

Text Analysis of American Airlines Customer Reviews

Rajesh Tolety, Oklahoma State University
Saurabh Kumar Choudhary, Oklahoma State University

ABSTRACT

Which airline should I chose to make my journey comfortable? This is the question which comes to everyone's mind every time one plans a trip because it's not only about reaching the destination but also about the travel experience on board. This is not only important for the passengers but also for the airline companies as they also want their customers to be satisfied and happy so that customers prefer them every time they fly.

The objective of this paper is to analyze the customer's reviews of American Airlines to categorize their experiences with respect to the reviews. The nature and the tone of the reviews are important metrics for American airlines to track and manage their performance and services. SAS® Enterprise Miner is used to understand the association between the customers' expectations and their experiences.

Our preliminary analysis using the text parsing and text filter nodes helped us to get a quick understanding of all the terms present in the reviews and the nature of relationship between them. The text cluster, text topic and text profile nodes are then used to group the terms from a similar context. We also explain what issues customers face while onboard. Results from our study will be helpful for American airlines to measure and track such issues in future.

INTRODUCTION

According to a survey report of Trip Advisor, about 43% of the airline passenger rely on online reviews of different airlines before booking a ticket. Text analysis features provided by SAS® Enterprise Miner™ are used to analyze and help interpret textual data about American Airline Customer Reviews. The text mining process followed in this paper is the one discussed by Chakraborty, Pagolu and Garla (2014)¹. The scope of this paper is limited to the textual analysis of data, validating the reported information from different websites. Text profiler and text builder node has helped in clustering the information gathered and indicating the presence or absence of a word or group of words. These rules are used to predict a target variable, i.e. whether a feedback is positive or negative.

DATASET

We managed to get the dataset from three different websites i.e. Trip Advisor, Consumer Affair and Airline quality using import.io. We tried extracting the data from twitter but the limitation we encountered with twitter is that it only provides the data of past 7 days. Below is the table which shows the type of dataset and variables we had for analysis. Also, the image below shows the type of variables we had in the dataset.

SAS Format Variable	DATA TYPE	DATA SOURCE
Customer Reviews	CHARACTER	TRIPADVISOR, CONSUMER AFFAIR and AIRLINE QUALITY
Customer Ratings	NUMERIC	TRIPADVISOR, CONSUMER AFFAIR and AIRLINE QUALITY
Review Date	DATE	TRIPADVISOR, CONSUMER AFFAIR and AIRLINE QUALITY

Table 1: Variables

PROCESS FLOW

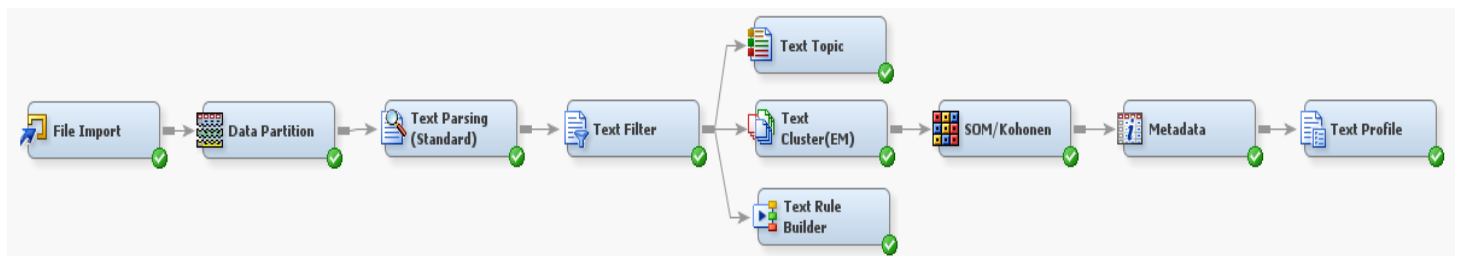


Fig 1: Process flow diagram

FILE IMPORT and DATA PARTITION:

File import node is used to import the data. Using the data partition node, we divide the whole data into two parts i.e., training (50%) and validation (50%). The analysis would be done on the training part and will be checked on the validation part to measure the accuracy. The validation statistics can then be used to assess the results from predictive models such as the text rule builder node.

TEXT PARSING:

The Text Parsing node parses a document collection in order to quantify information about the terms. The node is used to parse the text data using different parts of speech and noun groups. Few words are issued by reviewing the words and their importance. Following table shows the list of terms that were discarded/kept based on their importance as judged by the default parameters in SAS® Text Miner. Also, figure 3 shows the number of frequencies per documents.

Term	Role	Attribute	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
+ be	...	Alpha	20609	3721N	+		22536	1
+ flight	...	Alpha	11582	3217Y	+		5720	2
+ have	...	Alpha	8840	2890N	+		22514	3
not	...	Alpha	9480	2857N			22557	4
+ airline	...	Alpha	4697	2205Y	+		16553	5
+ do	...	Alpha	5337	2177N	+		22554	6
+ get	...	Alpha	4498	1834N	+		22555	7
+ fly	...	Alpha	2878	1779Y	+		10150	8
+ american	...	Alpha	3381	1746Y	+		876	9
no	...	Alpha	3315	1704N			22669	10
more	...	Alpha	1953	1589N			22574	11
+ hour	...	Alpha	2817	1458Y	+		11467	12
+ time	...	Alpha	2644	1443Y	+		1471	13

Fig 2: All terms with their frequency

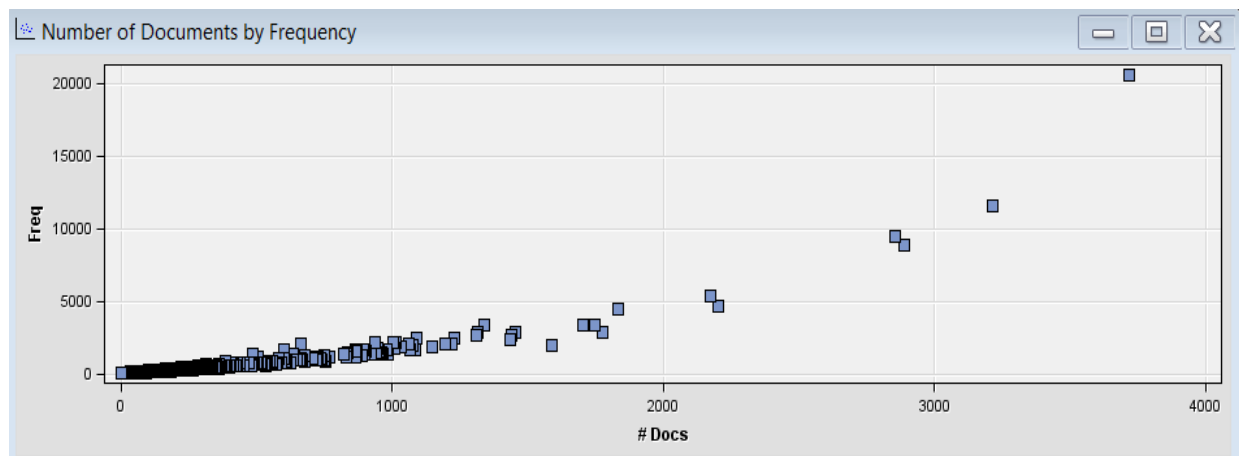


Fig 3: Number of Documents per Frequency

TEXT FILTERING:

Text Filter node is used to further reduce the total number of parsed terms that will be analyzed. The idea is to eliminate extraneous information so that only the most valuable and relevant information is considered. User defined synonym list is created using interactive filter to give a definitive name which can identify a set of words to generalize the terms. Using the spell check option we can correct the misspelt words as we can see in below table. The misspelt term 'passanger' is corrected to 'passenger', 'comunication' to 'communication' and so on. The import synonym option in the text filter node can be used to group terms together as synonyms either by adding a table or by manually selecting the

terms and marking them as synonyms. Table below shows an example of the exported synonym list that was created to use in this analysis.

Parent # Docs	Term	# Docs	Parent	Role
66.0	descent	2.0	decent	
289.0	passenger	1.0	passenger	PROP_MISC
289.0	pasenger	1.0	passenger	
289.0	passanger	2.0	passenger	
5.0	substantiallty	1.0	substantially	
12.0	30 day	1.0	30 days	TIME
70.0	comunication	1.0	communicat...	
5.0	tony	1.0	ton	

Fig 4: Table with spell check using SAS® default dictionary

	TERM	FREQ	# DOCS	KEEP ▼
☒	flight	18151	4300	☒
☐	plane3	1	1	
☐	flight	12704	3770	
☐	planet	7	7	
☐	flightâ	2	2	
☐	aircraft	1	1	
☐	plane	2	2	
☐	floght	1	1	
☐	aircraft	434	311	
☐	plane	2724	1352	

Fig 5: Synonym Grouping

CONCEPTS LINKS:

SAS® Enterprise Miner has a very useful feature of concept links which helps us to understand the association between various terms used in the dataset. Concept Link diagrams are visual representations of how terms are related to one another. When we generated concept link diagrams on the customer reviews, we find many interesting links as follows:

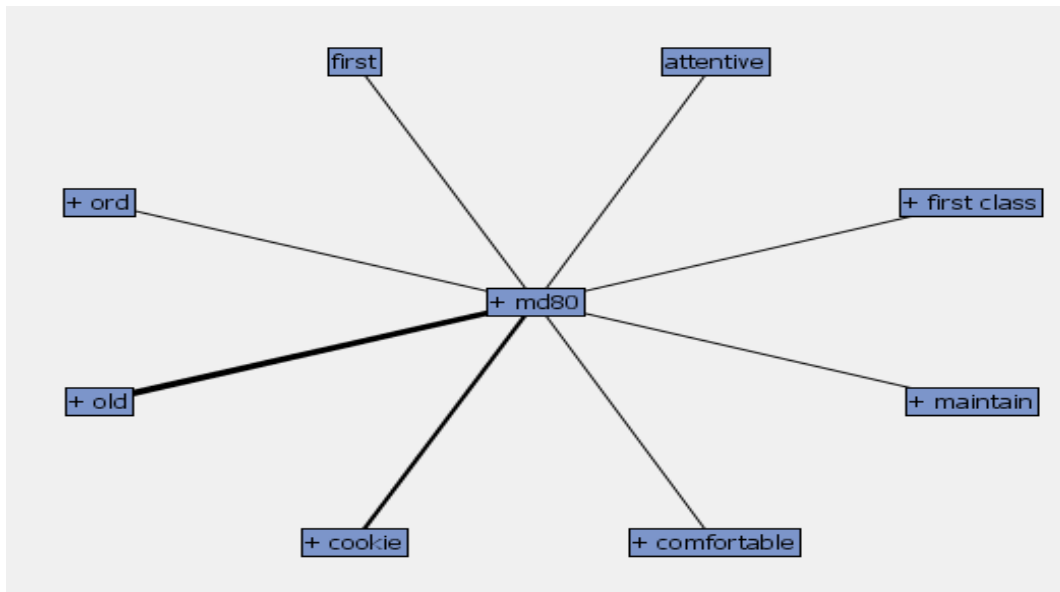


Fig 6: Concept Link Diagram for md80

The term which is being analyzed is at the center and the width of the link determines the strength of the association. Wider the link, stronger is the association, i.e. the two terms were present in the same document for more number of times.

- The term 'md80' is associated strongly to the term 'old'.
- The term 'business class' is strongly associated to many terms such as 'seat', lounge' and 'upgrade'.

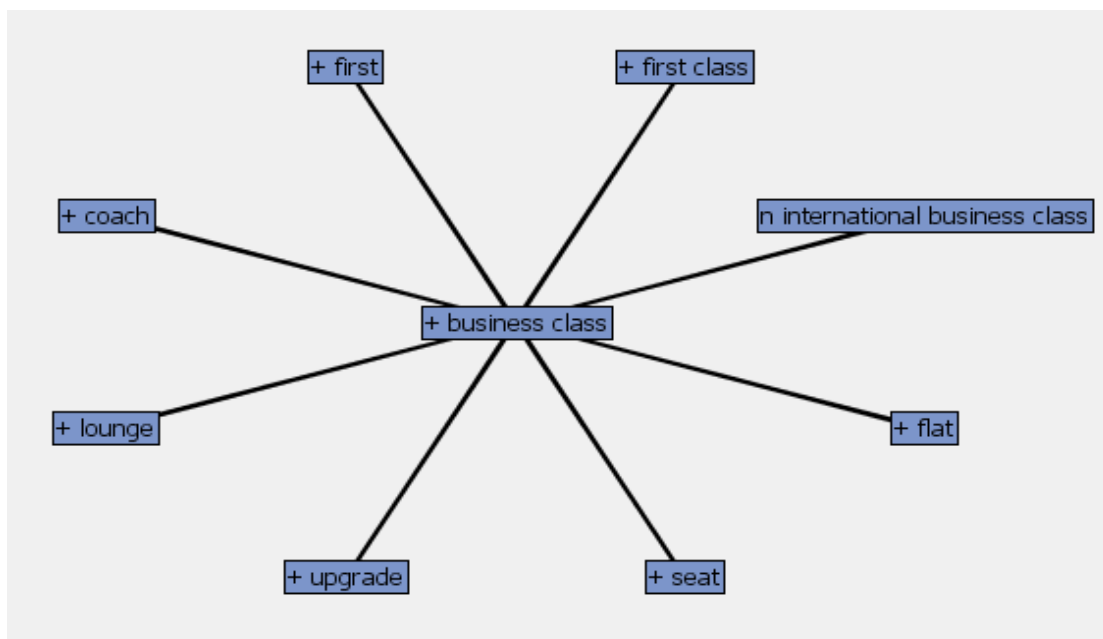


Fig 7: Concept Link Diagram for Business Class

TEXT CLUSTERING:

Text Clustering assigns each document to a cluster using Singular Value Decomposition (SVD) to reduce the curse of dimensionality. We have used hierarchical clustering in this analysis. Below is the descriptive distribution pie chart of the text clusters.

After we have done clustering, we found the frequency and percentage of the terms in the reviews. Nine clusters are generated with each containing 20 descriptive terms which describe the cluster. We can see that the classification is based on different contexts such as one containing all the terms which are related to seating comfort, other cluster containing reviews regarding flight delays and so on.

From the cluster frequency by RMS and distance between clusters graph we can say that the clusters are well separated from each other and the frequency is also well distributed.

Cluster ID	Descriptive Terms	Frequency	Percentage
1	+bag +check +luggage +baggage +lose +clai...	581	11%
2	+flight economy +'american airlines' +fly +do...	958	19%
3	+flight +seat +service +good +food +class +...	736	14%
4	+hour +delay +airport +leave +late +wait +ga...	865	17%
5	+customer +ticket 'us airways' ni +phone +sp...	1034	20%
6	vegas +las +'las vegas' +domestic 'domestic ...	99	2%
7	+seat +sit +row +extra +leg +room +move ...	360	7%
8	+ticket +pay +fee +charge +credit +card +pu...	311	6%
9	+'new york' +city economy +domestic +intern...	224	4%

Fig 8: Cluster Generated along with Cluster ID's

TEST RULE BUILDER:

The text rule builder node is used to generate a set of rules using subsets of terms to predict a target variable. Here the target variable is binary i.e. whether the feedback is positive or negative. While collecting the data, since we had the customer rating as well, based on its value we classified it as positive or negative. All the ratings which had value less than 5 were classified as negative and rest were positive.

Target Value	Rule #	Rule	Precision	Recall
NEGATIVE	1	hour & ~excellent & ~friendly & ~comfortable ...	99.51%	35.37%
NEGATIVE	2	miss & ~good	98.85%	44.95%
NEGATIVE	3	customer & ~great	97.35%	57.49%
NEGATIVE	4	suppose	97.45%	59.93%
NEGATIVE	5	uncomfortable	97.12%	64.63%
NEGATIVE	6	cancel & ~food	97.02%	68.12%
NEGATIVE	7	horrible	97.09%	69.86%
NEGATIVE	8	rude & ~excellent	96.36%	73.87%
NEGATIVE	9	pay & ~friendly & ~lounge & ~great	96.35%	78.22%
NEGATIVE	10	poor & ~well	96.44%	80.31%
NEGATIVE	11	force	96.47%	81.01%
POSITIVE	12	excellent	92.06%	22.57%
POSITIVE	13	good & ~customer & friendly	92.08%	36.19%
POSITIVE	14	comfortable & ~bad	89.40%	52.53%
POSITIVE	15	segment	89.94%	55.64%
POSITIVE	16	great & ~wait	87.05%	65.37%
POSITIVE	17	on time & ~miss & ~rude & airline	87.13%	68.48%
POSITIVE	18	enjoyable	87.19%	68.87%
POSITIVE	19	md80	86.67%	70.82%
POSITIVE	20	fresh	86.79%	71.60%

Fig 9: Text Rule Builder Rules

The text rule builder in this case generated a set of 20 rules. With the presence or absence of a word or group of words in a review, it can be classified either positive or negative. The results can be interpreted as follows:

- Rule 1 specifies that with the presence of the term 'hour' and with the absence of terms such as excellent, friendly and comfortable, we can say with a precision of 99.51 that the review is a negative one.
- Similarly, rule 17 specifies that with the presence of terms like 'on time' and 'airline' and with the absence of terms like 'miss' and 'rude', we can say with a precision of 87.13 that the review is a positive one.
- If we go through rule 19, it states that the presence of word 'md80' alone guarantees with a precision of 86.67 that the review is a positive one. This result is in contrast with the concept link, according to which the term 'md80' is strongly associated with the term 'old'. If we go through few of the reviews, we will find that in spite of the fact that 'md80' is an old flight, passengers don't hesitate to fly in this. They find the attendants very friendly and the seating also comfortable.
- In a similar way, considering the results from text rule builder and observing the concept links, detailed analysis can be done on every individual entity.
- The training and validation misclassification rate for the model are 15.16% and 19.04% respectively.

TEXT PROFILE:

The Text Profile node enables you to profile a target variable using terms found in the documents. As a special case of this, a target time variable can be used to display how terms change over time. The segments which are obtained from the SOM/Kohonen are considered as the target variable and the text reviews are profiled against them. The table shows the set of terms which collectively describe the segment. We can see that there is a strong relationship between the 2nd and the 6th segments as both of them are based on similar context like the features provided by the credit cards and the facilities offered by the airlines.

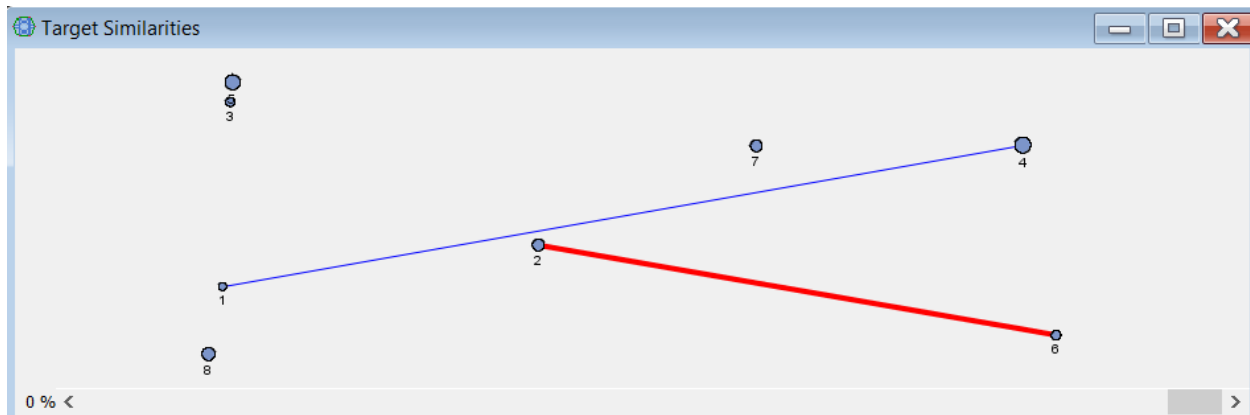


Fig 10: Target Similarities

Name	Value	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Freq	Corpus	SOI Seg ID
Corpus		consistent	outside	coast	american...	rochester...	row/Loc	march/Loc	pleasantly	5168	1	
SOM SE...1	review	good	customer	service	american...	airline	fly	charlotte/...		154		
SOM SE...2	card	credit	mile	offer	great	charge	board	food		653		
SOM SE...3	luggage	bag	las	vegas	las vegas...	lose	check	claim		320		
SOM SE...4	us airway...	agent	cancel	airline	rude	help	reservation	customer		1168		
SOM SE...5	new york/...	economy	domestic	domestic...	internatio...	los	city	angeles		1064		
SOM SE...6	seat	great	option	comforta...	food	cabin	screen	extra		368		
SOM SE...7	class	san	first	trip	juan	san juan/...	phoenix/L...	diego		636		
SOM SE...8	ticket	miami/Loc	connection	voucher	line	miss	passenger	pay		805		

Fig 11: Profiled Variables

CONCLUSION

This research was intended to analyze customer reviews of American Airlines using SAS® Enterprise Miner 13.2. Exploratory analysis combined with text analytics provided a sound understanding of text data.

- Using the text rule builder node in SAS® Enterprise Miner, we can classify the reviews into positive or negative. This type of analysis can be extremely useful to the audience that wants value for their money and also for those people who like to choose the flight based on certain criterion.
- Concept links can be used to analyze the occurrence of a term with other terms and also the strength of the association between the terms.
- Using the text rule builder node in SAS® Enterprise Miner, we can classify the reviews as positive or negative.
- We can use the model from the text rule builder and the score node to classify the new reviews.
- The airlines can do this analysis in regular time intervals in order to know what customers think about their service.

REFERENCES

- Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS® by Goutam Chakraborty, Murali Pagolu, Satish Garla.
- SAS® Institute Inc. 2014. Getting Started with SAS® Text Miner 13.2. Cary, NC: SAS® Institute Inc.

ACKNOWLEDGEMENTS

We thank Dr. Goutam Chakraborty (Director for Business Analytics program, Founder of SAS OSU Data mining certificate program) Oklahoma State University, for his support, guidance and encouragement throughout our research work.

AUTHORS

Saurabh Kumar Choudhary is a full time Graduate student at Oklahoma State University. He is pursuing his Master of Science in Business Analytics. He holds a Bachelor Degree in the field of Electronics and Telecommunication and is an author of research paper in an esteemed International Journal. Saurabh has also successfully completed his analytics project works at school and aspires to make every data valuable with his skills.

Rajesh Tolety is a Graduate Teaching Assistant and full time student at Oklahoma State University. He is pursuing his Masters in the field of Business Analytics. Holding a Bachelors in the field of Information Technology and having worked for 3 years in the field of providing cloud based management software, he completely understands what value data bring to the table. He is interested in the field of predictive modeling and text analytics.

CONTACT INFORMATION

Your comments and questions are encouraged and valued. Contact the authors at:

Rajesh Tolety
Master of Science in Business Analytics
Oklahoma State University
rajesh.tolety@okstate.edu,
<https://www.linkedin.com/in/rajeshtolety>

Saurabh Kumar Choudhary
Master of Science in Business Analytics
Oklahoma State University
saurabh.k.choudhary@okstate.edu,
<https://www.linkedin.com/in/saurakc>

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.