Summarize product reviews from Youtube videos, comments, search-suggestions and blogs

Phase 1: Literature Survey and Data Sources:

Abstract:

Summarizing the product from the reviews has recently attracted a lot of attraction due to its extensive application. ECommerce companies like Amazon, flipkart contain massive amounts of reviews data. Consumers and sellers spend a large amount of time reading through long reviews to find out what is perceived as good and bad about a product. It is important for customers and sellers to understand what exactly the negative review was about. For e.g.; If amazon delivered a product late by a week, the 'service' aspect is bad rather than the product.

The problem of summarization of reviews can be solved using Aspect Extraction of Opinion mining. Extracting/identifying the aspects of the product can help the business to attract more customers and improve the service. Therefore summarizing the product reviews into various aspects and understanding its sentiment can help the business and customer tremendously. Most of the existing aspect extraction task is done on the small-labeled dataset and applying these methods on large scale datasets may produce irrelevant results and not be scalable. Also, labeling such large data is a huge challenge and not practical. To overcome this problem, an unsupervised learning approach is more suitable. Furthermore, Existing methods of aspect extraction based on unsupervised approaches use either fixed number of aspects or only frequency based approaches which may extract a huge number of the aspects but most of them might not be relevant to the domain and on the other hand it might exclude infrequent relevant. This study aims to cover such limitations by exploring an efficient approach of combining frequency-based approach (word level) and a syntactic-relation based approach (sentence level) which is enhanced further with a semantic similarity-based approach to extract aspects that are relevant to the domain, even terms (related to the aspects) are not frequently mentioned in the reviews.

1. Business problem

1.1 Introduction:

Summarizing the product reviews can be used in various applications. Business/sellers can use this to understand what aspect of the product/service the customer talks about the most and what is sentiment associated with that aspect. Likewise customers/users can go through an individual aspect of a product rather than reading through all the long reviews. Hospitality industry can use this to identify sentiment for each aspect category

i.e., hotel staff, food variety, price, taste, location etc. Such type of aspect extraction is really critical sometimes.

For e.g.;
Take this review ‘Food was pretty average but ambience and service was awesome’.
Here the user is appreciating the service and ambience of the place.
Now take this review ‘Food was okay but staff were very rude’.
Here even though food is okay but service is horrible and this type of feedback is really critical for the hospitality industry.

1.2 Challenges:

The main challenge here is working with large-scale unlabeled data and extracting all the aspects without using any fixed aspect vocab. Furthermore, it should be able to extract all the aspects relevant to the product/domain and exclude irrelevant aspects.

2. Earlier methods

Most of the earlier approaches and researches regarding the motion magnification can be categorized into the following two approaches.

2.1 Vocab Based Approach:

In this approach, a predefined aspect vocab is used. Few researchers use only a fixed aspect list while others use a fixed list of aspects to find the relevant terms of that aspect. Further clustering is performed by finding the similarity between review sentences and aspect terms. While this approach is reasonably good for small scale data but It is observed that learned aspects tend to work better on large-scale data.

2.2 Frequency Based Approach:

This is the most used method by researchers to extract the aspects. Despites being very simple, it is quite effective. This method basically first high occurrence words and choose the noun and noun phrase as a candidate for aspect. If the frequency of the candidate aspects exceeds some threshold, it is considered as an aspect. Once the aspect is extracted then the aspect's sentiment word is selected based on the nearest adjective and finally polarity value is assigned to that aspect.Although the frequency-based method is an efficient one, it has obvious limitations. One of the limitations is that the approach may select words that are not aspects (i.e., pick up many words that do not contain any subjectivity) because it relies only on word frequencies.Furthermore, aspects that are not frequently mentioned will not be detected using this method.

2.3 Syntactic Relation Based Approach:

The syntactic relation-based method also called the rule based method aims to analyze the syntactic structure of the sentence and the relations among the words to identify the aspect's sentiment words.

2.4 Topic Modeling Based Approach:

2.5 Approach expected to be followed in this project:

We are going to take reference of experiments given in this paper.
Our approach is the hybrid approach which combines frequency-based approach (word level) and a syntactic-relation based approach (sentence level) which is further enhanced using semantic similarity based approach. The core idea of this approach is to extract all the aspects related to the domain without using any fixed list of aspects and also extract that aspect which is not frequent. And produce sensible results on large-scale data.

This approach consists 3 task -

1. Aspect Extraction
2. Summarize reviews based on aspects.
3. Estimate aspect sentiment rating

3. Dataset

3.1 Challenges in obtaining dataset:

It is very difficult to obtain or scrape the large-scale data under legal compliance. Apart from that we need to have some labeled data in order to evaluate the performance of the model. The data set should also have a wide range of product/service as well as varying distribution of reviews for each product.

3.2 Use API to Collect data:

We can collect places API to collect listing/places reviews data but the only problem with this API is that it lets you collect only 5 star reviews data points and the quota is really small. There are other similar APIs available as well but it seems to be quite costly.

3.3 Scrape the dataset:

This is one of the easiest ways to collect unlabeled data. It might require extensive preprocessing but it is a good approach for the experimentation purpose. There are various public scraper tools available out there with various limitations but it will get your work done with a bit of tweaking.
We are using one such scraper tool among many to scrape amazon product reviews data. This API can scrape up to 500 products and 1000 reviews each product.

3.4 Publicly Available Dataset
Often large companies make their data public under legal compliance to help the innovators and contribute to the open source society. This 'open data' helps both companies as well as academic researchers mutually. We have one such unlabeled data available from amazon as well.
Amazon Customer Reviews Dataset
About the Data
The dataset contains the customer review text with accompanying metadata, consisting of two major components:

1. A collection of reviews written in the Amazon.com marketplace and associated metadata from 1995 until 2015. This is intended to facilitate study into the properties (and the evolution) of customer reviews potentially including how people evaluate and express their experiences with respect to products at scale. (130M+ customer reviews)
2. A collection of reviews about products in multiple languages from different Amazon marketplaces, intended to facilitate analysis of customers' perception of the same products and wider consumer preferences across languages and countries. (200K+ customer reviews in 5 countries)

3.4 Sources of dataset :

Sources of dataset:

For this project we have used places API to get the domain specific data and get the sense of the real world. We have collected 5 locations data which belongs to the automotive industry.

4. Loss functions and metrics

Though our primary goal is to extract aspects of unlabeled large scale data. Due to unavailability of labels on that data, we won't be able to compute any performance

metric but to evaluate the performance of the model we have collected some small-set labeled data.

4.1 F - Measure (F1 score): It is the combination of precision and recall of the model, and it also is defined as the harmonic mean of the model's precision and recall.
We want our model to have high precision and the analogy behind it is that the model should be able to precisely predict the correct aspect category of the review sentence. It shouldn't be categorizing the *service* aspect to the *food* aspect. At the same time the model should be able to recall all the sentences related to that aspect. Therefore, F - Measure evaluation metric

4.2 Manual Inspection: As we are using large-scale unlabeled data so we cannot apply conventional evaluation metric on our data due to unavailability of labels.
Therefore,human inspection is required to tune the model.

5. Real world constraints and expectations

5.1 Latency requirements:
There is no need for real time generation of aspects or summarizing the reviews at the moment. So we can create a data store to save the model's output and create an IR system to retrieve the result to display on a dashboard.

5.2 Expectations:

- Proposed model is able to extract all the relevant aspects and should be able to exclude irrelevant aspects.
- It should be able to handle large-scale datasets with sensible results.

6. References

- Unsupervised Semantic Approach of Aspect-Based Sentiment Analysis for Large-Scale User Reviews | IEEE Journals & Magazine | IEEE Xplore
- [1812.03361v2] An Unsupervised Approach for Aspect Category Detection Using Soft Cosine Similarity Measure (arxiv.org)
- Aspect Extraction in Customer Reviews Using Syntactic Pattern - ScienceDirect
- Performing effective Aspect Based Sentiment Analysis | Board Infinity - YouTube
- Understanding Aspect Based Sentiment Analysis - YouTube
- Sentiment Analysis is not enough!! | by Dhruv Gangwani | DataDrivenInvestor

- Aspect-based Sentiment Analysis — Everything You Wanted to Know! | by Intellica.AI | Medium
- Aspect Based Sentiment Analysis with Machine Learning | 47Billion | by Chetan Borse | 47Billion | Medium
- Aspect Extraction and Opinion Analysis (achyutjoshi.github.io)

Phase 2: EDA and Feature Extraction

Data Description:

The data for this experimentation is collected using google place API. It consists of 5 automotive service location's review data.

Total number of records in in the dataset: **109923**
Note: Dataset does not contain any labeled output.

Dataset contains the following columns:

- location id - Indicates which location the record belongs to.
- comment title
- comment

Sample Data:

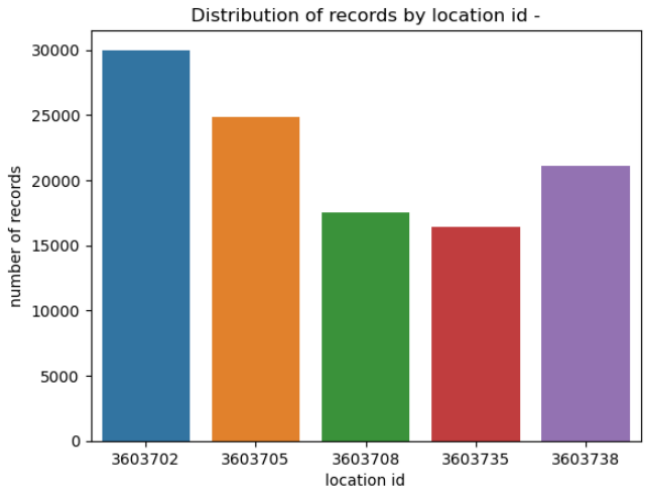
	location_id	comment_title	comment
3571	144196_72	NaN	I actually pushed my Onstar SOS button and was...
38266	144196_75	NaN	They answered my question in a courteous manner.

Null count :

```
location_id          0
comment_title    109923
comment           55010
```

- There is no comment title available for any of the comment
- Half of the comment data is also missing

Distribution of records by location id :



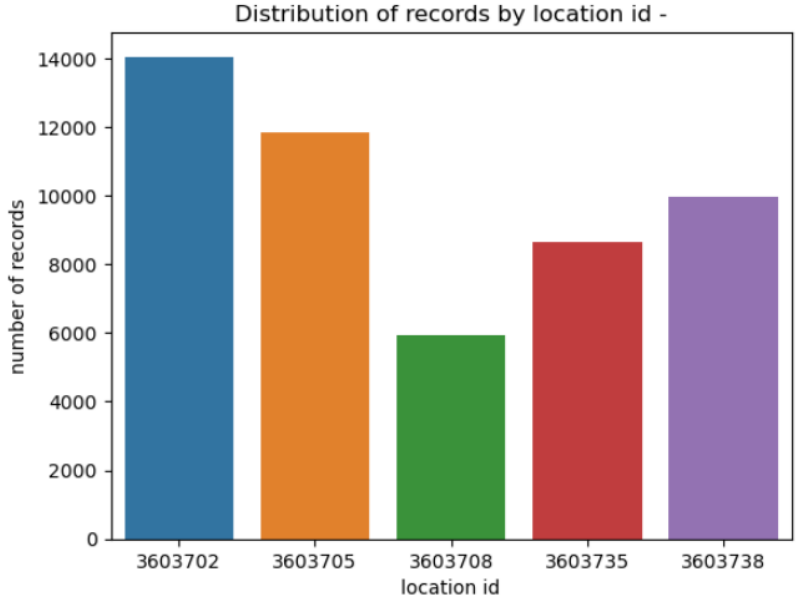

```
value counts:

3603702    30000
3603705    24904
3603738    21113
3603708    17505
3603735    16401
```

Number of records containing duplicate comments: **4519**

Number of records after removing null comments and duplicates: **50394**

Distribution of records by location id (after removing null comments and duplicates) :




```
value counts:

3603702    14043
3603705    11821
3603738     9957
3603735     8642
3603708     5931
```

Basic Data Cleaning Steps:

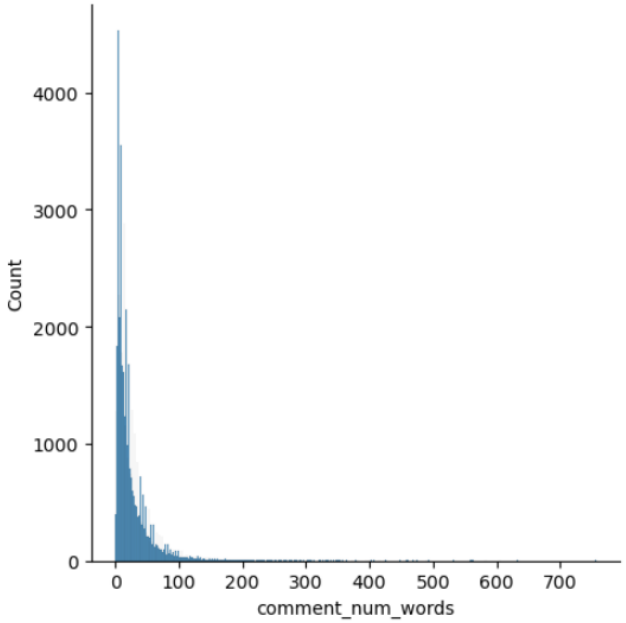
- Remove Null and Duplicate comments
- Expand contractions
- Lowercase the reviews
- Remove digits and words containing digits
- Remove punctuations
- Change the emoticons with words

Univariate Feature Analysis:

As *comment* is our primary feature in this project, we can perform all sorts of analysis which can decode the insights of text data. Below are the few example -

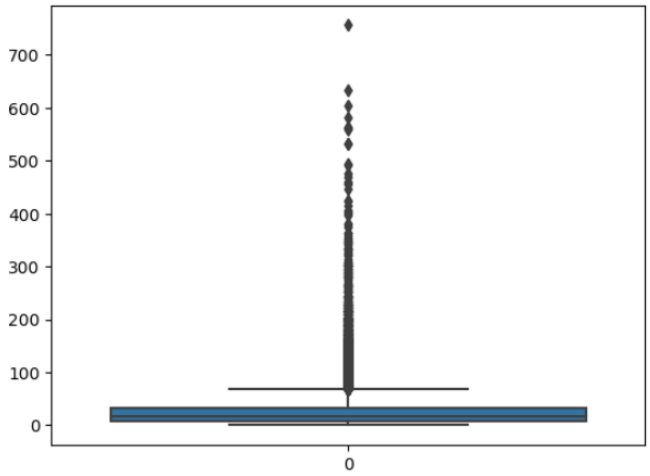
1. Word count and character count
2. Box plot of word length
3. Stop word counts
4. WorldCloud
5. Unigram Analysis
6. Bigram Analysis
7. Trigram Analysis
8. Polarity Analysis

1. Word count



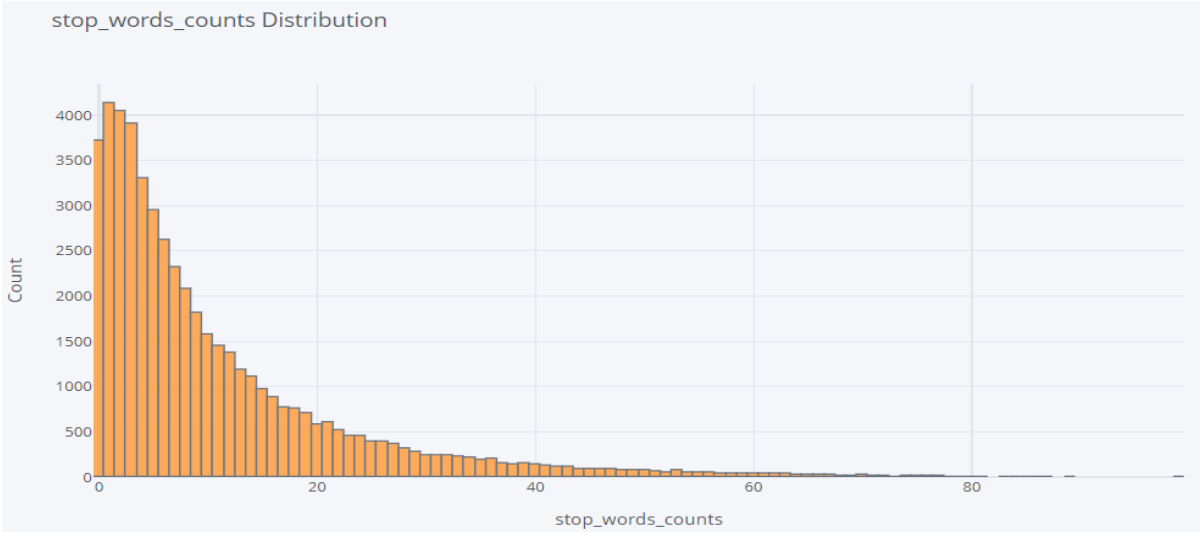
- We can observe that most of the comments have word length less than 150 words but there are a few comments which have very high word length. Those reviews may not be useful as it may contain only noise.

2. Box plot of word length



- Box plot of word length can help us to detect those records where comment length is abnormally high.
- We can observe that 99 percent of data having word length less than equals 152 words. So it is better to remove those comments from the dataset which is having skewed word counts as it won't add much to the overall result.

3. Stop word counts



- Again, just like word count distribution plot, stop count distribution plot also follows pareto distribution.

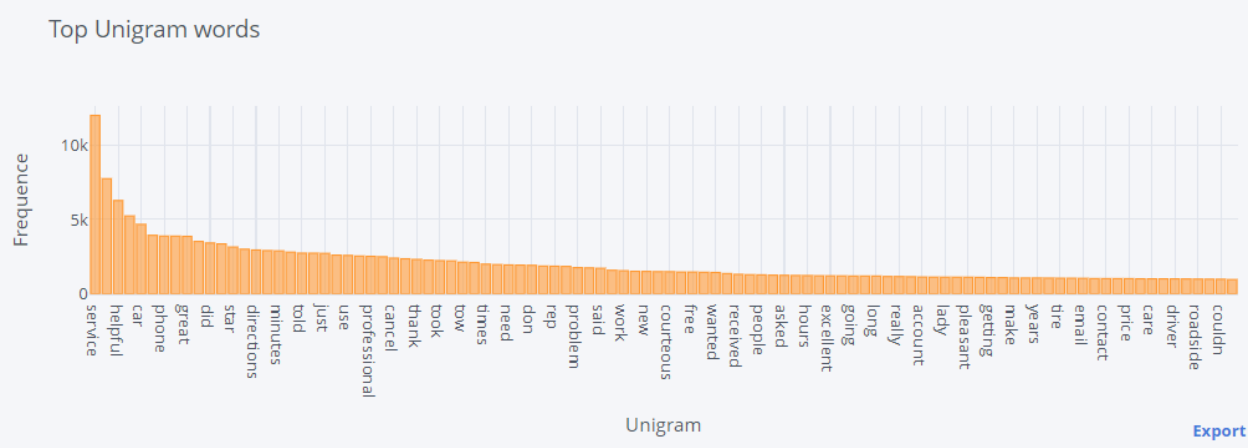
4. WorldCloud

- WordCloud visually represents the frequency of different words present in a document.
- It gives importance to the more frequent words which are bigger in size compared to other less frequent words.
- It is important to remove all the stop words before visualizing the WordCloud plot

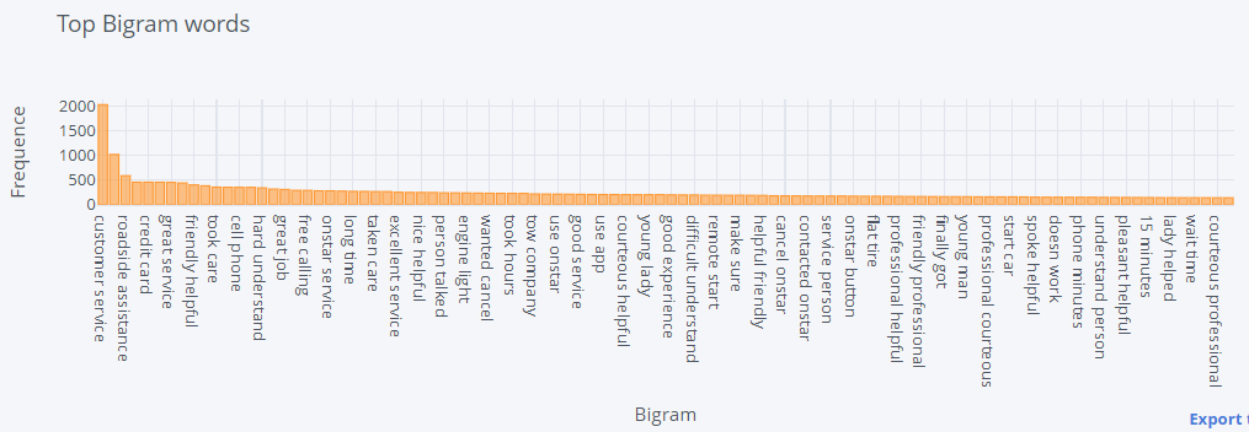
- Plot -

- We can observe that words like *Service, OnStar, Helpful, Call, Vehicle* are the most used words.
- It is a bit difficult to get top 100 frequequent words along with their frequency just using this word cloud plot. There comes the need of unigram frequency analysis.

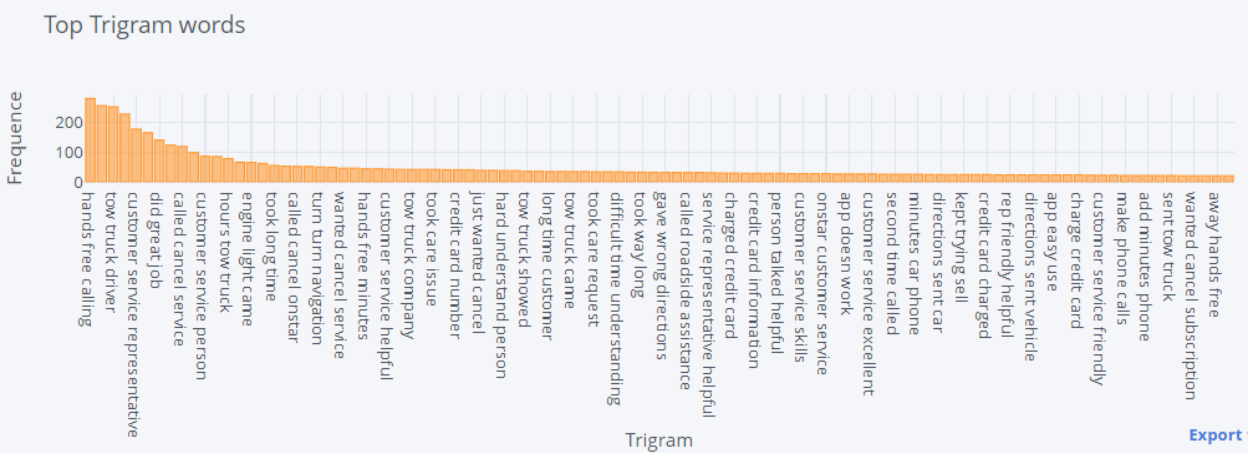
5. Unigram Analysis



6. Bigram Analysis



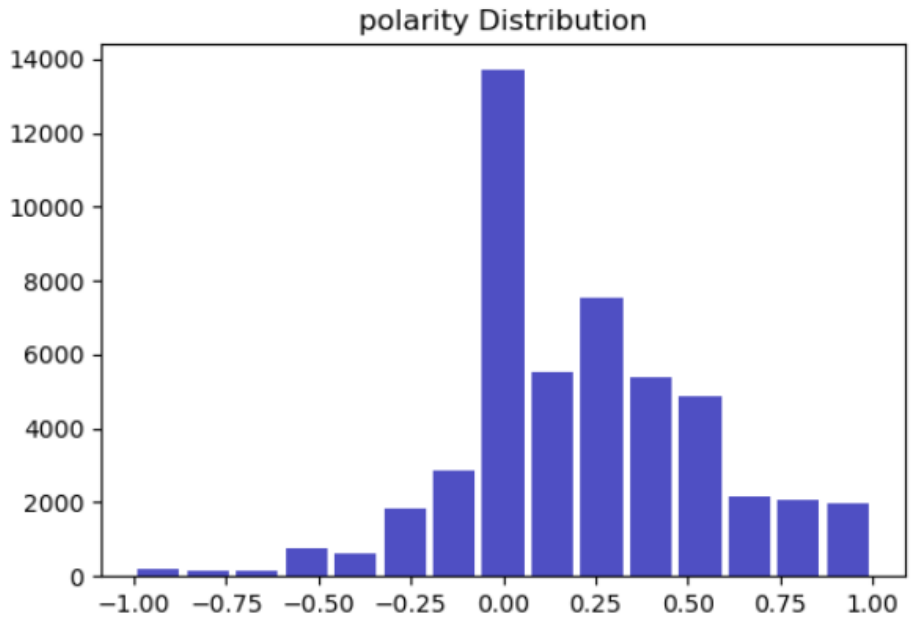
7. Trigram Analysis



- N-gram analysis is the frequency based approach to understand the most used words.
- This type of analysis can help to get the idea of what sort of Aspects people are talking about.
- We can infer from n-gram analysis that comments have mostly belows topics -
    1. Customer service
    2. Application (App)
    3. Assistance and Speed
    4. Service provider (OnStar)
    5. Auto Parts (like flat tire, engine light)
    6. Skills and competence
- These insights can help us to choose appropriate aspect terms.

8. Polarity Distribution Plot

- Help us to infer weather people are taking good or bad
- We can observe that a lot of the comments have neutral sentiments.
- We have more comments having positive sentiments.



Feature encoding:

As our data consists of text data only, we can encode it using below 2 approaches:

1. Frequency Based (E.g BOW)
    a. It represents the word using its occurrence.
    b. The dimension of the features could be very high because of the large vocab size and It usually creates the sparsity.
    c. It also doesn't capture the semantic essence of the word.
2. Semantic Relation Based (E.g W2V and BERT)
    a. Word embeddings aim to mapping semantic meaning of words into a geometric space. The geometric space formed by these vectors is called an embedding space

- Our experiment approach is trying to find the aspects using both frequency based approach and semantic relation based approach. So we will incorporate both techniques one way or another.
- For frequency based approach, we will count only frequency of noun and noun adjectives using syntactic grammar rules.
- For Semantic Relation Based approach, we will use custom trained w2v, pre trained w2v and pertain bert embeddings.

High-dimensional data visualization:

Below are the following techniques that can be used for visualizing high dimensional data.

1. PCA
2. tSNE

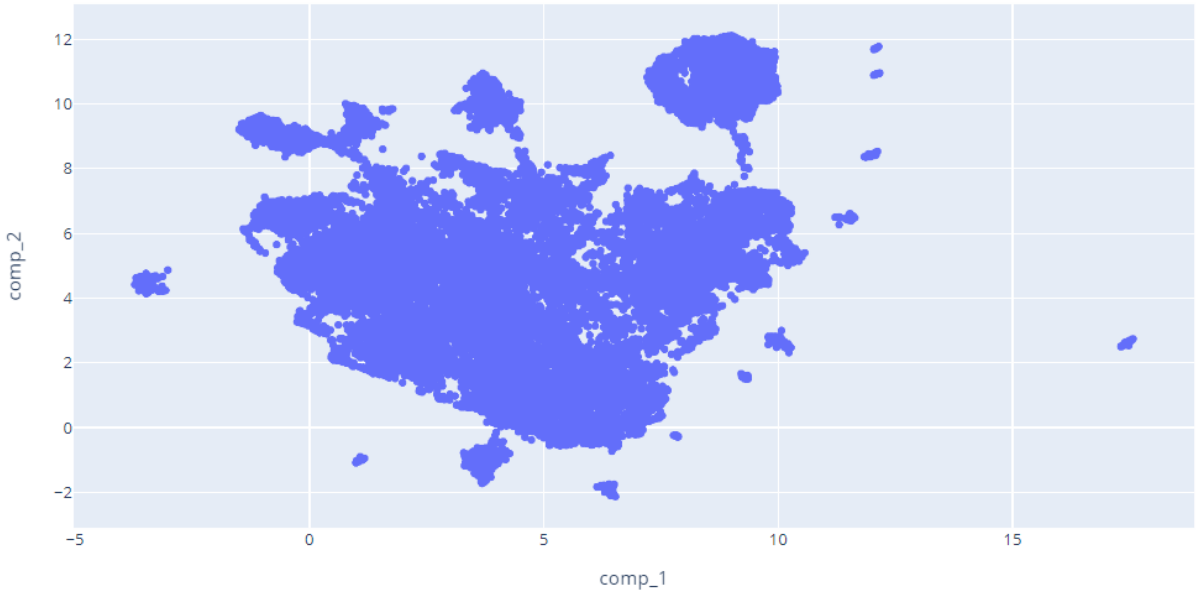
3. uMAP

For this experiment we are using the UMAP technique to visualize high dimensional text data. uMAP is very similar to tSNE but works slightly faster than tSNE. The main advantage of tSNE types of technique over PCA is that it captures the distance of the neighborhood. That means if two words have the same meaning then It will try to preserve this information in the projection.

Plot -

- Split the reviews into sentences
- Get the embeddings using pretraing bert sentence encoder
- Custer visualization using uMap

Cluster Visualisation using BERT Embeddings of sentences:



Phase 3: Modeling and Error Analysis

Our experiment consists of 3 tasks:

1. Aspect Extraction
2. Aspect Inference
3. Aspect's Sentiment detection

Each step involves a model and corresponding error. This report we will explore each step one by one and explain the experiment that we have conducted.

1. Aspect Extraction

For this step we have taken the reference from this paper and written the algorithm from scratch.

Our approach is the hybrid approach which combines frequency-based approach (word level) and a syntactic-relation based approach (sentence level) which is further enhanced using semantic similarity based approach. The core idea of this approach is to extract all the aspects related to the domain without using any fixed list of aspects and also extract that aspect which is not frequent. And produce sensible results even on large-scale unlabeled data.

Core idea for Extracting Aspects -

1. Hybrid combination of both frequency based approach (word level) and syntactic-relation based approach (sentence level) followed by semantic similarity based approach.
2. Model should be able to Eliminate all the irrelevant aspects and should be able to extract all the semantically related aspects.

Understanding the approach in detail -

There is three main component of this step:

1.1. Extracting aspect terms by extracting noun and noun-adjective pairs

- Extract noun and noun-adjective pairs as those have potential to be classified as aspects.

  Two approaches are used:
  1. Frequency based approach (FBA)
  2. Syntactic-relation based approach (SRBA) - Uses Dependency parser
  - Both of these approaches work in parallel, also called Parallel Hybridization Approach (PHA)

1.2. Extracting aspect terms using similarity based approach

- Using pre-trained embedding, get those words which are semantically similar to domain aspects.

1.3. Merging both the dictionary and consolidating a final aspect dictionary

Below is the flow chart of mentioned approach -

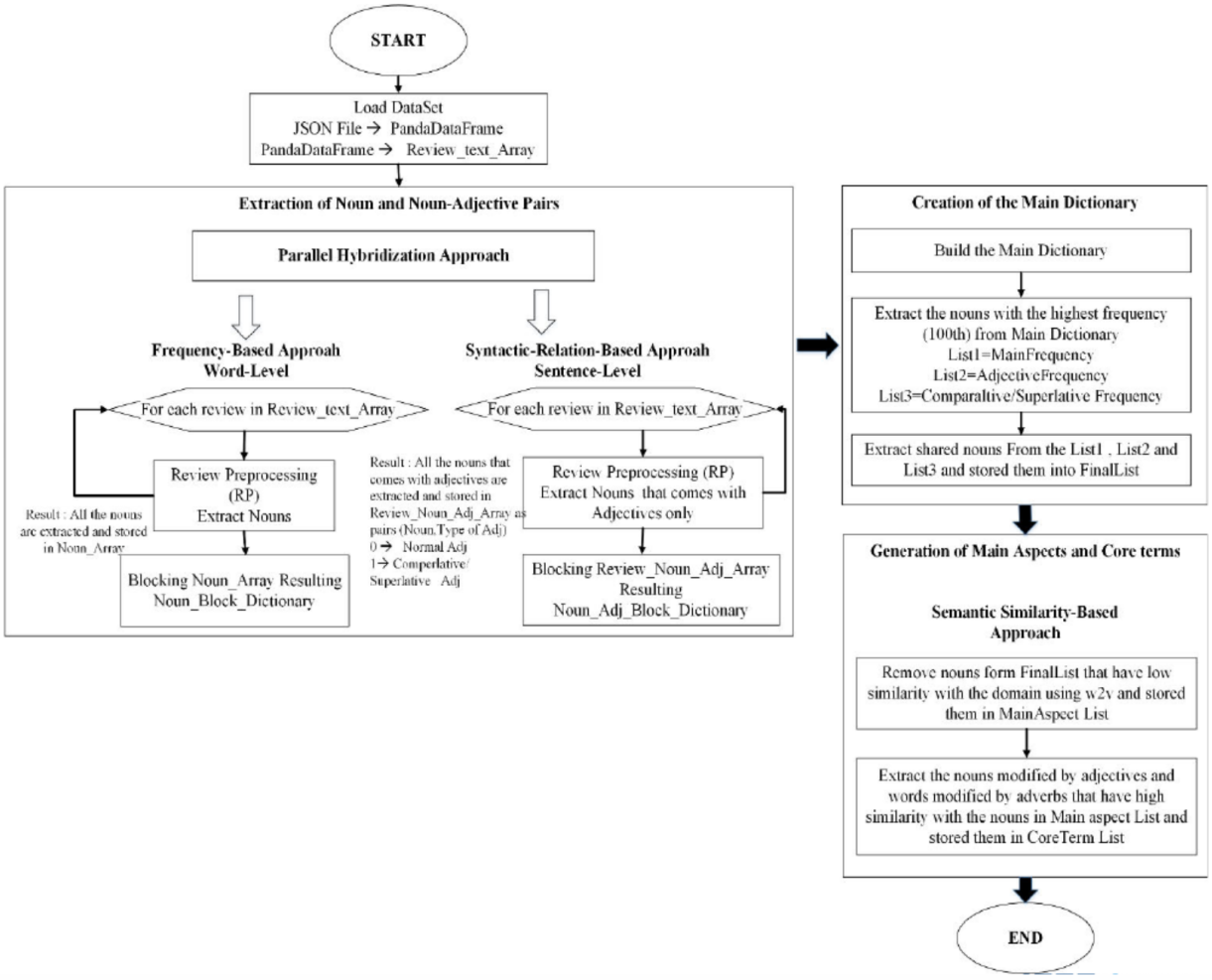


FIGURE 1. **The general approach of the semantically enhanced aspect extraction method.**

Below are the list of domain relevant aspects -

```
['Sales',
 'Service',
 'Parts',
 'Leasing',
 'Finance',
 'Customer',
 'Competence',
 'Pricing',
 'Administration',
 'Speed',
 'Suggestion',
 'Appearance',
 'Cleanliness',
 'Test Drive',
 'Appointment',
 'Tactics',
 'Appearance',
 'Amenities',
 'Cleanliness',
 'Location',
 'Parking',
 'Experience',
 'Recommend']
```

Choosing only 8 aspects for the experiment. Below are the list -

1. Finance
2. Administration
3. Competence
4. Sales
5. Pricing
6. Service
7. Customer
8. Experience

For the similarity based approach, we have experimented with 3 models.

1. Pre trained Word2Vec embedding trained on google news data
2. Custom trained Word2Vec model on our own dataset
3. Using pre trained bert embeddings

As we don't have the labeled data for this experiment, we have to manually inspect the result by eyeballing it.

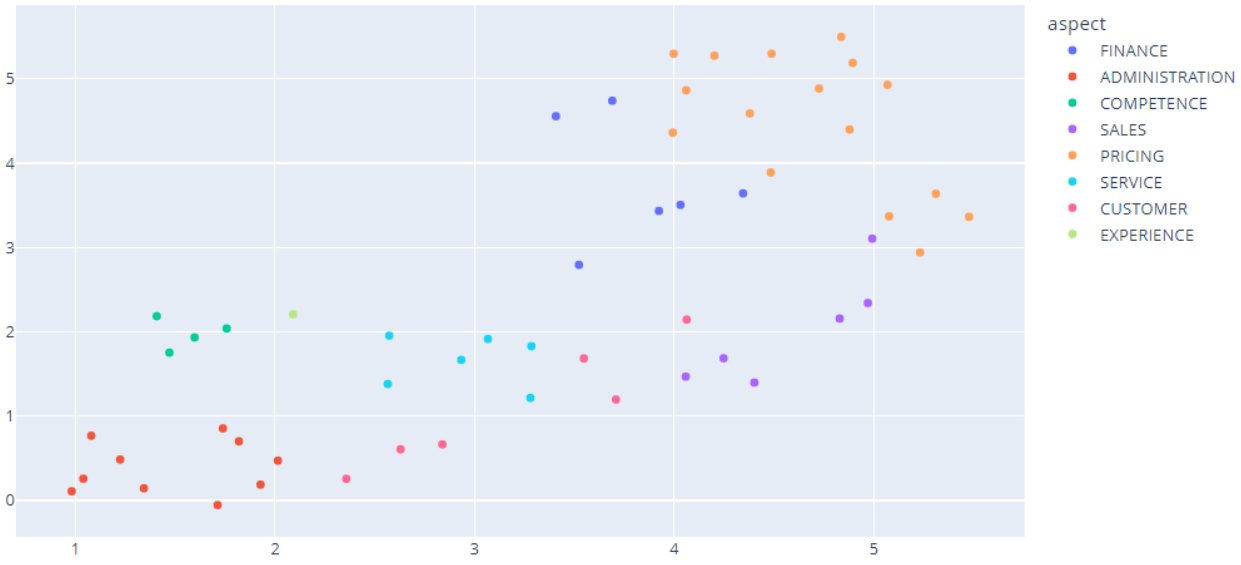
Findings -

- Both pre-trained word2vec embeddings and custom trained word2vec embeddings didn't work well as per the experiment.
- Whereas, bert embedding gave the most appropriate result among all three.

Below are the final list of aspects and aspect' terms extracted using mentioned approach and various experiments -

	aspects	aspect terms
0	FINANCE	[money, credit, asset, income, billing, bank]
1	ADMINISTRATION	[advisor, representative, authority, adviser, department, staff, assistant, administration, agency, agent]
2	COMPETENCE	[knowledge, capability, confidence, skill]
3	SALES	[buyer, advertising, dealership, business, dealer, promotion]
4	PRICING	[fee, afford, dollar, purchase, rate, deal, buy, price, cost, expense, payment, charge, discount, budget, amount, buying]
5	SERVICE	[system, provider, job, service, aervice, assistance]
6	CUSTOMER	[person, company, product, customer, employee, someone]
7	EXPERIENCE	[experience]

Aspect's cluster visualization :



2. Aspect Inference

Objective is to infer the most appropriate aspect for each sentence of the comments.

To achieve this task, we have followed the below steps

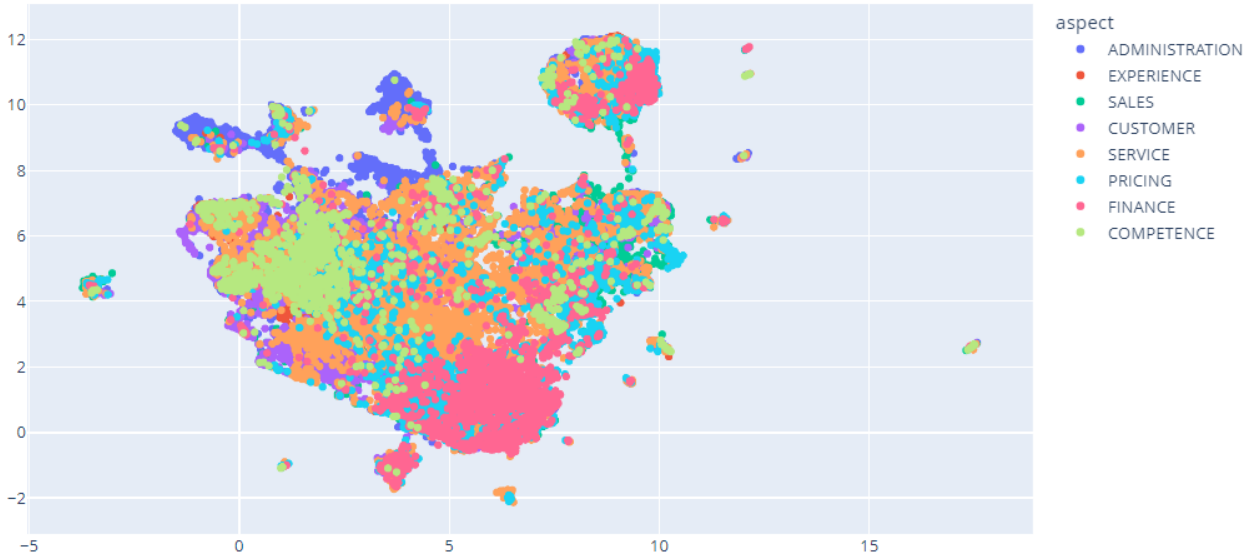
- Split the reviews into sentences using NLTK.
- Encoded each sentence using pre-trained BERT.
- Encoded each aspect term using the same pre-trained model.
- Find the cosine similarity b/w sentence and aspect terms.
- Assign the closest aspect to the sentence if the similarity score is greater than the threshold otherwise assign it as *miscellaneous* aspect category.

Getting an appropriate threshold could be a bit tricky at first. But with some hit and trail and after eyeballing the result it could be done.

Sample Result:

	sentences	aspect	aspect_score	aspect_seed_word
38521	Not sure if this was an issue with the app or the satellite service associated with that function.	SERVICE	0.267794	provider
33259	they took care of my order in a matter of minutes.	CUSTOMER	0.299488	customer
18991	I downloaded the My Chevy app, but it says I'm not active	MISCELLANEOUS	0.209404	dealership
38065	Charged for a free trial was told was ended.	PRICING	0.359957	fee
42735	Easy to deal with and very knowledgeable	COMPETENCE	0.614274	knowledge

Cluster Visualization of Aspects



3. Aspect's Sentiment detection

- Using pre trained DistilBERT model to find the sentiment of a sentence and its corresponding aspect
- Ref: Hugging face [sentiment pipeline]

Sample output

sentences	clean_sentences	Ind	aspect	aspect_score	aspect_seed_word	aspect_sentiment_score
The representative that helped me was amazing.	the representative that helped me was amazing	0	ADMINISTRATION	0.604803	representative	5
She was kind, professional and went over and b...	she was kind professional and went over and be...	1	EXPERIENCE	0.302708	experience	3

Pros and cons of this model

Pros -

1. Model is able to capture the semantically similar aspects unlike other approaches like frequency based approach or syntactic relation based approach only.
2. Works great when we don't have unlabeled large scale data because of blocking techniques used in frequency based approach (word level) and syntactic-relation based approach (sentence level).

3. This model is very simple and intuitive and can be easily used for other domains/industry, given some domain aspects.

Cons -

1. As we are using an unsupervised learning approach to assign the aspect and calculate sentiment of each sentence, the result is not very accurate.
2. Because of lack of labeled data we cannot fine tune the bert embeddings further to get the better result.

Phase 4: Advanced Modeling and Feature Engineering

We have seen that an unsupervised based approach is simple, intuitive and gives a pretty decent result. But the result can be significantly improved if we would have had labeled data.

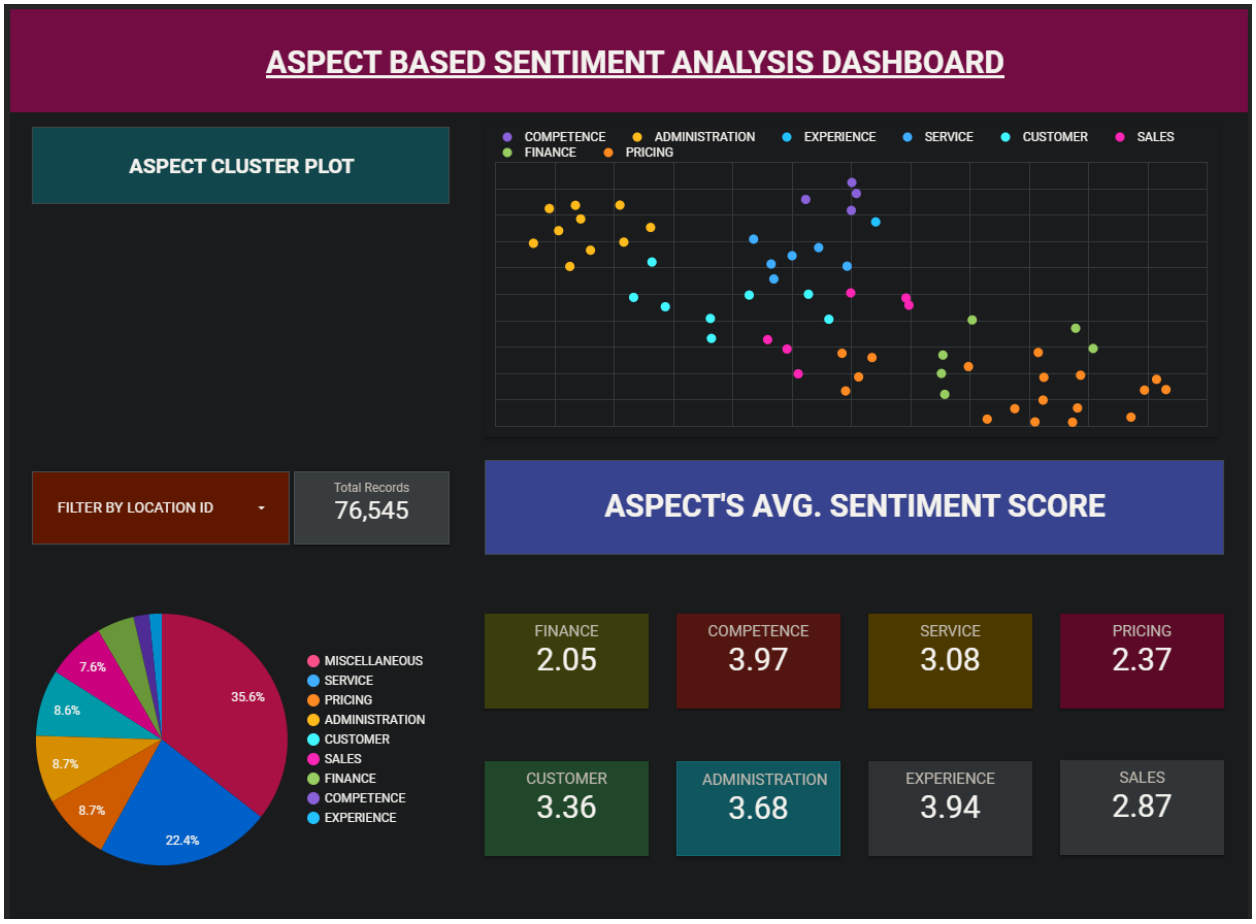
That's why in order to improve the overall result we only experimented with the pretrained embeddings. And finally finalized that embedding which gave the most appropriate result.

- For Embedding we have used *all-MiniLM-L6-v2* model.
- For Sentiment calculation we have used *nlptown/bert-base-multilingual-uncased-sentiment model.*

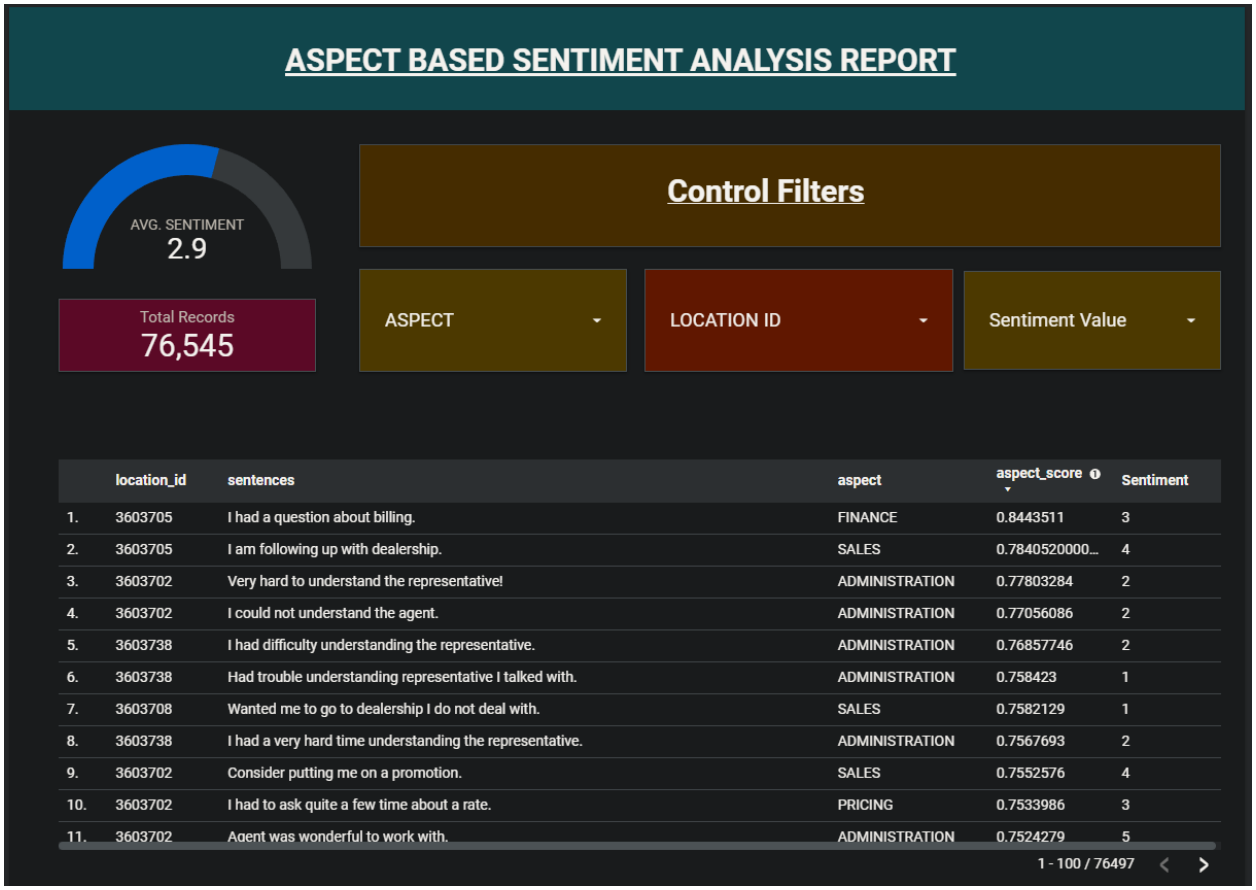
Phase 5 and 6: Deployment and Productionisation &Project

Our eye catching experiment for this project is an interacting report. It automatically calculates each aspect's average sentiment score by selecting a location id.

We have used DataStudio to create the report. Here is the link to the dashboard.



Snapshot 1



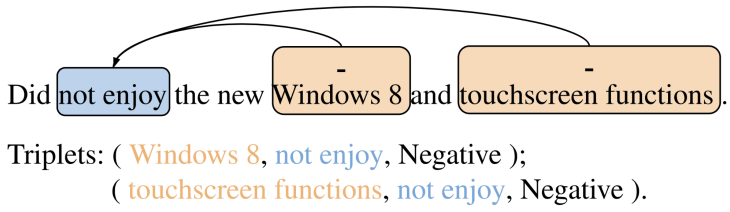
Snapshot 2

Further approach is to label some data with the help of domain experts and use some supervised learning based approach.

One such approach is [Span-Level Interactions for Aspect Sentiment Triplet Extraction].

The goal of this approach is to extract sentiment triplets of the format (aspect target, opinion expression and sentiment polarity)

Snapshot from paper:



These types of models are excellent in order to get the best results but the drawbacks from this type of approach is that it requires exhaustive labeling and may not be applicable to large scale data and the learning can not be transferred to other industries/domains easily.