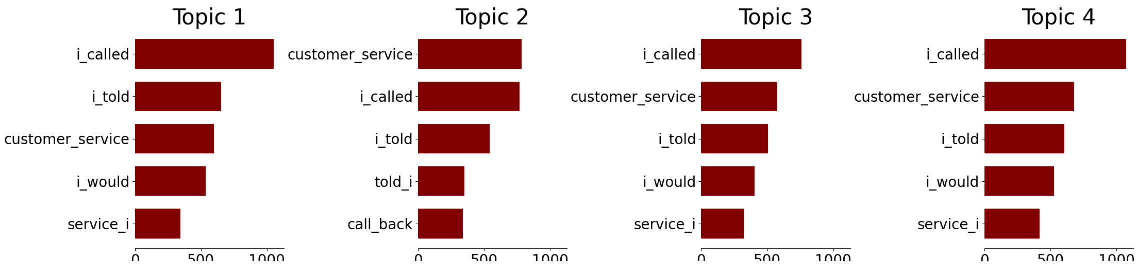


Figure 10: Bigrams - BoW - LDA

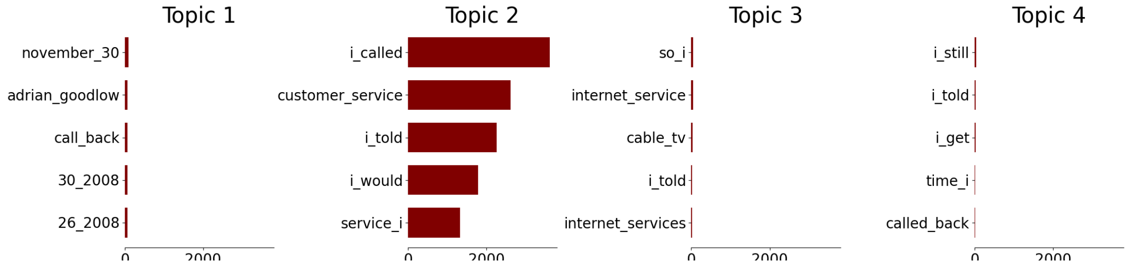


Figure 11: Bigrams - BoW - LSA

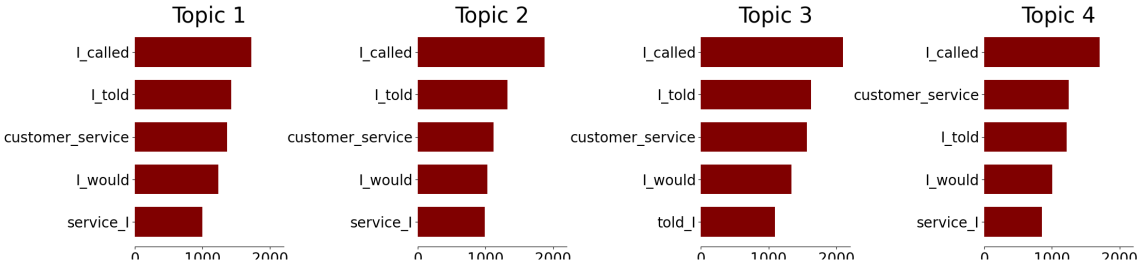


Figure 12: Bigrams - TFIDF - LDA

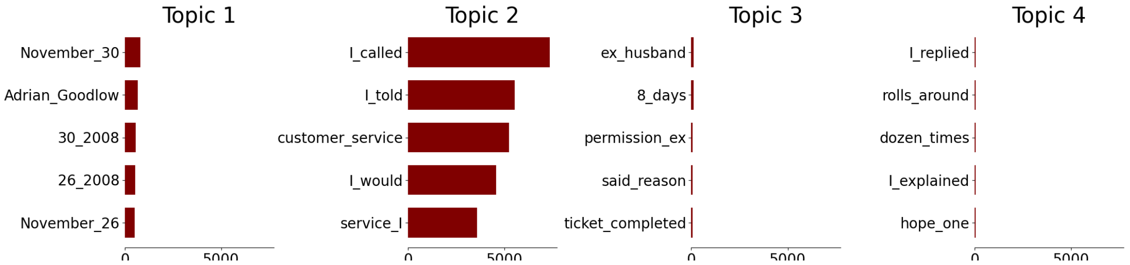


Figure 13: Bigrams - TFIDF - LSA

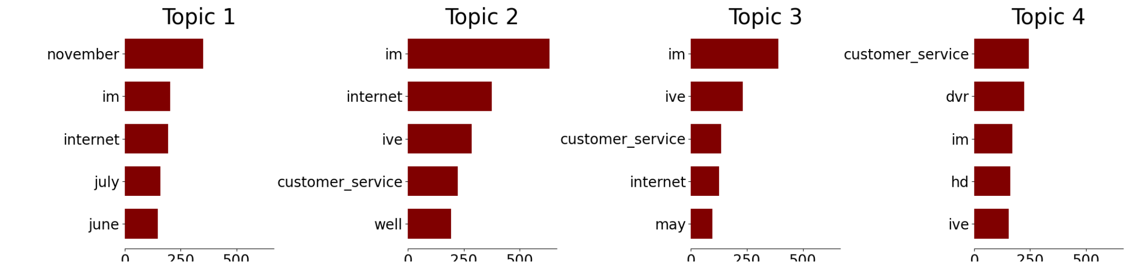


Figure 14: Noun Phrases - BoW - LDA

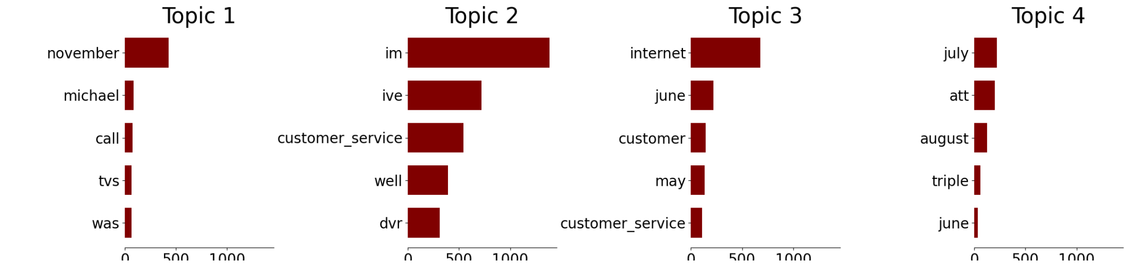


Figure 15: Noun Phrases - BoW - LSA

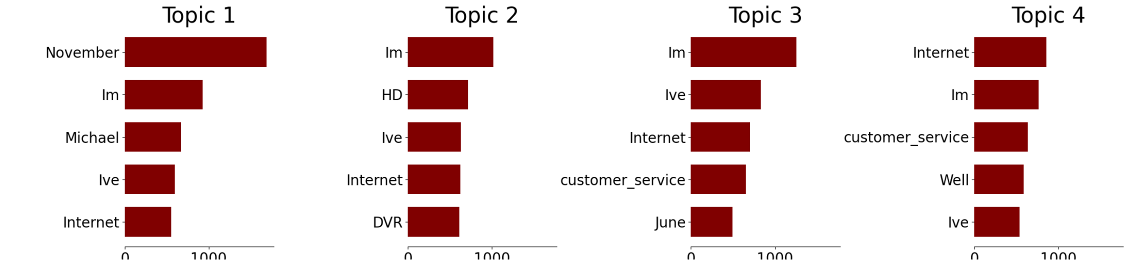


Figure 16: Noun Phrases - TFIDF - LDA

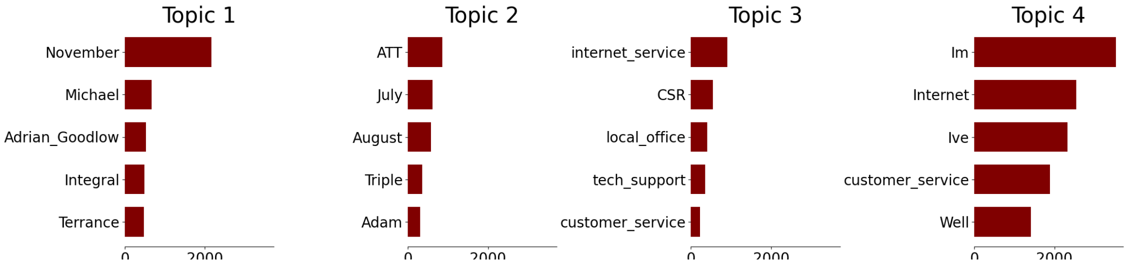


Figure 17: Noun Phrases - TFIDF - LSA

Interpretation of the topics

By looking at the derived topics in addition to the word clouds, we can arguably pinpoint the customer service as the biggest problem of the company. A quick internet search revealed that Comcast is indeed notorious for bad customer service. Some sources, for example, https://www.theverge.com/2014/8/19/6004131/comcast-the-worst-company-in-america claiming Comcast as the worst company in America. Additionally to the customer service issues, there also seem to be problems with their internet service and also their cable service.