NLP Analysis of HISCOX Online Customer Reviews

12/15/2020

Table of Contents

Executive Summary

Hiscox USA is the leading small business insurance carrier in the United States and serves thousands of customers with essential specialty insurance products such as Cyber, Professional Liability, General Liability, and Business Owner policies. Sales occur across all customer channels including direct agent and digital mediums (desktop and mobile). To improve the customer experience, the company utilizes Feefo to collect quantitative data and free text feedback at the conclusion of a customer's interaction with Hiscox. The marketing team is interested in sentiment analysis associated with customer reviews and what contributes to those ratings. The team performed quantitative Natural Linguistic Programming analyses. The findings were as follows:

Finding 1. Over 95% of customer reviews were positive in both score and sentiment analysis. Furthermore, these high ratings are consistent year-over-year. However, without knowing the total number of surveys being sent to customers, we cannot know if these results are a representative sample of all customer experiences.

Finding 2. Discrepancies do exist between some Star score ratings and the actual language from the corresponding review, but do not constitute a significant number of responses. Furthermore, there are very few (≈5%) where CoreNLP scores differ markedly from Star rankings.

Finding 3. Using Natural Language Processing (NLP) techniques, we found the top two to three ideas -- both positive and negative -- that resonate with a plurality of survey respondents. Year over year results show a consistent enthusiasm for Hiscox's customer service but suggest occasional issues with payment systems, agent timeliness and policy changes.

Keywords: Hiscox, insurance, NLP, customer, review, sentiment, analysis

Scope

Hiscox has been collecting customer reviews since 2016 using solutions from Feefo, a third-party vendor that provides "genuine feedback from genuine customers" via a survey request that is automatically delivered via email after a customer uses Hiscox's online portal. The purpose of this review platform is to provide customers unbiased opinions of a company's performance. (Corporate clients get feedback on overall sentiment, changes in website traffic, and customer insight.) Customers are given a single prompt and use free-from text to submit their concerns and opinions regarding their experience.

Requests from marketing were wide-ranging. Suggested lines of inquiry included:

- What is the customer experience from chat compared to that of live agents in call center?
- What is the overall customer interaction experience throughout the journey end-to-end?
- What is the experience with the brand? How is brand perceived?
- What are the common themes and pain points in reviews/chats?
- What are the topics in the journey which drives positive/negative sentiments?
- Can we infer customer profiles (types of business, insured's products, devices used in writing reviews, etc.)?
- What is the overall customer interaction experience throughout the journey end-to-end?

Due to time constraints, our team chose to pursue two business objectives (listed below) that seemed obtainable, given the data provided.

Business Objectives

The marketing team has two specific business objectives. These are highlighted in the form of questions.

1. *Do differences exist* between Service Star feedback scores and related customer comments? For instance, are there occasions where a customer leaves a top Star rating, but the overall tone of the review is mixed or even negative? The marketing department believes that inconsistencies

between star ratings and text sentiment prevents them from taking corrective actions associated with service level, product features, and customer experience.

2. Is there volatility in sentiment and scoring *over time*? Overall, are ratings are getting better or worse? Can any change be attributed to service, product, or other variables?

3. Can we discern high-level areas of customer concern and praise from the free-from text entries?

Data Quality

The Hiscox marketing department delivered two data sets: a free form text entry following an automatic survey (via Feefo), and a click to chat record of interactive chats between customer and agent. It consisted of 10,219 unique reviews, spanning late 2016 (when the contract with Feefo began) through the third quarter of 2020. (Personal identifiable information were redacted prior to loading.)

The dataset contains a number of fields of interest relevant to customer opinion:

- "Service Feedback" – consisting of negative ("-"), positive ("+"), double negative ("--") or double positive ("++") marks, which represent perceived service ratings.
- "Service Comments" – customer feedback in free-form text
- "Vendor Reply" – responses to the customer by service center agents in response to a Feefo review
- "NPS" - representing net promoter score
- "Service Star Feedback" – an overall service rating chosen by the customer, ranging from 1 to 5 stars. A 1 represents the lowest rating and a 5 represents the highest rating.

Real-world data is typically messy, and this was no exception. Table A shows the full set of data under consideration. The categories of `Vendor Reply`, `Tags` and `Nps_Score` do not contain a sufficient number of entries to be of significant value. And though there are only two null values

for Service Comment, this feature contains several entries where comments are comprised of only punctuation or numbers, which are subsequently dropped from consideration. Additionally, non-standard words, jargon and acronyms like CSR or E&O needed to be transposed to standard English for the Natural Language Process algorithms to function correctly. In free-form text entry, spelling errors are common and difficult to deal with in large quantity. This data set also contained 18 reviews that used Spanish which were also deleted. Finally, erroneous characters like ampersands and HTML tags were also removed from the reviews prior to Tokenization.

Exploratory Data Analysis

Following data cleaning procedures, we used common Exploratory Data Analysis (EDA) techniques to summarize the data set. The "token length", or number of words in a review, ranged from three to 428, with the vast majority of which were under 40 words in length (Figure 1). Collating all Service Star ratings shows that over 95% of all reviews since 2016 have been rated 4- and 5-Star by customers (Figure 2), which are considered to be net-positive impressions.

Data veracity, one of the "V"s of big data, is critical to successfully analyzing it for actionable insights (Bellazi, 2014). To validate the ratio of positive to negative reviews for the sake of due diligence, we looked at another third-party review platform vendor, TrustPilot, a publicly available database to which Hiscox also subscribes, and analyzed the ratio of positive to negative reviews (Figure 2). Confirming the ratio of positive reviews to negative with an outside source gives us confidence that the distribution of our sample is likely indicative of the customer experience as a whole. In our data set, the number of reviews by month, arranged by year, show a fairly consistent participation rate, with summer months tending to report slightly more responses (Figure 3). Exceptions would include 2016, when the program was not rolled out until December of that year, and 2020, where it is likely that COVID issues have resulted in depressing the quantity of responses. (We do not have data as to how many customers were sent survey forms, and so are unable to quantify and compare *response rates* over time.)

Methodology

Following initial EDA, we looked to address the client's requests using three different types of Natural Language Processing techniques: sentiment analysis, topic modeling and n-gram analysis. These are standard methods commonly used to begin an initial investigation of free-from text (Manning & Schultz, 1999)

Sentiment Analysis:

To address the *perceived discrepancies between review language and star rating*, we used an NLP method called Sentiment Analysis (SA) and the Stanford CoreNLP Processing Toolkit (Manning, et.al., 2017). For SA, we did not remove stop words or use lemmatization -- techniques useful for n-gram and topic modeling -- as these pre-processing steps often make words lose their context and reduce information content. Using the Natural Language Tool Kit (NLTK) with Stanford's CoreNLP pre-trained models we input the free-form customer reviews to get the sentiment scores for each individual review, then compared those scores to the Star Scores in the Feefo data, looking for discrepancies. The Stanford CoreNLP algorithm did an adequate job of classifying reviews that were primarily negative in sentiment, but some reviews were tagged improperly and given a negative rating by the classifier when the tone of the review is at least somewhat positive. A comparative sample of reviews (Table B) and the absolute difference between CoreNLP scores and Star Ratings are listed below (Table C). Overall, we observe that the Stanford CoreNLP tagged about 60% of the reviews as positive (scores of 4 or 5), about 30% of the reviews were tagged as neutral, and around 10% of the reviews are labeled negative. The plots of both show more than a passing similarity to each other (Figure 4).

The client specifically asked for an assessment of service rating trends over time. To that end we used a time-series analysis of weighted averages that shows the trendline of customer ratings since the project's inception (Figure 5). This trend averages around 4.75 for most years. 2016 is an outlier because we have very few observations from this year with which to analyze. When compared to the similarly weighted Stanford CoreNLP Scores for Sentiment Analysis, we see that the overall scores, though lower, mirror the trends in the Star ratings. This suggests the Stanford model is compatible with the Star ratings, if not yet finely tuned.

N-Gram Analysis:

The client is interested in commonalities among customer reviews, and an N-gram analysis is a common tool used to explore these relationships. An N-gram model is built by counting how often word sequences occur in corpus text using a window of size N (Manning & Schultz) . After counting the frequency of each term, we ranked them to see the most common terms.

N-Grams from positive reviews (3-5 star) between 2016 and 2020 seem to be consistent. Most positive reviews seem to mention customer service which is "good", "fast", and "helpful", while others mention "fair pricing", "easy quotes", and information they may provide for small businesses. N-grams with negative (1- and 2-Star) reviews seem to be inconsistent from year to year. In 2016, "charge" and "slow resolution" were the most frequently cited issues. In 2017, issues with "online payment information" topped the list of customer concerns. 2018 shows customers calling "coverage" and "card payment" the biggest problems, but in 2019 it was "slow customer service" that concerned them the most. 2020 complaints appear to be issues related to "policy changes".

The most common bigrams and trigrams (with counts) associated with negative reviews are listed below by year (Figs. 6 - 9).

Topic Extraction:

Using N-gram analysis we also extracted the most common topics of customer feedback by year for both positive reviews (Table D) and negative reviews (Table E).

Several of the most frequently cited topics are mentioned every year. They appear to concern customer service, pricing, the ease of getting a quote and insurance for small businesses, all items we would expect for reviews of an insurance company.

Conclusions

The vast majority of responses -- over 95% -- were *positive in both score and sentiment analysis*. This ratio of positive to negative comments can be validated by examining a similar, external source (TrustPilot). Furthermore, these high ratings are *consistent year-over-year*. Ratings are not getting worse, but they generally don't improve, either. It should be noted that we do not

have access to data which might show average customer interactions industry-wide, useful for benchmarking customer service comparisons with Hiscox's competitors. Neither do we know the total number of surveys being sent out to customers. Thus, higher customer ratings may be a result of a response rate lower than average (and primarily it is satisfied customers who are offering reviews).

Only 9% of Star Scores agree perfectly with CoreNLP Scores, but only a small percentage (5%) differ significantly (≥3) in the comparison of expected sentiment with the customer's rating. 57% differ by a factor of only one. This is to be expected as the accuracy of CoreNLP sentiment analysis is subject to the quality and quantity of the training data and may also reflect the industry-specific nature of the Hiscox texts.

N-gram analysis worked well for finding bigrams and trigrams that are naturally meaningful. Associating them with either positive or negative reviews allows us to see phrases that hint at potential areas of concern for decision makers. Furthermore, the same method was tweaked to find the top two or three most-common ideas each year that permeate customer reflections. These are the "uppermost thoughts" in customer's minds after interacting with Hiscox.

Concerns/Issues

Customer feedback and comments seem to be overwhelmingly positive. The client, however, is interested in both positive and negative feedback. Thus, while we have isolated the data per year into two camps, positive customer reviews with a Service Star rating of four and five stars and negative customer rev with a Service Star rating of three stars or less -- it is difficult to parse small text blobs for nuanced meaning at scale.

Although text used for n-gram analysis was preprocessed, results rendered from our initial efforts were not insightful. We discovered we needed to go back into our preprocessing script and add stop word removal and eliminate lemmatization to make results for our bigram analysis more interpretable. Further, some bigram terms were obviously being truncated, so we expanded our analysis to trigrams, which yielded a richer trove of information.

There is controversy in applying parametric methods to non-parametric data like the CoreNLP sentiment scores and Star ratings (Carifio & Perla, 2008) (Jamieson, 2004), especially when there is only one Likert-style question in the survey. Both scores are ordinal scale data, and don't have defined intervals between them. One can't say a star score of four is exactly twice as good as a score of two, for instance. One potential way of discerning accuracy in the Sentiment Analysis is to see if the bulk of scores "line up" as it were, with the Star Scores. Our thought process is this: it's reasonable to assume most five-star reviews will be reflected as such in the language used. Then it would be possible to search out the instances where the variance is greatest -- both positive and negative -- and hand-review an appropriate proportion of these for accuracy, noting any significant departures. For future reference, it would be appropriate for the marketing department to revisit best practices for survey design to allow for more meaningful comparisons.

For Further Study

Investigate creating an industry-specific lexicon that can augment the one used by the Stanford CoreNLP kit. It would be extremely useful to have a custom model that can accurately classify the small reviews that are somewhat ambiguous in assessment before attempting further Sentiment Analysis en masse. This would allow the use of machine learning models that can automatically send pertinent information to management.

Topic modeling analysis with respect to time may show a correlation of the top terms identified with specific events that might have triggered these reactions. Taking a deep dive into the comments that mention these keywords can provide Hiscox with a clearer picture of the situation as well as a base for taking concrete action items to resolve customer issues. One example of this is the collection of TrustPilot reviews of the Hiscox U.K. arm. That ratio of satisfied to unsatisfied customer reviews is much more dire: *nearly half of all reviews report a strong negative reaction* to Hiscox. Further investigation shows that several are tied to unmet expectations of claims for Business Interruption insurance, triggered by the nation-wide lockdown in the Spring of 2020 that closed nearly all businesses for an extended period of time. Prior to that event, there was a surge of very unhappy clients in 2019 during what appears to be a significant rise in homeowner premiums. These are, of course, broad events easy to delineate.

However, good Sentiment Analysis can also detect much more subtle trends provided the model used is a finely-tuned one.

Finally, of the most frequently mentioned terms was "customer service", and we wrote our code to categorize both positive and negative reviews that mentioned it. By going through these reviews by hand, Hiscox can gain insight into the specific situations.

Citations

Bellazzi, R . Big data and biomedical informatics: a challenging opportunity. Yearb Med Inform 2014; 9: 8–13.

Carifio J, Perla R. Resolving the 50-year debate around using and misusing Likert scales. Med Educ. 2008 Dec;42(12):1150-2. doi: 10.1111/j.1365-2923.2008.03172.x. PMID: 19120943.

Jamieson S. Likert scales: how to (ab)use them. Med Educ. 2004 Dec;38(12):1217-8. doi: 10.1111/j.1365-2929.2004.02012.x. PMID: 15566531.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.

Manning, Christopher D., Schultz, Hienrich. 1999. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA.

Tables

Table A.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10219 entries, 0 to 10218
Data columns (total 12 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   Customer Name                           10216 non-null  object
 1   Service Date                            10219 non-null  object
 2   Feedback date                           10219 non-null  object
 3   Tags                                    20 non-null     object
 4   Service Feedback                        10219 non-null  object
 5   Service Comment                         10217 non-null  object
 6   Vendor Reply                            456 non-null    object
 7   Nps Score (Promoting product to others) 3474 non-null   float64
 8   Service Star Feedback                   10219 non-null  int64
 9   month_year                              10219 non-null  period[M]
 10  month                                   10219 non-null  int64
 11  year                                    10219 non-null  int64
dtypes: float64(1), int64(3), object(7), period[M](1)
memory usage: 958.2+ KB
```

Table B.

	clean_reviews	ratings	StanCoreNLP_Score	diff_ratings_StanCoreNLP_Score
2161	I ? m a new business owner and was unsure of e...	5	3	2
398	I was disappointed with the insurance , becaus...	4	2	2
4645	It was fine and your folks are always responsi...	4	4	0
5637	So far good easy to obtain .	5	4	1
7329	Prompt , responsive , helpful and reasonable	5	4	1
7377	Great Customer Service	5	4	1
1016	Excellent customer service . Very informative ...	5	4	1
1025	Easy to use . No hassle way to acquire .	5	3	2
3606	It was an excellent experience when we started...	5	4	1
8424	I was very impressed with how well informed th...	5	3	2

Table C.

```
1    57.0
2    28.0
0     9.0
3     5.0
4     0.0
Name: diff_ratings_StanCoreNLP_Score, dtype: float64
```

Table D.

Compilation of Top Bigrams & Trigrams from Positive Reviews

2016	2017	2018	2019	2020
fast service	excellent customer service	excellent customer service	customer service excellent	excellent customer service
good service	would highly recommend	easy to (get) quote	would recommend hiscox	easy to get quote
fair price	small business	fast and easy	small business owner	fast and easy
friendly helpful				

Table E.

Compilation of Top Bigrams & Trigrams from Negative Reviews

2016	2017	2018	2019	2020
charge incorrectly	update payment information	switch payment card	get certificate insurance	policy change
charge monthly	payment information online	coverage approximately less	call back multiple (times)	rewrite entire policy
expect faster resolution				requirement sent week ago

Figures

Figure 1.

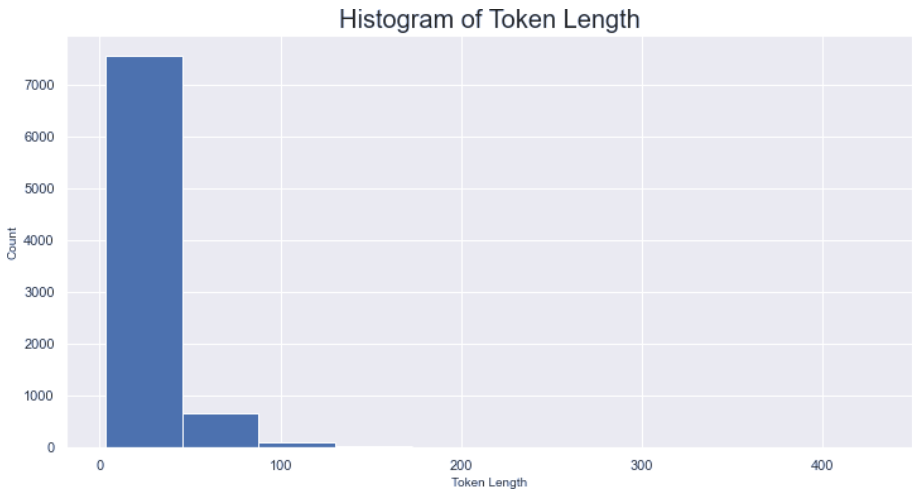


Figure 2.

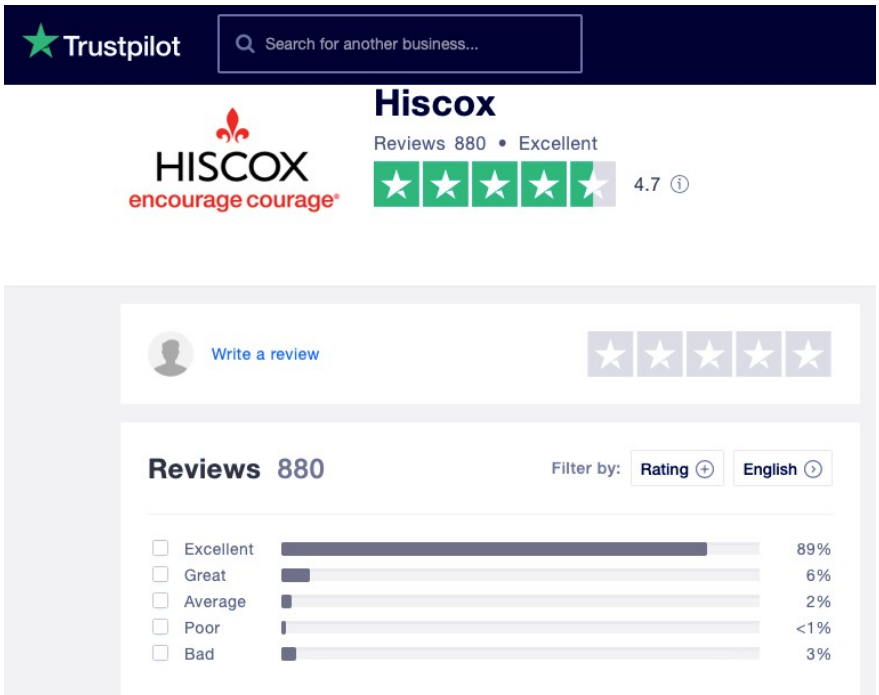


Figure 3.

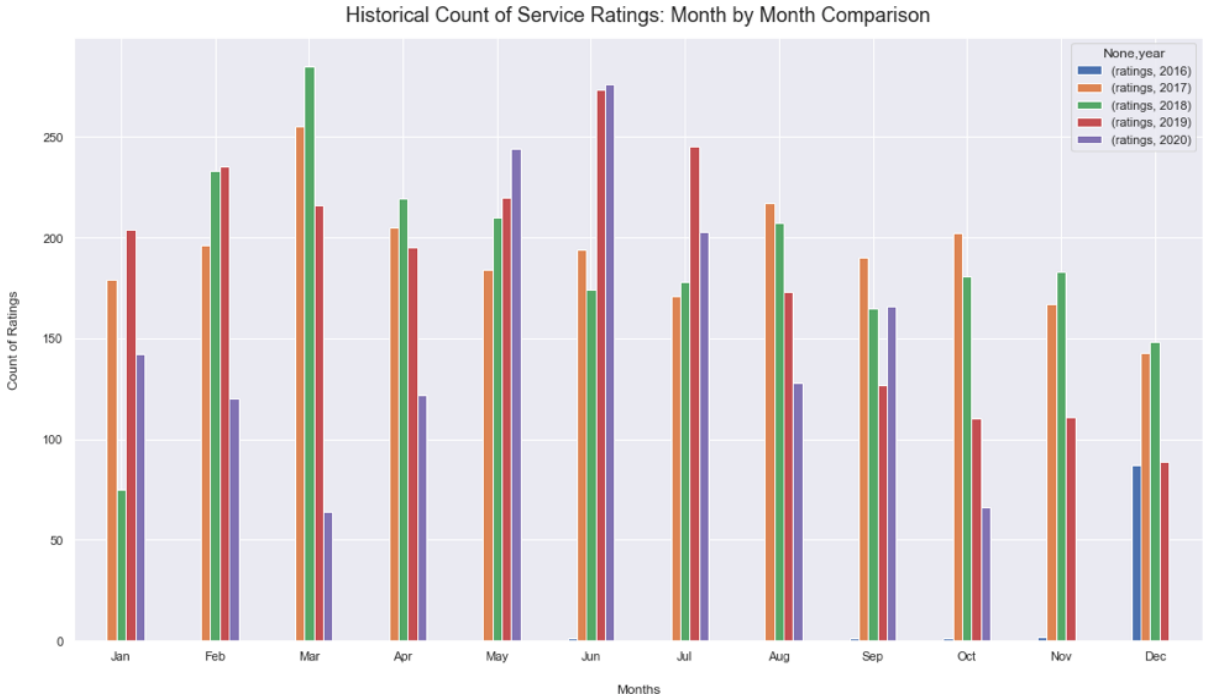
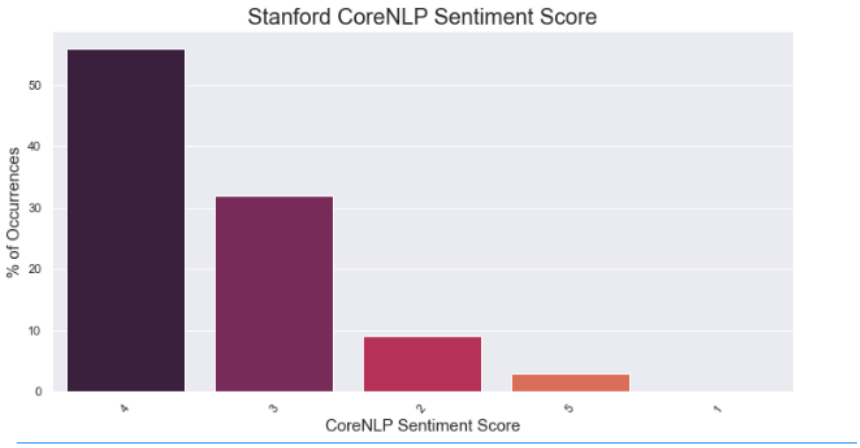


Figure 4.



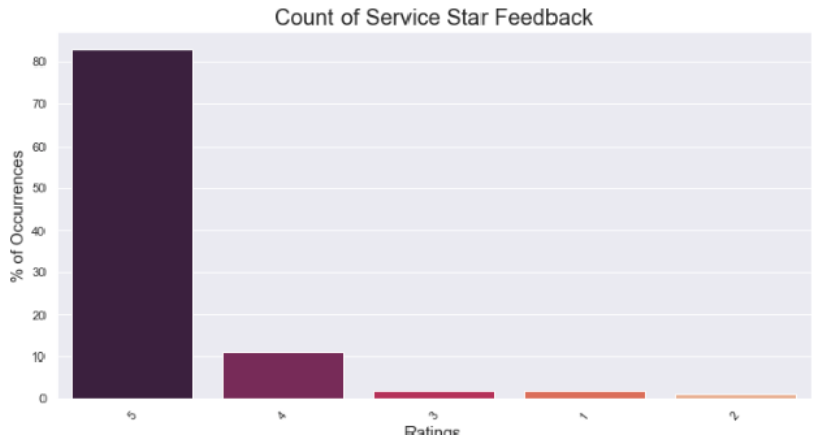


Figure 5.

Figure 6.

```
Top 10 Bigrams for 2017
-----------------------------------
call back ---> 6
customer service ---> 5
insurance company ---> 5
tell would ---> 5
still wait ---> 4
get pay ---> 3
receive policy ---> 3
never receive ---> 3
sign insurance ---> 3
credit card ---> 3
Top 10 Trigrams for 2017
-----------------------------------
update payment information ---> 2
payment information online ---> 2
terrible customer service ---> 1
customer service cancel ---> 1
service cancel policy ---> 1
cancel policy agent ---> 1
policy agent give ---> 1
agent give application ---> 1
give application email ---> 1
application email send ---> 1
```

Figure 7.

```
Top 10 Bigrams for 2018
-----------------------------------
customer service ---> 4
would get ---> 4
call back ---> 4
request change ---> 3
get policy ---> 3
login policy ---> 2
exist customer ---> 2
go process ---> 2
able find ---> 2
willing listen ---> 2
Top 10 Trigrams for 2018
-----------------------------------
switch payment card ---> 2
find coverage approx ---> 1
coverage approx less ---> 1
approx less line ---> 1
less line could ---> 1
line could disappointed ---> 1
could disappointed hiscox ---> 1
disappointed hiscox incredibly ---> 1
hiscox incredibly poor ---> 1
incredibly poor experience ---> 1
```

Figure 8.

```
Top 10 Bigrams for 2019
-----------------------------------
call back ---> 13
customer service ---> 6
still receive ---> 5
say would ---> 4
several time ---> 4
multiple time ---> 4
certificate insurance ---> 3
cancel policy ---> 3
back get ---> 3
cancel hiscox ---> 3
Top 10 Trigrams for 2019
-----------------------------------
call back get ---> 3
get certificate insurance ---> 2
call several time ---> 2
call back multiple ---> 2
back multiple time ---> 2
still receive refund ---> 2
day never get ---> 2
poor customer service ---> 2
go throw three ---> 2
purchase business insurance ---> 2
```

Figure 9.

```
Top 10 Bigrams for 2020
-----------------------------------
customer service ---> 5
still wait ---> 3
storage unit ---> 3
week ago ---> 2
another company ---> 2
business address ---> 2
insurance policy ---> 2
question ask ---> 2
change policy ---> 2
rewrite entire ---> 2
Top 10 Trigrams for 2020
-----------------------------------
rewrite entire policy ---> 2
storage unit tell ---> 2
customer service rep ---> 2
still wait document ---> 1
wait document send ---> 1
document send requirement ---> 1
send requirement week ---> 1
requirement week ago ---> 1
week ago still ---> 1
ago still nothing ---> 1
```