Airlines Sentiment Analysis

Introduction

- The dataset is taken from Kaggle's open data source. https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment
- We did a Sentiment analysis of the different travelers reviews of their flights as good or bad.
- Through the text analysis, we analyzed the various feedback for six different airlines given in the month of FEB 2015

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
airline	Rejected	Nominal	No		No	.	.
airline_sentime	Target	Nominal	No		No	.	.
text	Text	Nominal	No		No	.	.
tweet_id	ID	Nominal	No		No	.	.

Airline Sentiment: This column describes overall response of the users in respect to their airline experience.

Text:This column describes in detail the comments made by the users on twitter.

Tweet_id: This column describes Id of the tweet users.

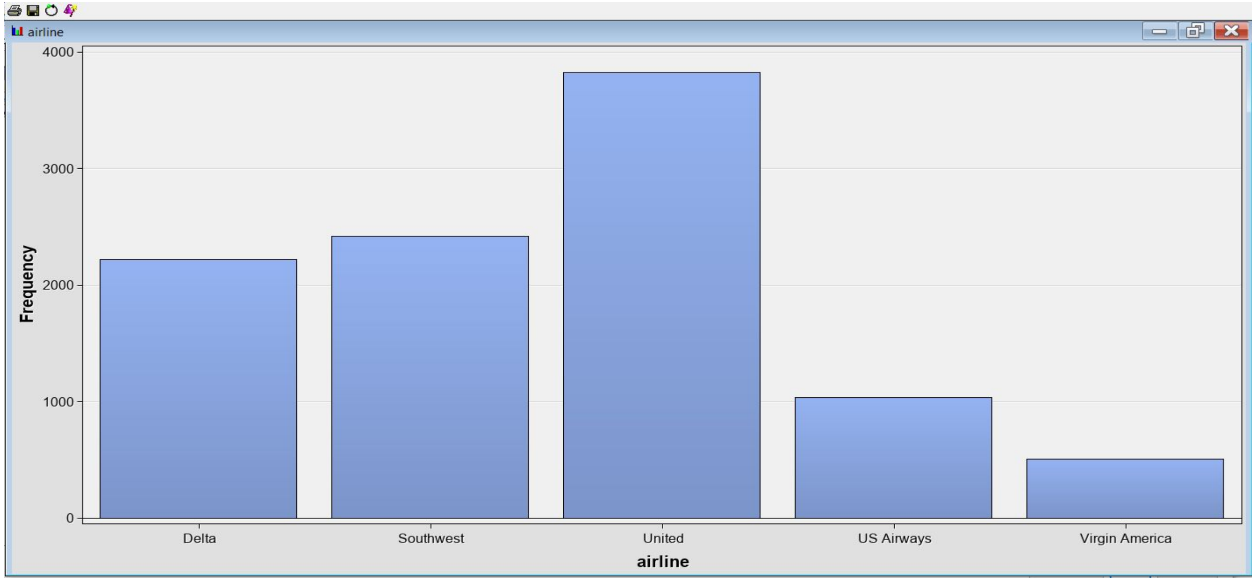
OBJECTIVE

- With this project presentation, we aim to analyze and provide a model that explains how 'Opinions' on the operation of top US airlines are formed among people.

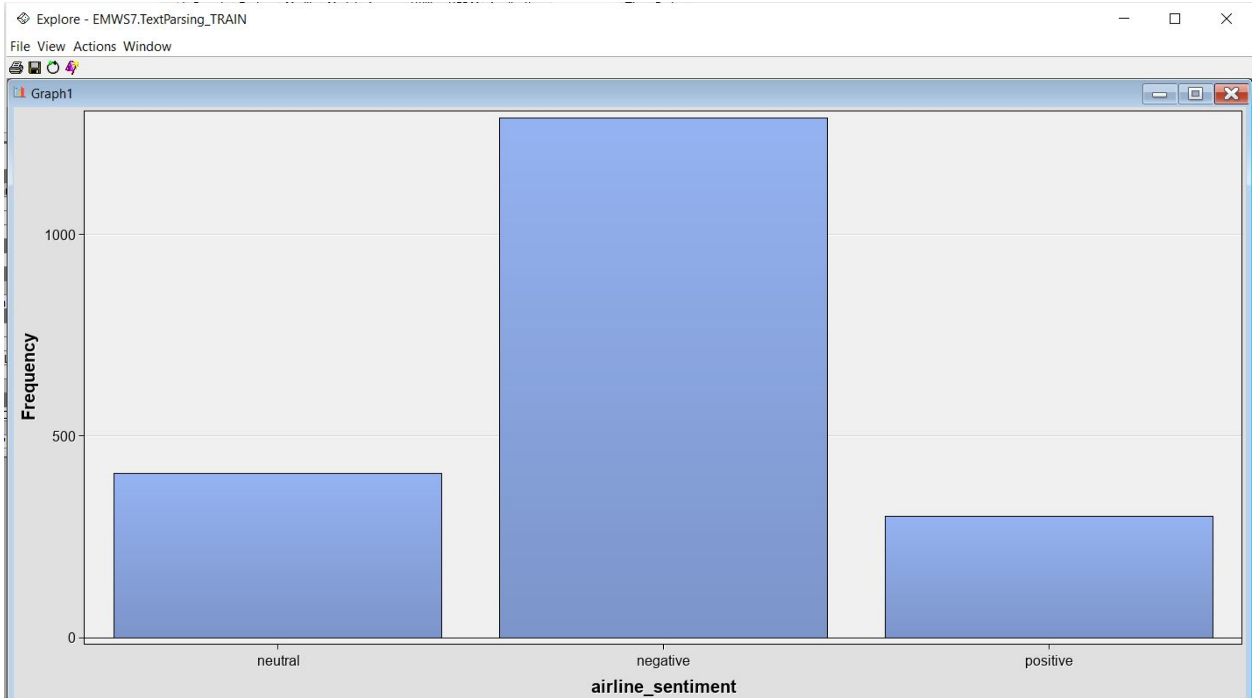
- Through the sentiment analysis done on the Twitter dataset, we aim to convey the information to the airlines so that they could use this information to deliver better service to their consumers.

- By determining whether the opinions created are positive or negative we aim to submit the final insights modeled to the airline industry for their business use.

PREPROCESSING & EXPLORATION



- There are 6 airlines, but we will not be using them as inputs to the model to avoid adding a bias.

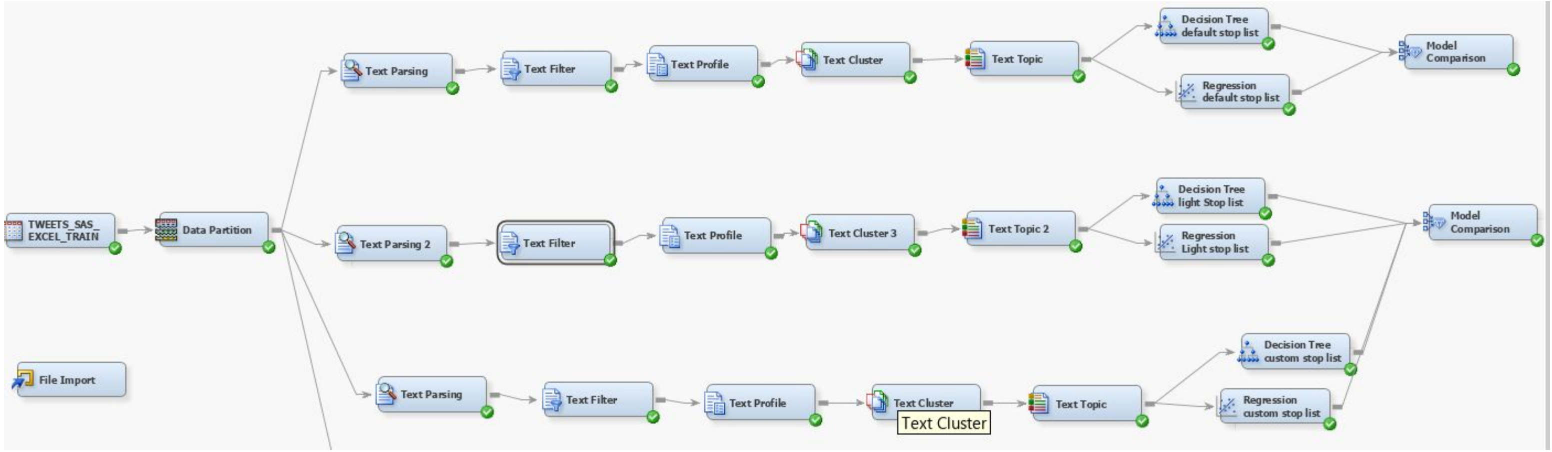


- Most of the sentiments are negative which says that people with negative sentiments are more likely to share their experiences.

Property	Value
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	50.0
Validation	30.0
Test	20.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	4/17/22 1:26 PM
Run ID	a7d3e006-7ade-44a4-badb-67f7cf824
Last Error	
Last Status	Complete
Last Run Time	4/17/22 1:29 PM
Run Duration	0 Hr. 0 Min. 2.11 Sec.

- **We're removing the neutral sentiment and and dealing only with the positive and negative sentiments.**
- **Partitioning the data in 50, 30, and 20 percent.**

PROCESS SELECTION



Stop List

STOP LIST 2

TERM	ROLE
's	Prop
(631)891-5722	Noun
-1366	Noun
0.21mbps	Noun
0/3	Num
44606	Num
42055	Num
44615	Num
0_0	Num
0xjared	Noun

Here we see 831 Stop words

STOP LIST 3

TERM	ROLE
's	Prop
(631)891-5722	Noun
-1366	Noun
0.21mbps	Noun
0/3	Num
44606	Num
42055	Num
44615	Num
0_0	Num
0xjared	Noun
44563	Num
1-2888155964	Num

Here we see 18651 Stop words

PROCESS SELECTION

Process 1: Model with default SAS Stop list

Process 2: Model with Light Stop list

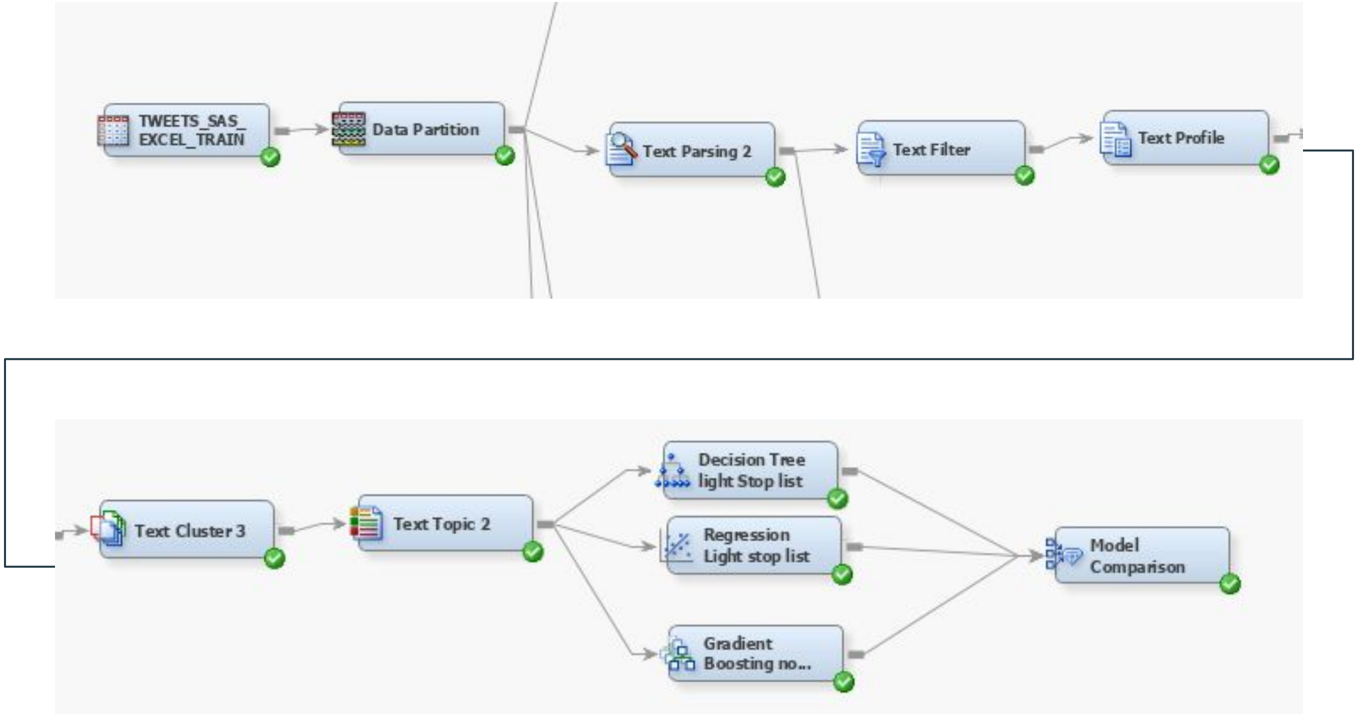
Process 3: Model with Custom stop list

Model Description	Train: Misclassification Rate ▲	Valid: Misclassification Rate
Regression Light stop list	0.120818	0.126805
Gradient Boosting light stop list	0.121512	0.138359
Decision Tree light Stop list	0.131739	0.151935
Regression default stop list	0.147686	0.152802
Decision Tree default stop list	0.16138	0.17591
Regression custom stop list	0.196048	0.196129
Decision Tree custom stop list	0.198475	0.19844

Selection criteria:

- ❑ It seems that the process with the light stop list is giving the best explainability and accuracy. Thus, we will be using the process going forward to fine-tune the model.
- ❑ Also, light stop list is working better perhaps because the data is pre-cleaned.

MODEL EXPLANATION



FINE-TUNING the MODEL

Prior to Fine-Tuning :

After trying all the combinations of Term weights and frequency weights, we are getting the best results with the Frequency weight of **Log** and Term weight of **Mutual information**.

Model Description	Train: Misclassification Rate ▲
Regression Light stop list	0.120818
Gradient Boosting no stop list	0.121512
Decision Tree light Stop list	0.131739

Model Description	Train: Misclassification Rate ▲	Valid: Misclassification Rate
Gradient Boosting no stop list	0.080083	0.136626
Regression Light stop list	0.089617	0.132293
Decision Tree light Stop list	0.100537	0.14818

Using SVD and Text Topic Node:

- Using the default value of **High** SVD resolution, **max** 100 SVD dimensions, and 4 clusters, we are getting better explainability:
- The accuracy results didn't improve significantly.

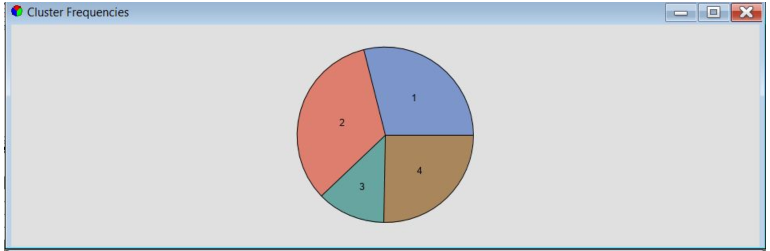
- Final Accuracy:

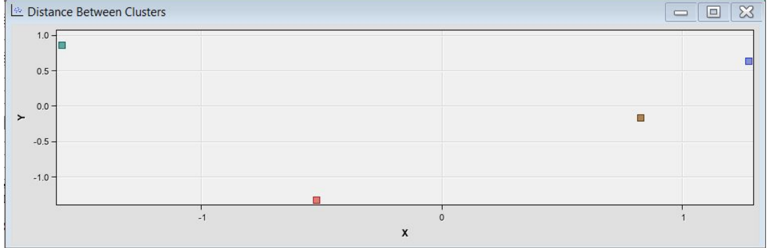
Model Description	Train: Misclassification Rate ▲	Valid: Misclassification Rate
Gradient Boosting no stop list	0.080083	0.136626
Regression Light stop list	0.089617	0.132293
Decision Tree light Stop list	0.100537	0.14818

Model Description	Train: Roc Index ▼	Valid: Roc Index
Gradient Boosting no stop list	0.951	0.887
Regression Light stop list	0.95	0.901
Decision Tree light Stop list	0.861	0.765

CLUSTERS

.. Property	Value
General	
Node ID	TextCluster8
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
⊟ Transform	
SVD Resolution	Medium
Max SVD Dimensions	100
⊟ Cluster	
Exact or Maximum Number	Exact
Number of Clusters	4
Cluster Algorithm	Expectation-Maximization
Descriptive Terms	20
Status	
Create Time	4/18/22 4:52 PM
Run ID	7c6b605c-008f-4989-b94a-0bad80fe7
Last Error	
Last Status	Complete





Clusters

Cluster ID	Descriptive Terms
1	southwestair jetblue +thank so +good virginamerica thanks much +appreciate +great +know +flight why just +delay +eve...
2	usairways americanair +unite no +hour flightled cancelled late +flight 'on hold' +delay +call +minute now +phone +wait +t...
3	not +do americanair +even +call +unite why +help +change +ticket +book +tell +phone usairways still +agent +get +sea...
4	+guy very +see +fly new +response +great southwestair +good +airline united just +service again +customer much than...

SENTIMENTS

PROCESS 2		
SL no.	POSITIVE	NEGATIVE
1	response	cancelled
2	thanks	on hold
3	great	flighted
4	great job	reschedule
5	great people	charge
6	experience	strand

Major sentiments from the positive cluster 1 and negative cluster 3

Cluster ID	Descriptive Terms
1	southwestair jetblue +thank so +good virginamerica thanks much +appreciate +great +know +flight why just +delay +ev...
2	usairways americanair +unite no +hour flightled cancelled late +flight 'on hold' +delay +call +minute now +phone +wait ...
3	not +do americanair +even +call +unite why +help +change +ticket +book +tell +phone usairways still +agent +get +s...
4	+guy very +see +fly new +response +great southwestair +good +airline united just +service again +customer much th...

TEXT TOPICS

Positive Sentiment Topic Terms

+thank,southwestair,jetblue,+appreciate,+response
jetblue,+flight,no,+be,+delay
southwestair,cancelled,+flight,on hold,+be
not,+do,+be,+appreciate,+unite
ðÿ,jetblue,southwestair,ï,best
so,much,jetblue,+appreciate,southwestair
virginamerica,+guy,thanks,+fly,â
+great,+guy,+great flight,thanks,+job
+good,+guy,jetblue,â,not
+hour,+flight,usairways,cancelled,no

Terms

Topic Weight	+	Term	
0.941	+	great	Adj
0.155	+	guy	Noun
0.135	+	great flight	Noun G
0.124		thanks	Noun
0.098	+	job	Noun
0.082		great service	Noun G
0.081		great job	Noun G
0.057	+	thank	Verb
0.046		great people	Noun G
0.043		united thanks	Noun G
0.042		jetblue	Pron

Negative Sentiment Topic Terms

Topics

cancelled,flightled,+flight,southwestair,on hold
+customer,+service,customer service,+bad,ever
+do,not,americanair,+get,+call
+thank,jetblue,southwestair,so,much
+unite,+hour,+delay,+plane,now

Terms

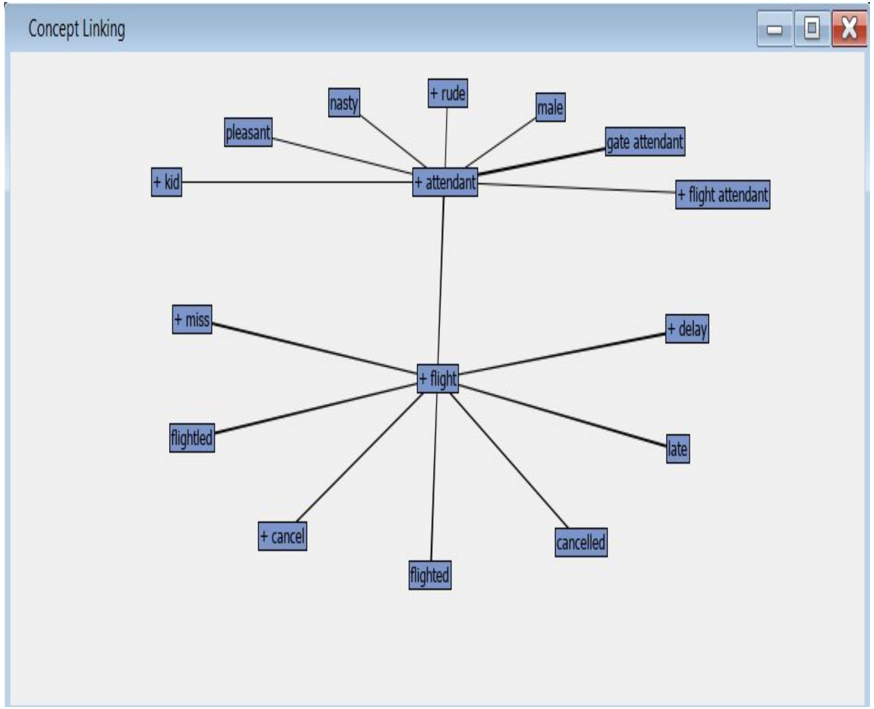
Topic Weight	+	Term
0.5		cancelled
0.387		flightled
0.362	+	flight
0.203		southwestair
0.197		on hold
0.191		flighted
0.178	+	hour
0.155	+	get
0.143	+	be
0.134		americanair
0.119		help

BUSINESS CASE

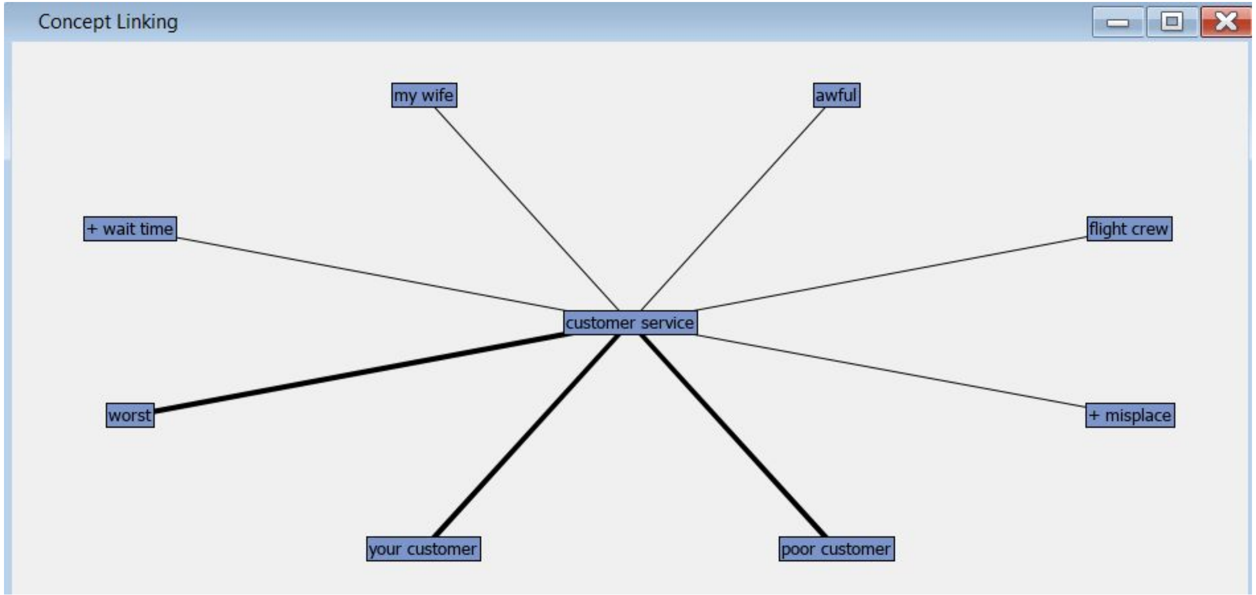
- ❑ Airlines can use the text analytics to understand where they can **improve the service.**
- ❑ They can track the Twitter to understand what their **customer needs** are.
- ❑ They can use the **concept links to understand relationship between terms** used in the tweets.
- ❑ They can use this model to **understand the polarity of a tweet** from their customers.

Concept Linking

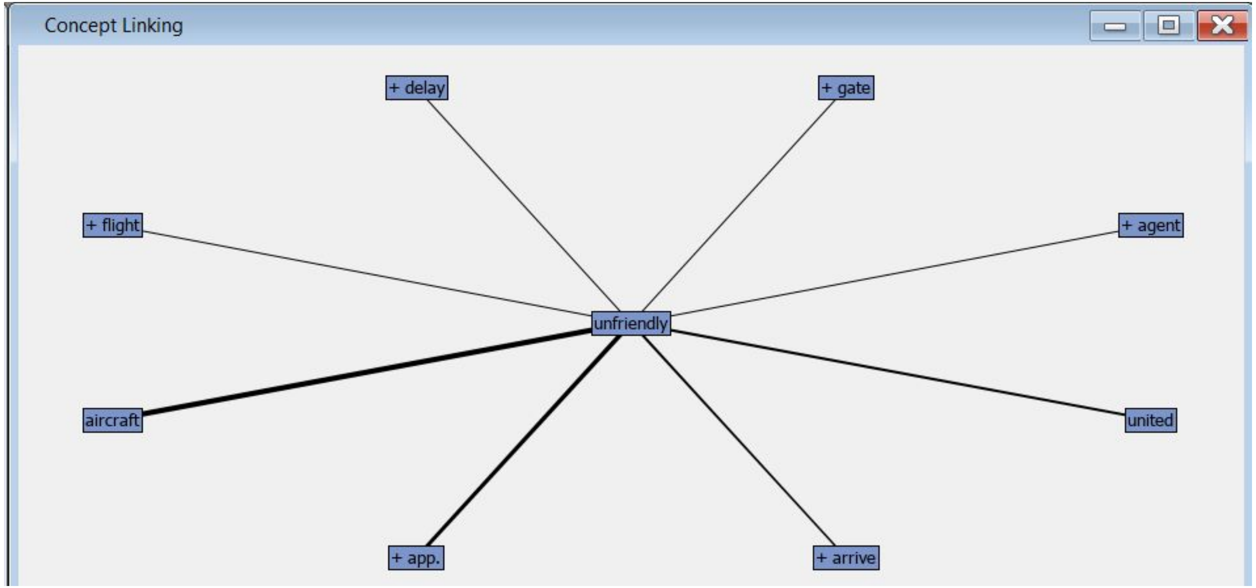
- Through concept linking we have explored the relationship between terms and their strength of association.
- It represents the terms associated with the positive and negative sentiments of the people towards airline industry.



Positive insights



Negative insights



Conclusion

Further Improvement

References

- Data set -https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment
- https://towardsdatascience.com/sentiment-analysis-on-us-twitter-airline-dataset-1-of-2-2417f204b971
- https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0015-2


Thank You!


QUESTIONS?