Text Mining to decipher Insights from Free-Response Consumer Complaints

A Project Report

Submitted in partial fulfilment of requirements for the degree of Master of Management

By Shaunak Handa

SR No. 05-10-04-10-81-17-1-14693

Under the guidance of

Prof. Parthasarathy Ramachandran

Department of Management Studies,

Indian Institute of Science

Bangalore – 560012

June 2019

Table of Contents

Acknowledgements

I would like to extend my deepest gratitude to Prof. Parthasarathy Ramachandran for his guidance and support throughout this project. It is an honor to work under him.

I take immense pleasure in thanking my external examiner Prof. Shashi Jain for his views on my work. His suggestions were very helpful in shaping the project.

Most of all, I would like to my thank my family for their blessings that saw me through all the difficult times.

Shaunak Handa IISc, Bangalore June 26, 2019

Abstract

The automobile industry with its recent advances including the electric vehicles, hybrid vehicles, navigation systems, cruise control and currently moving towards self-driving vehicles look to be promising factors which are going to influence the driving experience in the upcoming years. Technological changes in the past have contributed significantly to automotive safety and are expected to continue the same, however the full effect of these impressive contributions is uncertain. Customer surveys and complaint databases containing free-response incident descriptions and running an analysis on them could act as valuable resources to ascertain the same.

With today's world moving towards using electronic media for storing textual information and documents due to ease, advancement and availability of technology, in addition to the reduced storage costs. Finding out relevant information from a collection of unstructured documents is time consuming for the reader as opposed to from a large collection, ordered or classified by group or category which is less time consuming. There still exists the problem of converting unstructured textual data into structured textual data (in layman terms identifying best such grouping). Summarizing and clustering of free-response databases is prohibitively time consuming and difficult, thus Text mining can help reveal useful information, extending human analysis capabilities for large free-response databases to support earlier detection of problems and more timely safety interventions.

The aim of this study is to extract clusters of vehicle problems from the free-response data in the National Highway Traffic Safety Administration's vehicle owner's complaint database by applying text mining, thus generating any associated trends from the data if possible. Textual data is one of the most unstructured data's available and requires some preprocessing. The Free-response complaint data were pre-processed using the standard Natural language processing techniques (Explained later). Later in the project, Clustering using one of its popular variants, K-Means, the pre-processed Free-response complaint data is finally separated into similar groups. The clusters were labelled using Word Cloud. K-Means being an unsupervised learning, validating clusters is possible by comparing insights produced post to the Vehicle Model or Manufacture level defects and any recalls that occurred during that time-period within the automobile industry.

List of Tables

List of figures

Chapter One

1.1 Introduction

1.1.1 Background

Data invariably helps us in many places or circumstances. Having data can be considered a boon. Having adequate data helps in many ways like analysis, forecasting and what not. It helps in understanding the situations better, it helps in understanding some patterns like in stocks, oil price and the list goes on. Earlier, data was very scarce. There wasn't adequate amount of data available for analysis, as a matter of fact any other job involving the need of data. Having enough data can even play an important role in saving lives. If provided with the data about weather patterns and all, tsunamis can be predicted and required safety measures can be taken. Things can be handled. All these things put emphasis on the importance of data.

With the advent of internet and many more modern technologies, promising results have been shown. A lot of data is now readily available over the internet and most of it is free to access for everyone. According to Forbes article by Bernard Marr, there are about two and a half million bytes of data created every day at the current pace. The article says that 90% of the data available over the internet is created only in the past two years (2017-19). This will of course commensurate with time.

This increase in data might seem like a boon and it is. But the only problem is that large portion (almost all of it) of the data available is unstructured. This is one of the biggest problems today. Unstructured data is the data which is not in a pre-defined way or simply not organized. This problem is in dire need to be addressed. A lot of unstructured textual is also available which need to be organized for usage in future. Hence the techniques or methods which can be used to mollify this problem are very much necessary.

1.1.2 Problem statement

As mentioned in the above section, organization of the unstructured textual data for usage in future is required. The situation can be eased up to a certain extent by Clustering and labeling them appropriately. The result of this clustering and labeling is structured data created from unstructured data, thus adding meaning to the unstructured data. This labeling would assist in identifying the documents or a cluster of words which would further help the users get a basic idea or overview of what these documents are or what is this cluster about. This will help structuring the textual data, thus improving the user readability. This project considers the clustering and labeling of complaint database identifying for trends support earlier detection of problems and more timely safety interventions.

1.1.3 Literature review

Previous work explores the utility of computerized text analysis techniques in decoding narrative data related to incidents like machine learning methods to categorize motor vehicle crashes (maintained by an insurance company), occupational injuries (filed with a workers' compensation insurance provider), and general population injuries (from the U.S. National Health Interview Survey [NHIS]; Lehto, Marucci-Wellman, & Corns, 2009; Lehto & Sorock, 1996; Noorinaeini & Lehto, 2006; Wellman, Lehto, Sorock, & Smith, 2004). A variety of Bayesian and regression methods were used (i.e., fuzzy Bayesian, naïve Bayesian, singular value decomposition [SVD] Bayesian, and SVD regression) to assign narratives to a preexisting set of categories. The results showed that these models were highly sensitive and accurate, producing results that agreed with expert assignments in as many as 90% of cases. Review of textual analysis in health care research found the data were beneficial for identifying injury cases, extracting additional information about cases, and assessing the accuracy of a numerical coding scheme for injury identification (McKenzie, Scott, Campbell, & McClure, 2010). Such semiautomated approaches using a combination of predetermined categories and statistical methods (e.g., Bayesian models and clustering) have clear value.

A more automated approach would use the textual data to identify coherent and correlated subsets of incidents, without committing to a predefined set of categories. Such an approach can highlight trends in particular incident clusters that could go unnoticed if the narratives are indexed by predefined categories. This study follows such an approach, using latent semantic analysis (LSA) to identify similar incidents (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Dumais, 2004; Dumais, Furnas, Landauer, Deerwester, & Harshman, 1988). LSA uses only word co-occurrence to assess the semantic similarity of incidents and does not consider grammatical structure or the sequential relationship between terms (Deerwester et al., 1990). This "bag of words" representation of documents makes it possible to treat text as numerical data.

LSA has been successfully used in classifying and retrieving text-based patient record data (Chute, Yang, & Evans, 1991) as well as extracting patterns and concepts from psychiatric narratives (Cohen, Blatter, & Patel, 2008). LSA has also been used to flag fall-related injury cases based on unstructured text-based medical records (Tremblay, Berndt, Luther, Foulis,

& French, 2009). In general, LSA has the capacity to analyze collections of hundreds of thousands of documents, but collections of millions of documents might require sampling or other simplifying techniques (Dumais, 2004).

Recently, there has also been a paper (Vishnu, Konda, 2019), which uses K-Means clustering for Automated Text Clustering and then Labeling using Hypernyms.

Text Mining to Decipher Free-Response Consumer Complaints: Insights from the NHTSA Vehicle Owner's Complaint Database, a paper in The Journal of the Human Factors and Ergonomics Society · September 2014, uses LSA to reduce the features. Thereafter, it uses Hierarchical clustering for clustering the complaints, but it only focuses on 2 severity levels of complaints mainly the fatal and the injured categories.

1.1.4 Flow of project report

I have described the data collection methodology and a brief discussion of the data set has been done in the later parts of chapter 1. This work takes a corpus that is a collection of documents which will be the data to be clustered and labeled. The next step involves calculating the Term Frequency-Inverse document frequency (TF-IDF) and transforming the data into one or more clusters. Before calculating the TF-IDF, the text data is preprocessed using some standard techniques. A step by step approach for the same has been discussed in Chapter 2. Also, Chapter 2, has an introduction to Clustering and its types has been given with a detail about K-Means, which has been used to cluster the complaints in our dataset. Chapter 3 describes the steps discussed in Chapter2 specific to the model for this project. Finally, Post clustering, Word Clouds are generated to label/identify the clusters and generate insights from the data, description for the same has been provided in chapter 4.

1.2 Data description

Source: Data is from The National Highway Traffic Safety Administration's (NHTSA) vehicle owner's complaint database.

- The NHTSA's vehicle owner's complaint database encompasses approx. 100,000 incident reports (as of December 29, 1997). (Time period for Complaints – 1983 to 1997), based on Vehicle Owner's Questionnaire (VOQ) complaint entries.
- The complaints have been filed through NHTSA's Internet Vehicle Owner's Questionnaire (IVOQ; NHTSA, Office of Defects Investigation [ODI], n.d.), hotline VOQ, consumer letters, and other channels since January 1, 1995. Primarily used for defect screening and investigations or unreasonable safety risk.
- The vehicle owner's complaint database includes information regarding the vehicle (e.g., manufacturer's name, vehicle make and model, and model year), incident (e.g., involvement of fire or crash), number of injuries and/or deaths, fuel type, and other descriptors of the incident.
- In addition to this categorical information, the database contains a field, "Description of the Complaint," that describes consumers' account of the vehicle problem and its consequences.

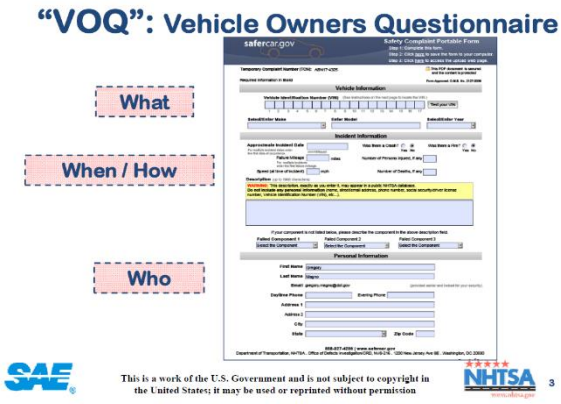


Fig- 1.1 Sample Questionnaire as prepared for free-response complaints

Variable Name	Type/Size	Description	Missing Values (%)
CMPLID	CHAR(9)	COMPLAINT ID	0
MFR_NAME	CHAR(40)	MANUFACTURER'S NAME(GENERAL MOTORS CORP)	0
MAKETXT	CHAR(25)	VEHICLE/EQUIPMENT MAKE(CHEVROLET)	0
MODELTXT	CHAR(256)	VEHICLE/EQUIPMENT MODEL(LEXUS, INNOVA)	0
CDESCR	CHAR(2048)	DESCRIPTION OF THE COMPLAINT	0
YEARTXT	CHAR(4)	MODEL YEAR, 9999 IF UNKNOWN OR N/A	2
FAILDATE	CHAR(8)	DATE OF INCIDENT (YYYYMMDD)	20
DATEA	CHAR(8)	DATE ADDED TO FILE (YYYYMMDD)	0
LDATE	CHAR(8)	DATE COMPLAINT RECEIVED BY NHTSA (YYYYMMDD)	0
FIRE	CHAR (1)	WAS VEHICLE INVOLVED IN A FIRE 'Y' OR 'N'	7
INJURED	NUMBER(2)	NUMBER OF PERSONS INJURED	26
DEATHS	NUMBER(2)	NUMBER OF FATALITIES	23
CITY	CHAR(30)	CONSUMER'S CITY	0
STATE	CHAR(2)	CONSUMER'S STATE CODE	0
ODINO	CHAR(9)	NHTSA'S INTERNAL REFERENCE NUMBER. THIS NUMBER MAY BE REPEATED FOR MULTIPLE COMPONENTS. ALSO, IF LDATE IS PRIOR TO DEC 15, 2002, THIS NUMBER MAY BE REPEATED FOR MULTIPLE PRODUCTS OWNED BY THE SAME COMPLAINANT.	0
CMPL_TYPE	CHAR(4)	SOURCE OF COMPLAINT CODE: CAG =CONSUMER ACTION GROUP CON =FORWARDED FROM A CONGRESSIONAL OFFICE DP =DEFECT PETITION,RESULT OF A DEFECT PETITION EVOQ =HOTLINE VOQ EWR =EARLY WARNING REPORTING INS =INSURANCE COMPANY	0

		IVOQ =NHTSA WEB SITE LETR =CONSUMER LETTER MAVQ =NHTSA MOBILE APP MIVQ =NHTSA MOBILE APP MVOQ =OPTICAL MARKED VOQ RC =RECALL COMPLAINT,RESULT OF A RECALL INVESTIGATION RP =RECALL PETITION,RESULT OF A RECALL PETITION SVOQ =PORTABLE SAFETY COMPLAINT FORM (PDF)VOQ =NHTSA VEHICLE OWNERS QUESTIONNAIRE	
PROD_TYPE	CHAR(4)	PRODUCT TYPE CODE: V =VEHICLE T =TIRES E =EQUIPMENT C =CHILD RESTRAINT	0

Table 1.1 Variable Name, Description and Missing %

The additional variables provided in the dataset had missing %percentage > 40% and this have been ignored for any analysis.

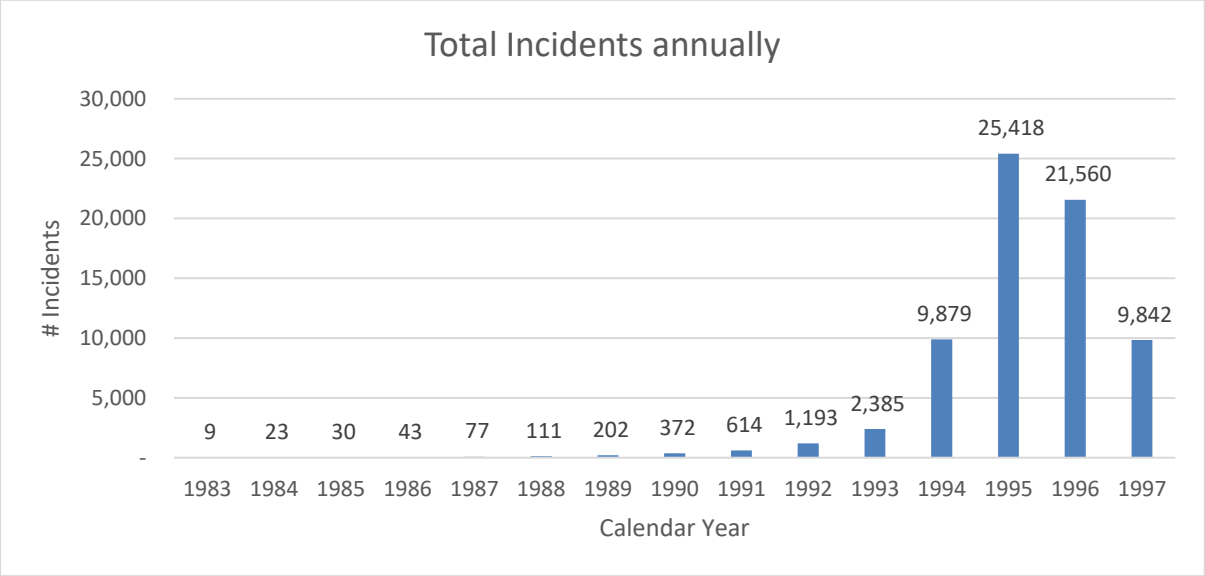


Fig- 1.2 Graph depicts the increasing # of annual incidents

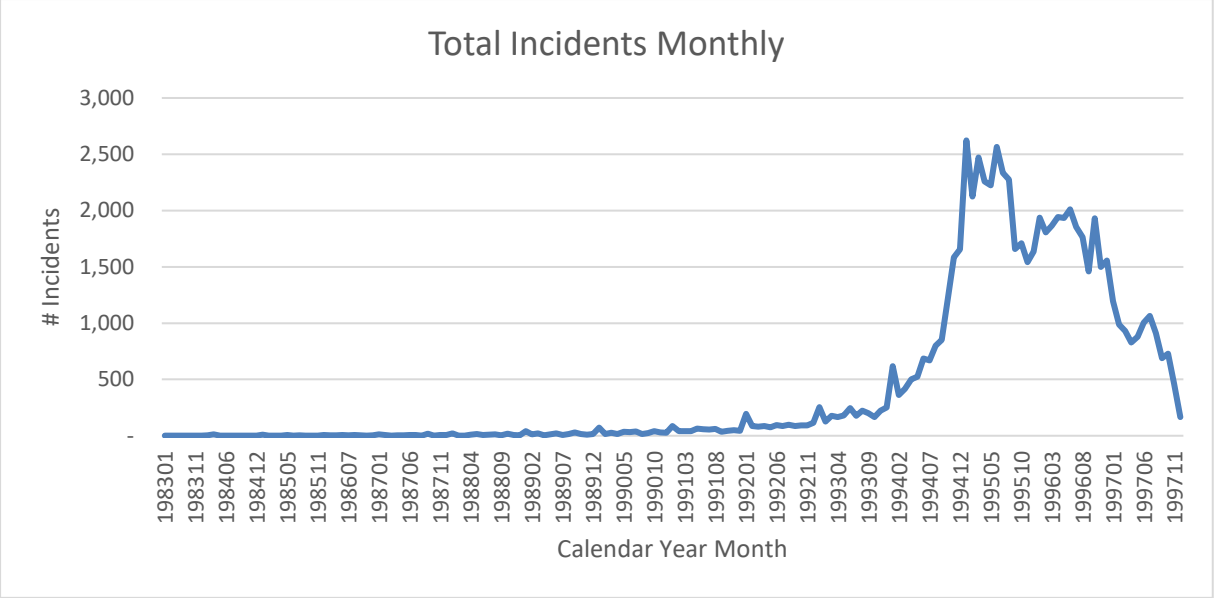


Fig- 1.3 Graph depicts the trend for monthly incidents

From figure 1.2 and 1.3, one can observe the rapidly increasing complaint volume from the period starting in the early 1990's. The incidents post 1990 have doubled annually, this can be due to sales of vehicles rising but this trend should have rung the bells calling for investigation into finding the causes and reasons behind this. As conveyed by figure 1.5, one can observe that the US car sales have not increased drastically keeping in accordance with the increase in the number of complaints.

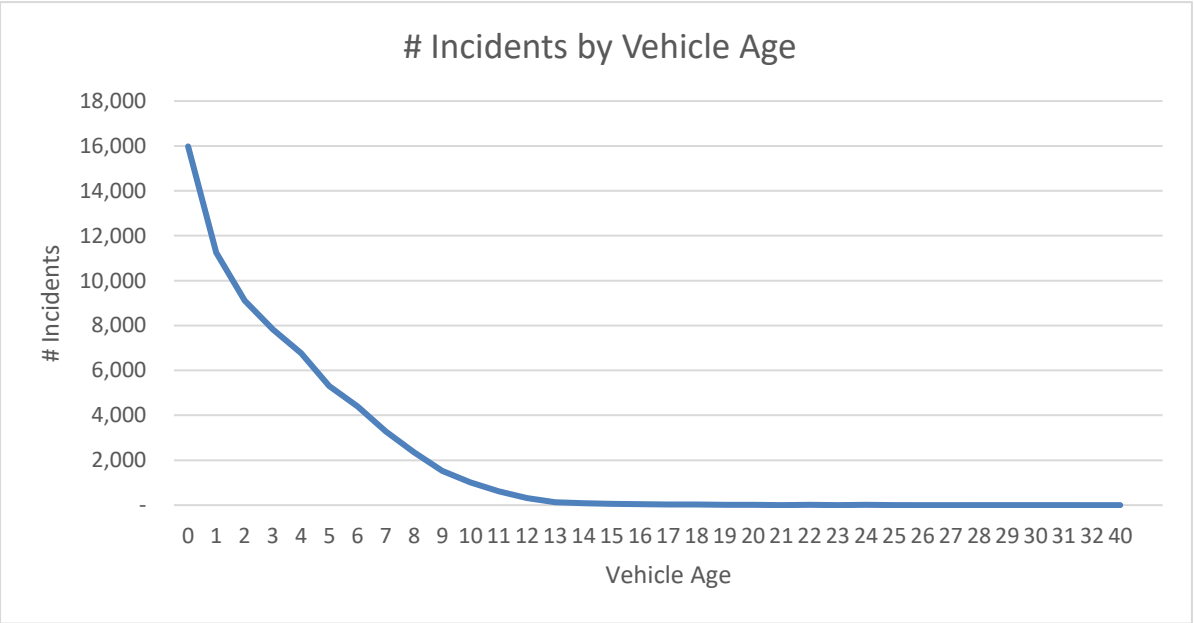
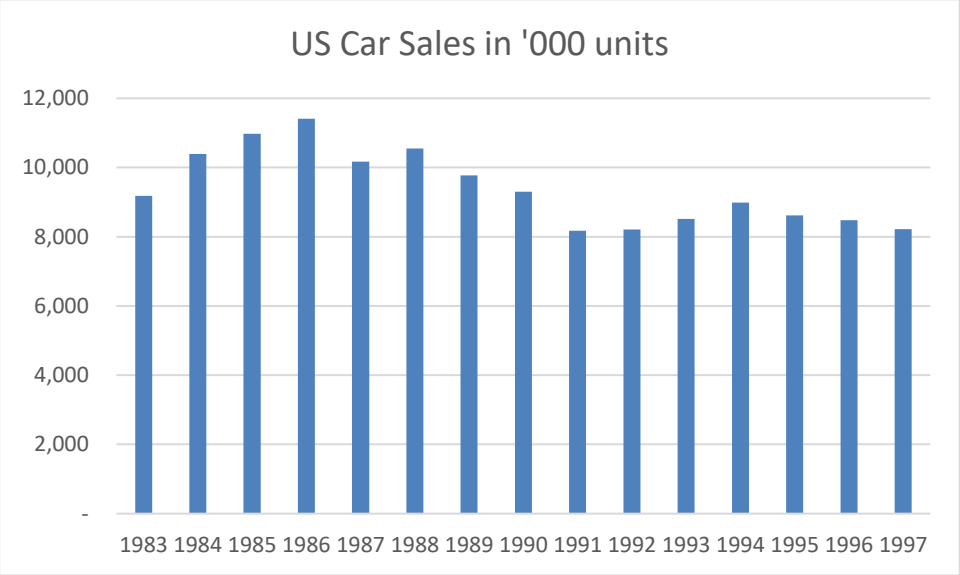


Fig- 1.4 Graph depicts the trend between age of vehicle at the time of incident and # of incidents

Figure1.4 represents a decreasing number of incidents with vehicle age that might imply faulty vehicles/parts within the vehicles or reckless driving rather than parts being replaced by being worn off.

ODI's most important field data:

- Prompts most new defect investigations
- Supports existing investigations
- Assess safety recall effectiveness
- NHTSA also uses complaints to target compliance testing
- Support research and rulemaking activity



*Credits - https://www.statista.com/statistics/199974/us-car-sales-since-1951/

Fig- 1.5 US Car sales from 1983-1997

Chapter Two

2. Methodology

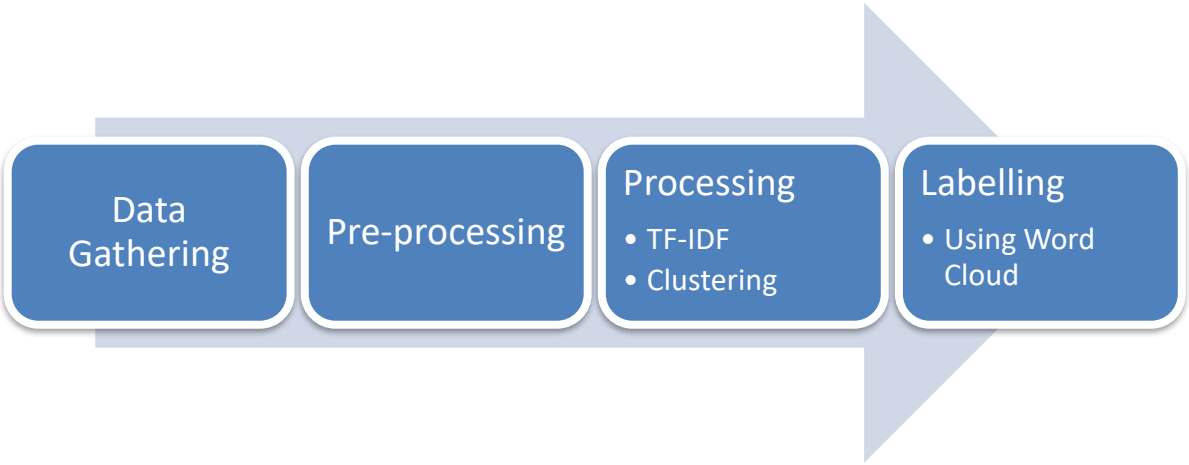


Fig- 2.1 Describes the methodology followed

The introduction section describes the data and its fields. Hence, the chapters proceeds with the Pre-processing steps in this section.

2.1 Pre-Processing

The below table provides a brief introduction to the standard steps followed for pre-processing textual data.

S.No	Step	Input	Output
1.	Converting all letters to lower case	The 5 biggest countries by population in 2017 are China, India, United States, Indonesia, and Brazil.	the 5 biggest countries by population in 2017 are china, india, united states, indonesia, and brazil.
2.	Tokenization	NLTK is a leading	[‘NLTK’, ’is’ , ’a’ ‘leading’,

		platform for building Python programs to work with human language data.	‘platform’, ‘for , ‘building’, ‘Python’, ‘programs’, ‘to’ , ‘work’, ’with’ , ‘human’, ‘language’, ‘data’, ‘.’]
3.	Remove Numbers	Box A contains 3 red and 5 white balls, while Box B contains 4 red and 2 blue balls.	Box A contains red and white balls, while Box B contains red and blue balls.
4.	Remove Punctuation	This &is [an] example? {of} string. with.? punctuation!!!!	This is an example of string with punctuation
5.	Removing White spaces	“ \t a string example\t “	‘a string example’
6.	Removing Stop Words	NLTK is a leading platform for building Python programs to work with human language data	[‘NLTK’, ‘leading’, ‘platform’, ‘building’, ‘Python’, ‘programs’, ‘work’, ‘human’, ‘language’, ‘data’, ‘.’]
7.	Stemming	For example, “fishing,” “fished,” “fisher”	all reduce to the stem “fish.”.
8.	Part of Speech/Entity tagging	Parts of speech examples: an article, to write, interesting, easily, and, of	[(‘Parts’, u’NNS’), (‘of’, u’IN’), (‘speech’, u’NN’), (‘examples’, u’NNS’), (‘an’, u’DT’), (‘article’, u’NN’), (‘to’, u’TO’), (‘write’, u’VB’), (‘interesting’, u’VBG’), (‘easily’, u’RB’), (‘and’, u’CC’), (‘of’, u’IN’)]

Table- 2.1 Describes the standard pre-processing steps

Stemming is a process of reducing words to their word stem, base or root form (for example, books — book, looked — look).
Part-of-speech tagging aims to assign parts of speech to each word of a given text (such as nouns, verbs, adjectives, and others) based on its definition and its context

Explaining the 4 simple yet effective Preprocessing steps this approach takes:

Tokenizing:

Tokenizing is the first and foremost step. Tokenizing refers to dividing a piece of text into tokens, words in our case, based on specific parameters. When the documents are input, they must be divided into sentences and these must further be divided into words using a Tokenizer. This is quite easy, and it returns a list of words which are in the documents. This list allows us to implement the process in an efficient manner. All the further operations are preformed on this set of tokens only.

Punctuation removal:

Punctuations like " !, ?, .." will only be a burden to deal with. Whatever analysis is done, it is on the words. It doesn't have to care about the punctuation. These punctuations only delay the process but not help it in any way unless analysis involves getting the emotion of the text. Which it is not, in our case. The analysis only looks for a label which best describes a particular set of words. Hence, there is no reason to keep these under consideration any further. So, this step strips of any punctuations to proceed further.

Stop words removal:

Stop words removal is the most important step. Stop words are those which have to be filtered off before proceeding with natural language processing. These are the words which are most common in the language. A stop word will only be slowing down the process as they don't really carry a lot of information. A stop word can be "the" or "is" or "at" or "on" etc. As tokenization consider each and every word or token, having these in the list will not be of use and moreover will serve as a disadvantage in the further process. Hence removing these will help a lot. So, this preprocessing performs the mentioned operations for making the process more efficient.

Stemming and Lemmatization:

Stemming basically refers to stripping of the end or beginning of the word by taking some common prefixes and suffixes into consideration.

2.2 Processing

Processing in this projects sense involves clustering the textual data into different clusters.

The clustering of the complaints happens in the following 2 steps broadly,

1) Finding TF-IDF vector
2) Clustering

After these two steps have been applied, a desired number of clusters of words have been achieved based on their significance calculated using TF-IDF. The next section describes the TF-IDF vector and Clustering.

TF-IDF

1. Quantifies documents by giving weights to the terms (or tokens) inside the documents by using a technique called TF-IDF (Term Frequency — Inverse Document Frequency)
2. The calculated TF-IDF indicates the importance of each term (or token) to a given document
3. How many times a given term appears in the document it belongs to is the TF (Term Frequency) part of TF-IDF. The higher the TF value is, the more important the term is for the document.
4. However, if the given term appears in all the documents then it is not that important in order to identify the document. For example, if every single document contains 'supply chain' then this term would not be helpful to identify any given document.
5. So, inorder to have a weighting system that would decrease the importance of a

given term as the number of the documents it shows up increases. And this is the 'IDF (Inverse Document Frequency)' part of the 'TF-IDF'.

6. TF-IDF is a weighting mechanism that calculates the importance of each term for each document by increasing the importance based on the term frequency while decreasing the importance based on the document frequency.

For more clarity, there is a calculation explained for one word within multiple documents. For our project, each complaint would refer to one document.

- Consider a document containing 100 words wherein the word cat appears 3 times.
- The term frequency (i.e., TF) for cat is then (3 / 100) = 0.03.
- Now, assume there are 10 million documents and the word cat appears in one thousand of these.
- Then, the inverse document frequency (i.e., IDF) is calculated as log (10,000,000 / 1,000) = 4.
- Thus, the Tf-idf weight is the product of these quantities: 0.03 * 4 = 0.12.

Considering there are three complaints/sentences as mentioned below. The objective is to create a sample TF-IDF matrix for each word occurring within these complaints/sentences.

1. Document 1 - I ate an apple yesterday, and I ate another apple today.
2. Doc 2 - I will eat an apple today.
3. Doc 3 - I ate a banana yesterday.

Features

Complaints

	a	an	apple	ate	banana	eat	i	today	will	yesterday
Doc 1		0.0811	0.0811	0.0811			0			0.0811
Doc 2		0.0676	0.0676			0.1831	0	0.1831	0.1831	
Doc 3	0.2197			0.0811	0.2197		0			0.0811

Fig- 2.2 Sample TF-IDF matrix for the above-mentioned sentences

Clustering:

Clustering can be considered the most important unsupervised *learning* problem; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data.

A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way".

A *cluster* is therefore a collection of objects which are coherent internally, but clearly dissimilar to the objects belonging to other clusters

Classification

Clustering algorithms may be classified as listed below:

- **Flat clustering**: Creates a set of clusters without any explicit structure that would relate clusters to each other; It's also called exclusive clustering Hierarchical clustering: Creates a hierarchy of clusters.
- **Hard clustering**: Assigns each document/object as a member of exactly one cluster.
- **Soft clustering**: Distribute the document/object over all users

Algorithms

- Agglomerative (Hierarchical clustering)
- K-Means (Flat clustering, Hard clustering)
- EM Algorithm (Flat clustering, Soft clustering)

Hierarchical Agglomerative Clustering (HAC) and K-Means algorithm have been applied to text clustering in a straightforward way. Typically, it uses normalized, TF-IDF-weighted vectors and cosine similarity. Here, the k-means algorithm using a set of points in n-dimensional vector space for text clustering

K-Means Clustering - Introduction:

- This clustering algorithm was developed by MacQueen and is one of the simplest and the best-known unsupervised learning algorithms that solve the well-known clustering problem.
- K-Means clustering is used to cluster n observations into k groups where $k \leq n$.
- In other words, K-Means clustering is used to make "K" clusters out of the given observations. Each cluster will be formed in such a way that all the observations in it are similar to each other in the very same cluster than other clusters.
- Let's take an example, given some random names and addresses, there can be a cluster of names and another cluster of addresses. Hence, it formed a cluster of only similar things. A name is conceptually more similar to another name than it is to an address.
- This can be called a lazy method as it performs most of the computation when it has to perform the classification of the new observation. This is an instance-based clustering technique. Its training experience is all the observations it previously encountered. Its performance increases with increasing number of observations fed to it.
- K-Means creates "K" clusters, given K an integer. If K=3, it creates 3 clusters. That is, it classifies each of the observation into either one of the three clusters.
- The k-means clustering algorithm is known to be efficient in clustering large data sets.
- The centroid of a cluster is formed in such a way that it is closely related (in terms of similarity function; similarity can be measured by using different methods such as cosine similarity, Euclidean distance, Extended Jaccard) to all objects in that cluster.

Working Algorithm for K-Means:

- Choose k number of clusters to be Determined.
- Choose k objects randomly as the initial cluster center

- Initially, k is chosen, and k base points are taken randomly. These are called centroids.
- Each centroid is a data point, or an observation. Whole k-means works on Euclidean distance property i.e. the distance between two observations in the Euclidean space.
- Euclidean space is where the points are plotted.
- Whenever a new observation is encountered, its distance to the k number of clusters is calculated. It is classified to be a part of the cluster whose centroid is the closest to the new point.
- After classification of every new observation, the centroid is recomputed. Repeat
    - Assign each object to their closest cluster
    - Compute new clusters, i.e. Calculate mean points.
- Until No changes on cluster centers (i.e. Centroids do not change location any more) OR No object changes its cluster

Residual Sum of Squares (RSS):

RSS is the objective function in K-means and our goal is to minimize it. Because N is fixed, minimizing RSS is equivalent to minimizing the average squared distance; a measure of how well centroids represent their documents. RSS is a measure of how well the centroids represent the members of their clusters, the squared distance of each vector from its centroid summed over all vectors.

$$RSS_k = \sum_{x \varepsilon w_k} | x - \mu(w_k)|^2$$

$$RSS = \sum_{k=1}^{K} RSS_k$$

Where

- w_k = Document Cluster k

- $\mu = Mean\ or\ centroid\ of\ the\ document\ cluster\ w_k$
- $x = Document\ vector\ in\ cluster\ k$

The algorithm then moves the cluster centers around in space in order to minimize RSS.

Let's take an example where k=2, the centroids are represented by "x" mark in the following figure.

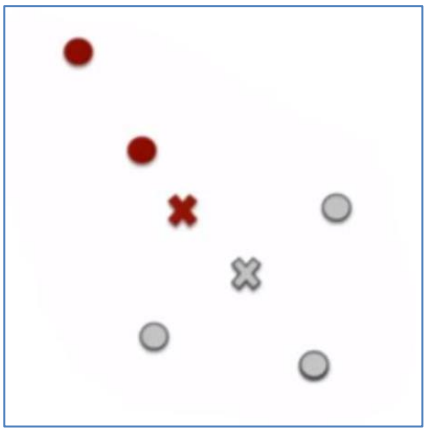


Fig- 2.3 Pre-Clustering Data points

After re-computing the centroids,

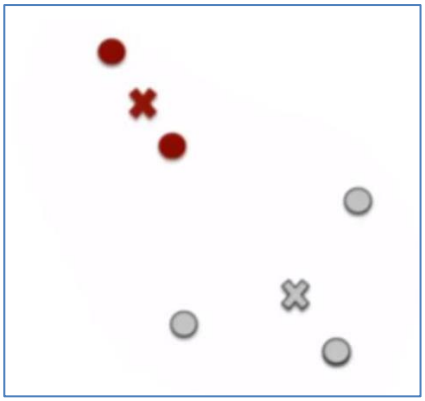


Fig- 2.4 Post-Clustering Data points

- This clustering is a major step as this is how one clusters the textual data into clusters which will be labeled.
- Like most of the algorithms, even K-Means works with only numerical data.
- So, to convert the textual data into numerical data, one calculates the TF-IDF of every word of the data that remains after preprocessing.
- The vector of TF-IDF values is generated, and fed to the K-Means algorithm to obtain "K" clusters.
- The document is represented in the form of vector such that the words (also called features) represent dimensions of the vector and frequency of the word in document is the magnitude of the vector i.e.
- Vector is of the form
    - <(t1,f1),(t2,f2),(t3,f3),….(tn,fn)>
    - Where t1, t2,.., tn are the terms/words (dimension of the vector) and f1, f2,…, fn are the corresponding frequencies or magnitude of the vector components.
    - Figure 2.2 depicts how a typical Vector for the words/terms and its corresponding frequencies would look like.
- This whole operation can be done using python its "sklearn" module.
- Now after generating the clusters, the next step involves to label them.

Up till now, the project discusses the steps to be followed for pre-processing and processing textual data to be finally clustered.

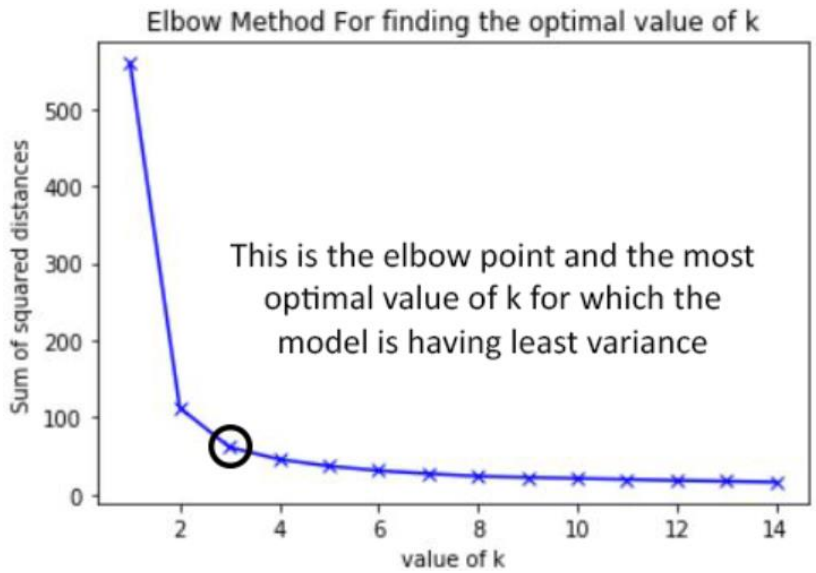
But, the bigger Question – How do you find the value for K??

- Since, K-means clustering is a type of unsupervised learning.
- It is used when you have unlabeled data (i.e., data without defined categories or groups).
- The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K

- Since it is an unsupervised learning algorithm, one can't give any value to k.
- The algorithm has to decide the number of clusters by itself. For deciding that, it follows a hit and trial approach.

Stopping Criterion:

- Taking the value of k starting from 0 and going up to the number of points and checking the variance (Euclidean Distance between that point to the centroid of the respective cluster which is plotted as function of K) after every iteration.
- "Elbow point "- where the rate of decrease sharply shifts.



*Credits - https://blog.cambridgespark.com/how-to-determine-the-optimal-number-of-clusters-for-k-means-clustering-14f27070048f

Fig- 2.5 Indicated the ideal expected graph for Elbow point Analysis

Labeling:

- This is the last phase of our procedure. This step involves labeling the obtained clusters.
- A good label is the one which best describes the data in it. So far our approach, tokenized the data, removed the stopwords, calculated TF-IDF and then finally clustered them.
- After obtaining the different clusters containing words, one must choose a best descriptor of these words.
- The above task is done so by creating Word Cloud for each of our clusters.

Word Cloud:

A text cloud or word cloud is a visualization of word frequency in a given text as a weighted list

Chapter Three

3. Modelling

3.1 Model Pre-Processing

- Chapter 1 gave a brief description about data; Chapter 2 spoke about the methodology generally followed for pre-processing textual data and then processing steps. Chapter 3 spoke about the model pre-processing steps customized for our data.

- Few steps like a spell checker and keeping only Verbs, Adverbs and Nouns have been added to the pre-processing phase as per our observation. T

- The pre-processing steps act like a feature generation/selection process in the case of textual data and aim to remove all the extra words that would be generated in the feature set without adding any extra predictive power to our set.

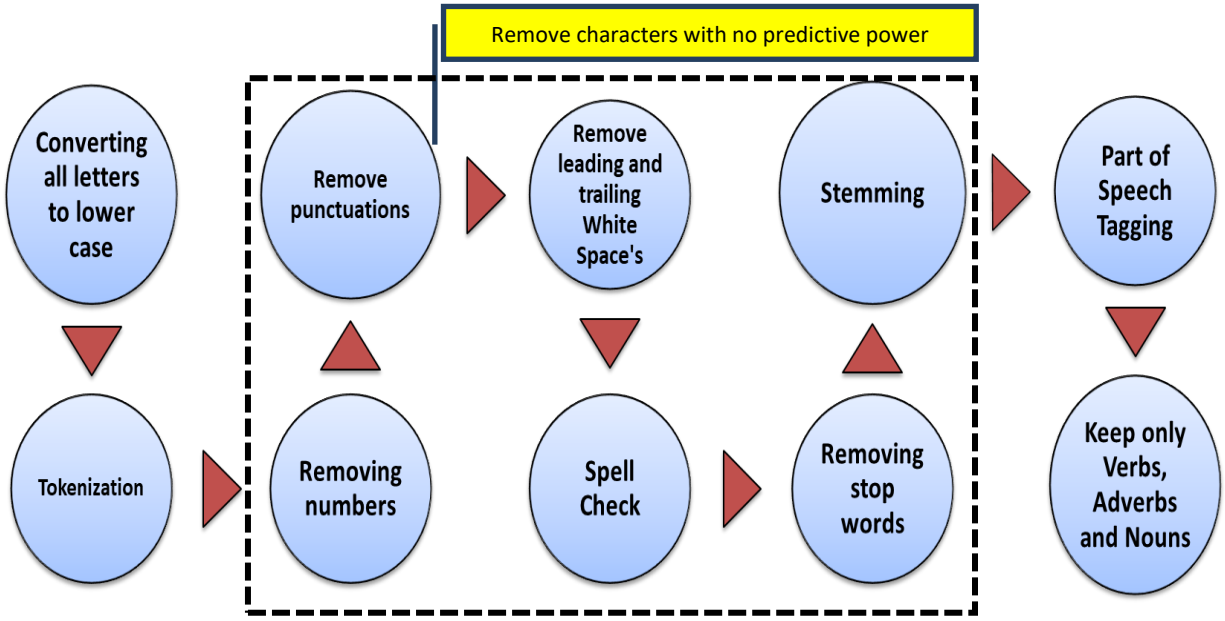


Fig- 3.1 Customized Pre-processing steps

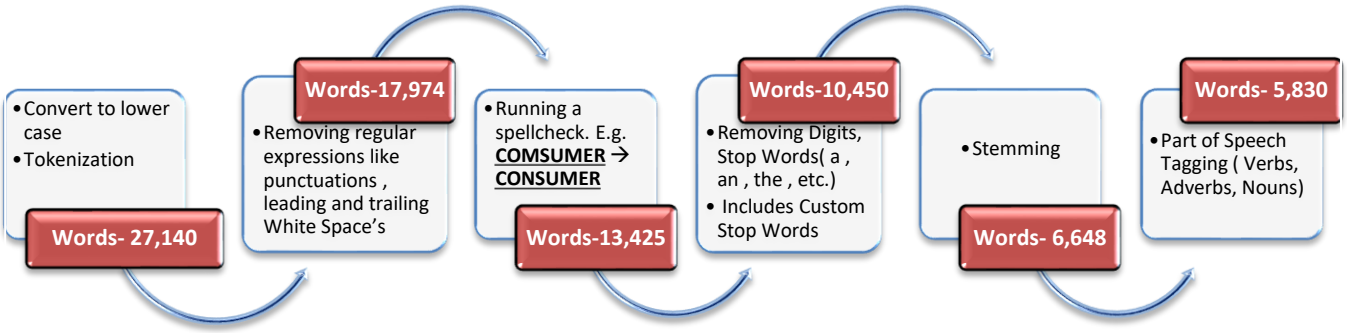


Fig- 3.2 Step by Step unique words

The figure 3.2 explains that the project starts with 27,140 unique words in the corpus of our complaints at the beginning, and at each step it goes on reducing these unique words to finally arrive at 5,830 unique words which would form our feature set of words from the corpus of complaints.

3.2 Model Processing

- TF-IDF vector for our analysis has a shape – (89,592 * 1,350). There are 1,350 features.
- To restrict the features to this number, a minimum limit for the feature of TF-IDF value as '0.001' is applied.

- The features consist of Unigrams, bigrams and trigrams.
    - Bi grams (co-occurring 2 words) include ‘Seat Belt’.
    - Tri-grams (co-occurring 3 words) include examples like “Anti Breaking System)

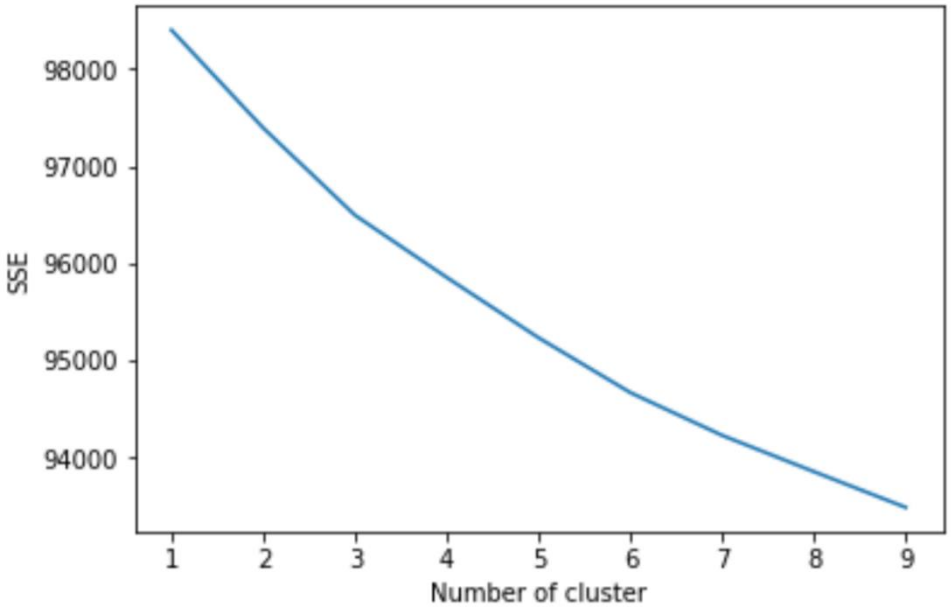


Fig- 3.3 Elbow Analysis graph for Clusters

Figure 3.3 depicts the SSE (Sum of Squared Errors within each cluster for K ranging from 1 – 10. As mentioned in Chapter 2, figure 2.4 our graph does not replicate the ideal graph. Thus, the analysis intends to find other methods to identify the number of clusters, also represented as ‘K’ in our project.

Stopping Criterion

- Since, the graph could not assist in identifying an Elbow Point:
- For K ranging from 1-10, Word Cloud were created for each cluster.
- Here, the clustering was stopped one step before a split occurred that resulted in 2 clusters with the same most frequent term.
- After multiple iterations of clusters for K values ranging from 1 -10, Cluster analysis identified **K = 8** as **our stopping criterion**
- For **K = 9**, there were **2 clusters** having **Brake** as the **most frequent word**.

Chapter Four

4. 1 Results and Insights

Finally, the analysis ends up having **8 clusters**, below is an image depicting the clusters.

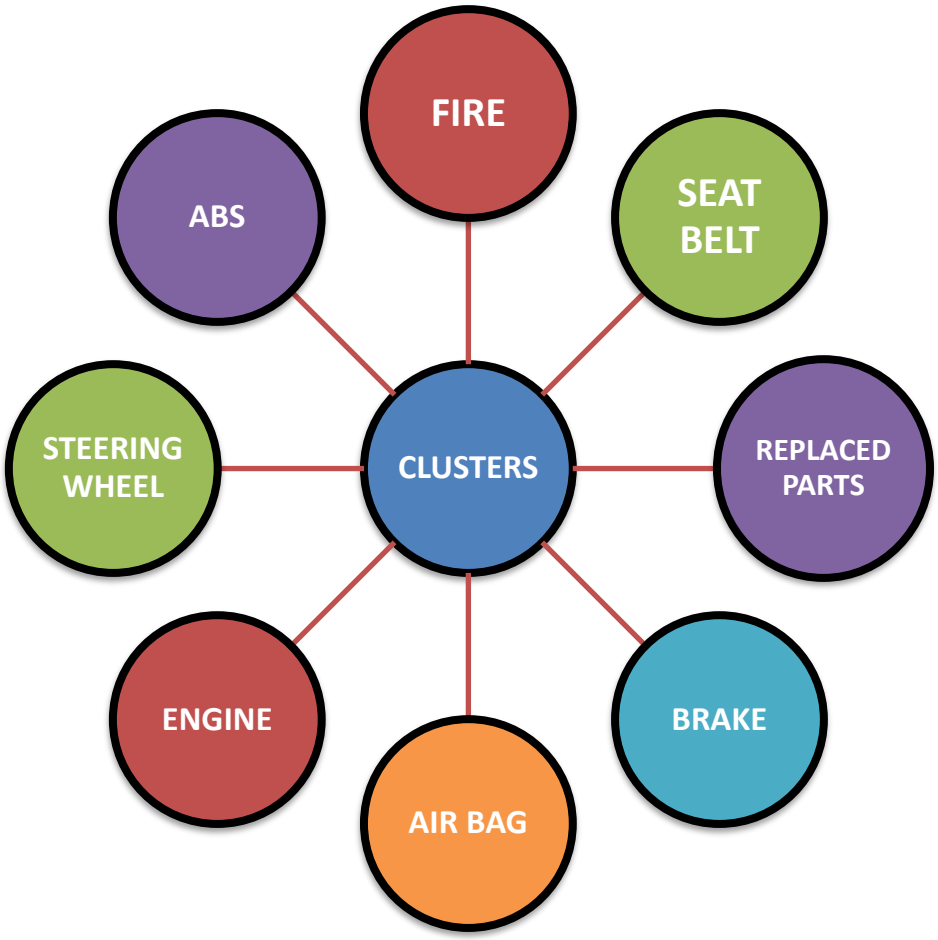


Fig- 4.1 Cluster Labels

The figure 4.1 depicts the labels for the 8 clusters that the analysis resulted in after the clustering. 6 of the 8 clusters reflect towards parts from the vehicles, whereas one of the clusters labeled as “FIRE” would reflect to incidents leading to fire in the vehicle. This would require more in-depth analysis of identifying the causes behind “FIRE”. One of the clusters are labeled as “Replaced Parts” which could refer to either a requirement of replacing the faulty part or else it can infer that even though parts were replaced, they lead to an incident. Both the cases would require further analysis.

Fig- 4.2 Word Cloud for the 8 Clusters

Figure 4.2 portrays the Word Clouds for the 8 clusters, the Word Cloud depicts the most frequent word which in majority of our cases ends up being our label. Other words within the Word Cloud also can help us infer the problems within the clusters.

- For example, 'air', 'bag' and 'deploy' as the most prominent words for "AIR BAG' cluster which might imply a problem with air bag deployment.

Manufacture Level:

1. Ford Motor, General Motors and Diamler Chrysler Corporation account for 78% of the complaints. (They accounted for 69% of the auto market share in 1990).
2. **SEAT BELT** emerged as one of the clusters. (The Takata Seatbelt Scandal, 1995 - 8.3 million vehicles recalled) [10]
3. However, **Honda** and **Toyota** account for **7%** and **5%** of the proportions within the **AIR BAG** cluster. (Honda accounts for **2.6%** and Toyota accounts for **2.2%** of total complaints).
4. **Nissan** accounts for **9%** of the proportion within the **FIRE** cluster. (Nissan accounts for 2.7% of total complaints). [9] (In 1994, Nissan to Buy Back 33,000 Defective Minivans)
5. Ford Motor accounts for **50%** of the proportion within the **FIRE** cluster. [7][8]
6. General Motors accounts for **48%** of the proportion within the **BRAKE** cluster. [4][5]

Model Level:

1. Dodge Caravan (Chrysler), Ford Tauras, Chrysler Voyager and Ford Explorer are the models with proportion of complaints > 2.5% within 7, 6,5 and 5 clusters.

4.2 Time Trend Analysis of Clusters

- One benefit of this analysis approach is that cluster labels are assigned to each complaint, facilitating the analysis of clusters with respect to other fields of data associated with the original complaint.
- Incident date and report date are particularly interesting in this case because they can describe time trends in incident frequency and indicate emerging issues that might require intervention.
- Figure 4.3 shows the number of incidents/complaints reported each year (from 1983 to 1997). The largest disturbance occurs in the 'BRAKE' Cluster around 1995 starting from 1991 which corresponds to the General Motors recall. The trend clearly shows indications of an increase in brake related complaints as early as 1991. An analysis of the complaints, similar to what has been done here, may have led to detection as early as 1991 or 1992.
- Figure 4.3 also shows the number of incidents/complaints in FIRE cluster of incidents reported each year (from 1983 to 1997). The largest disturbance occurs in the Tire cluster around 1996, which corresponds to the Ford faulty ignition switches causing fire in the vehicle issue. NHTSA launched a formal investigation into the incidents in 1996 (NHTSA, 1996); however, the time trend clearly shows indications of an increase in fire-related complaints as early as 1992. An analysis of the complaints, similar to what has been done here, may have led to detection as early as 1992 or 1993.

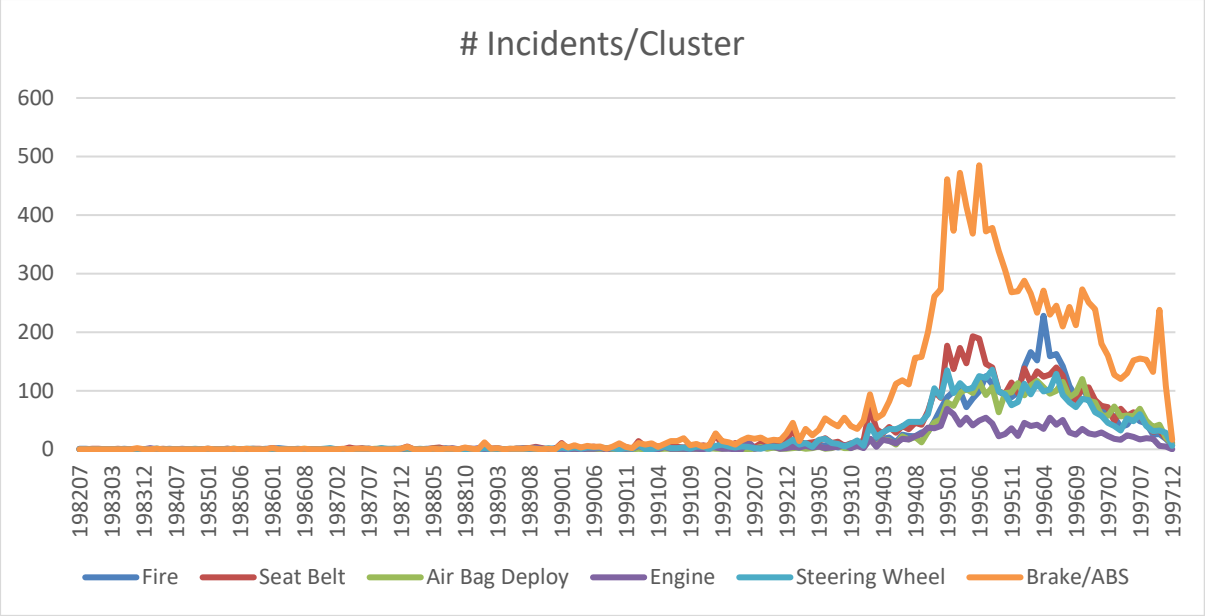


Fig- 4.3 # of incidents per Cluster

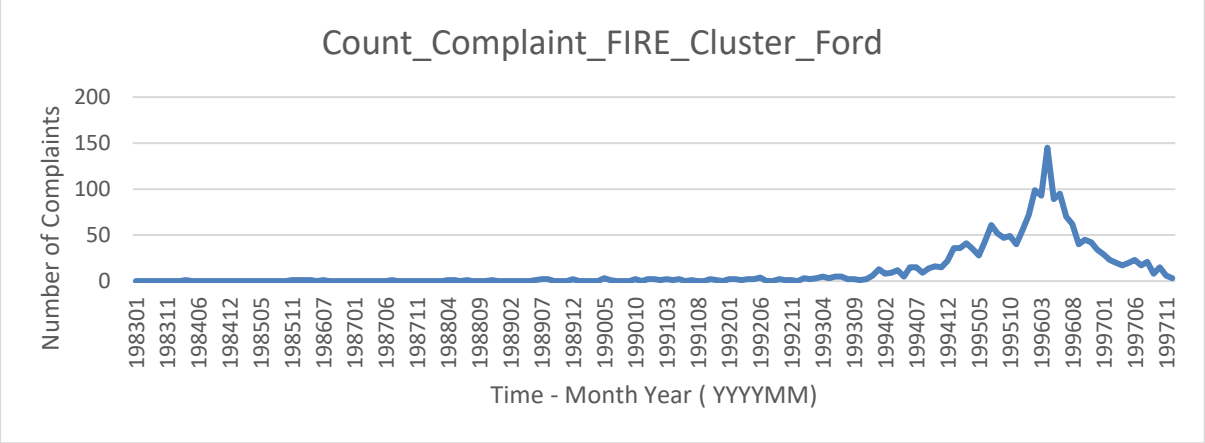


Fig- 4.4 # of incidents for the FIRE cluster specific to FORD

In 1996, Ford Motor Co. announced the largest safety recall by a single car manufacturer yesterday, saying it would replace the ignition switch in 8.7 million Ford and Mercury vehicles made between 1988 and 1995. [7][8]

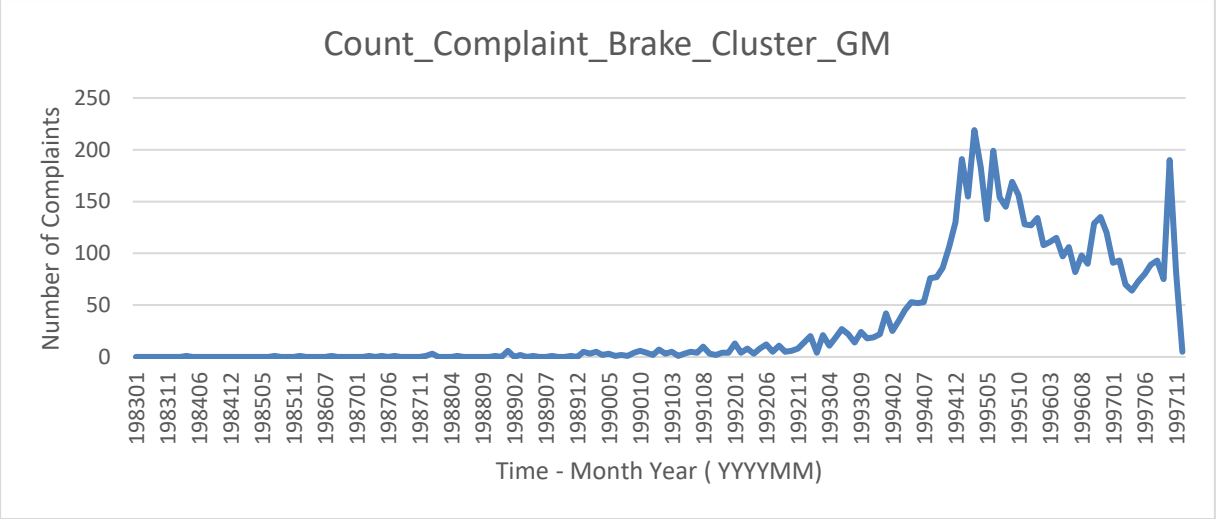


Fig- 4.5 # of Incidents for BRAKE Cluster – Specific to GENERAL MOTORS

G.M. recalled 1.1 million four-wheel-drive vehicles sold in the 1991 through 1997 years. [4][5]

4.3 Conclusion

Manual summarization of free-response databases is difficult and time-consuming, thus Text mining, specifically cluster analysis, can assist us in revealing useful information with ease and less time. With the growing number of unanticipated incidents like recalls and faulty manufactured parts as the industry adopts and tests new technologies, the value of such automated techniques is set to increase. Computerization of vehicles introduces failure modes that are difficult to detect during design, particularly for those systems that change the role of the driver (Merat & Lee, 2012). Traditional methods for testing in the automotive industry involve simulator-based testing and crash statistics, though the incident reports have important limitations as would be discussed in the upcoming section, they still offer an important complement to the traditional methods. The approach followed and outlined above could provide useful in identifying for any emerging problems in the vehicle owner's complaint database. Text analysis approach is not limited to this and can be applied to more generally a variety of consumer complaint databases or even Twitter feeds to capture the temporal and spatial trends in opinions, habits, or events (Golder & Macy, 2011). Assigning predefined categories to incidents will import a bias in the analysis which can let emerging issues go unnoticed. Based on the exploratory analysis outlined here in the project, it can indicate emerging themes months or years before they are revealed in crash or naturalistic driving data. This method could be thought of as to systematize (in several steps including preprocessing, and clustering) and run periodically to automate the detection of emerging problems and identifying and addressing trends before more lives are endangered.

Even though TF-IDF and K-Means clustering do not consider word dependencies, they still act as powerful tool in identifying themes in the narrative data and guiding analysts to more in-depth investigations of the identified documents. The efficient handling of complex narrative data and extracting cluster of complaints on the free-response incident descriptions contained in this field by using an analytic framework, which in other terms would have been very difficult for a human analyst to manually cluster and track them over time.

Limitations

The broader picture with this study is to utilize text mining and it's approaches for analyzing textual data, which in our case is contained within free response consumer review or complaint databases. The following are some of the limitations that were not considered in the scope of this project

- Since, technology is evolving at an exponential rate, it was difficult to accommodate for the technological changes which would have changed the type of complaints and their frequencies over the decade long time period. Thus, this idea behind this was to conduct the analysis in an entirely exploratory fashion with no a priori hypothesis made regarding the outcomes.

- Human intervention is still an integral part of any form of automation. To avoid misinterpretation of results/output and over reliance, it requires careful attention to text mining and its limits.

- Using TF-IDF and K-Means for clustering, one needs to be careful while ascribing causal relationships between component failures and incident outcomes, because they ignore word dependencies, sequences and nuances of word meaning considering the complaints as bag of words. These techniques can't differentiate between scenarios where a human analyst can easily discern easily like where faulty components are responsible for the incident from cases where an incident triggers a component failure.
    - The occurrence of an abrupt braking leading to a crash where the airbag fails to deploy is an example of the latter one, whereas the occurrence of a brake failure causing the vehicle to run into leading vehicle is an example of the former.

- Raw incident frequencies have been used while calculating and presenting the time trends which in other words implies not considering any changes in the count of registered vehicles or licensed drivers. One would expect some trends to be masked

and some others inflated due to the lack of this adjustment. Since the trends within the clusters are considered relative to each other, the patterns that stand out are those that indeed merit attention. Also, when the focus is on spikes from a year to the next, the effect of denominator changes would be minimal. Follow-up investigations on the identified problems can incorporate adjustments

- One of the major limitations involving the data considered for analysis is the database not being an exhaustive or unbiased record of incidents since it was compiled based on a voluntary reporting system. One can also expect the complaints to be influenced by media reports or news coverage regarding recalls of vehicles by the manufacturers. In many cases, the report could be filed by a person absent from the scene (like lawyer). Also, depending upon the technical expertise of the individual reporting complaints, both above factors could make it more difficult in making inferences about the reports.

- Many of the decisions throughout the analysis were based on heuristics. For example, the choice of stop words was partially driven by finding the most frequent noninformative terms (e.g., “vehicle”). The choice of stop words can influence the outcomes, both in the preprocessing phase and in clustering. For example, the term “passeng” (the stemmed form of “passenger”) occurs in several of the clusters. One could argue that “passenger” is a stop word in the context of vehicle complaints and should be removed, as would be the case with term frequency inverse document frequency weighting, which removes very frequent and very infrequent words (Oza, Castle, & Stutz, 2009). Such a decision might change the composition of some of the clusters. Parameters such as the stopping criterion in clustering and the sparsity cutoff introduce additional degrees of freedom. As such, even though the modeling framework used here is quantitative, major qualitative considerations enter the analysis.

- This process of clustering and labeling is aimed at dividing textual data into chunks of labelled blocks of data. This is very useful in organizing data and further helps in understanding data quickly and efficiently. When properly done, a quick glance will

be enough to understand the brief point of the data under consideration. However, this process is highly dependent on the performance of the clustering mechanism used. If the cluster is comprising of vaguely related words, the label assigned will invariably be vague as well. For the future, one can aim at developing or tweaking the present clustering models which will help in forming better clusters, which will lead to a concise label.

Future work

Injury and Death Involvement	Classification(Severity Leve)
At least one death reported	Fatal Incidents
No deaths and at least one injury reported	Injury Incidents
No injuries or deaths reported	Minor Incidents
No injuries reported, missing fatality data	
At least one injury reported, missing fatality data	
No deaths reported, missing injury data	
Missing both injury and fatality data	
Overall	

Fig- Future Work

- One can extract information about severe incidents (involving deaths and injuries) based on the above severity classification. Clusters of complaints can be identified and compared across subsets of data pertaining to fatal incidents and incidents involving injury
- Further extension of this work could be to utilize text mining and hierarchical clustering in analyzing a free-response database and comparing the results with other clustering techniques. One step further could involve, moving from bag of words techniques towards contextual analysis that could be further used for ascribing causal relationships between component failures and incident outcomes

References

A. Deniz Iren, Hajo A. Reijers. "Leveraging business process improvement with natural language processing and organizational semantic knowledge", Proceedings of the 2017 International Conference on Software and System Process - ICSSP 2017

B. Automated Text Clustering and Labeling using Hypernyms, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 2 (2019) pp. 447-451

C. Ghazizadeh, M., & Lee, J. D. (2012, October). *Consumer complaints and traffic fatalities: Insights from the NHTSA vehicle owner's complaint database*. Paper presented at the Human Factors and Ergonomics Society 56th annual meeting, Boston, MA.

D. National Highway Traffic Safety Administration, Office of Defects Investigation. (2011). *Vehicle owner's complaint database*. Retrieved from http://www-odi.nhtsa.dot.gov/downloads/

E. National Highway Traffic Safety Administration, Office of Defects Investigation. (n.d.). *Internet Vehicle Owner's Questionnaire*. Retrieved from https://www-odi.nhtsa.dot.gov/ivoq/

Online References

1. https://www.nltk.org/book/
2. https://scikit-learn.org/stable/
3. http://www.tfidf.com/
4. https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
5. https://www.nytimes.com/1999/07/22/us/gm-admits-brake-flaws-after-inquiry.html
6. https://www.baltimoresun.com/news/bs-xpm-1999-07-22-9907220192-story.html
7. https://www.nytimes.com/1995/05/22/us/us-said-to-want-huge-recall-of-cars.html
8. https://www.latimes.com/archives/la-xpm-1996-04-26-mn-63066-story.html
9. https://www.washingtonpost.com/archive/politics/1996/04/26/ford-issues-safety-recall-of-87-million-vehicles/9ec564b3-dccb-42e3-bf78-

bcbf09d01147/?utm_term=.835b939b128a

10. https://www.latimes.com/archives/la-xpm-1994-02-04-fi-19061-story.html
11. https://www.investopedia.com/slide-show/car-recalls/
12. https://www.nytimes.com/2016/08/27/business/takata-airbag-recall-crisis.html