Sentiment Analysis: US Airline

Presented by Niraj Bangari, Jared Choy, Sandeep Nair, and Gianni Spiga

TABLE OF CONTENTS

01

Introducing Our Data

Observing the sentiment of tweets from 2015.

The Data

“A sentiment analysis job about the problems of each major U.S. airline. Twitter data was scraped from February of 2015 and contributors were asked to first classify positive, negative, and neutral tweets, followed by categorizing negative reasons (such as "late flight" or "rude service").

—Crowdflower's Data for Everyone Library

The Data

- Our dataset comes from *Crowdflower's Data for Everyone library* by way of Kaggle.
- In our dataset, we have 14,640 rows
- Variables in the dataset include:
  - Text of Tweet (string)
  - Airline Tweet was mentioning
  - Sentiment of Tweet (string)
  - Reason for Negative Tweet (string)
  - Username
  - Timestamp
- We will primarily interested in the text variable for our predictor and the sentiment of the tweet for our response.

Our U.S. Airlines of Interest

American Airlines

DELTA

Southwest

United Airlines

U·S AIRWAYS

Virgin america

Our Goal

To perform different classification techniques on text data to accurately predict the intended sentiment of a user's tweet.

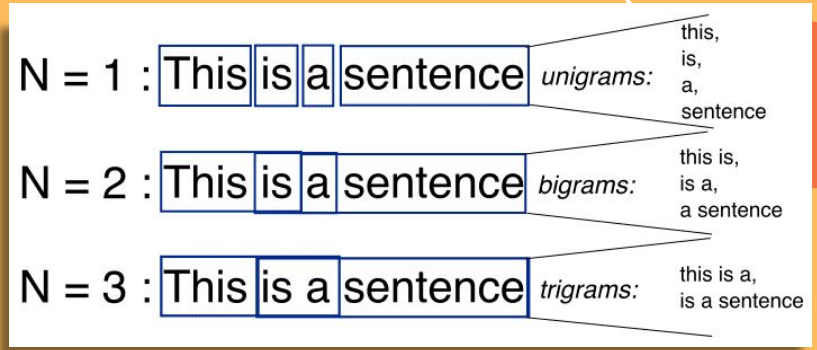
02

Preprocessing & Exploratory Data Analysis

Discovering initial results of our data.

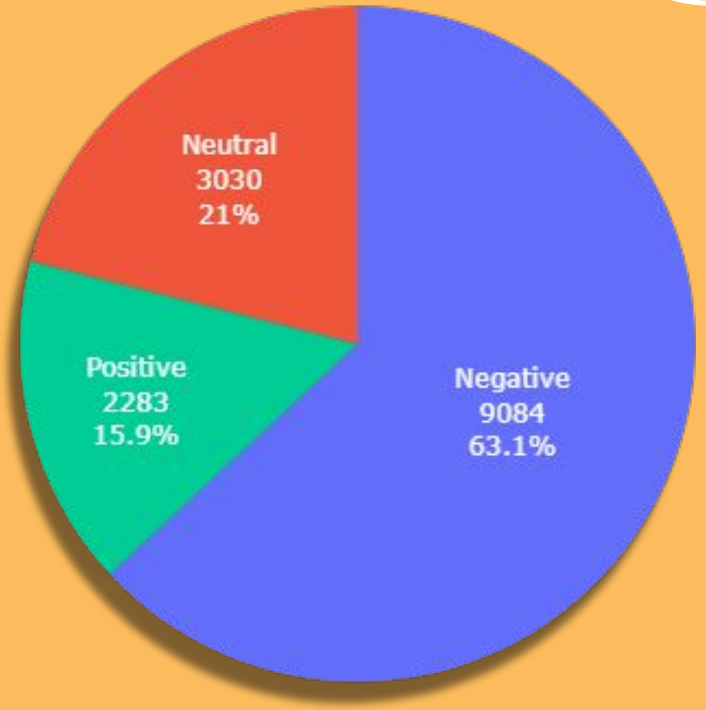
Pre-Processing

- Tokenize: Convert tweets into n-grams.
- N-Grams:
  - N-grams are contiguous sequences of 'n' items (words, characters) from a given sample of text.
  - In our analysis, we only use Unigrams (1-gram) and Bigrams (2-gram)
  - Large 'n' can lead to sparsity.
- Token Cleaning: Remove special characters and duplicates. Converted all letters to lowercase



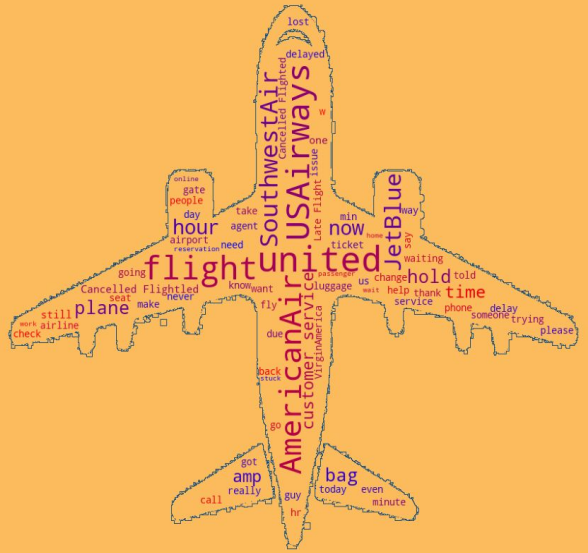
Frequency of Our Labels

- From the pie chart, negative class has the most labels, followed by neutral, and then positive.
- This class imbalance, while noticeable, does not give us enough concern to reparametrize our labels.

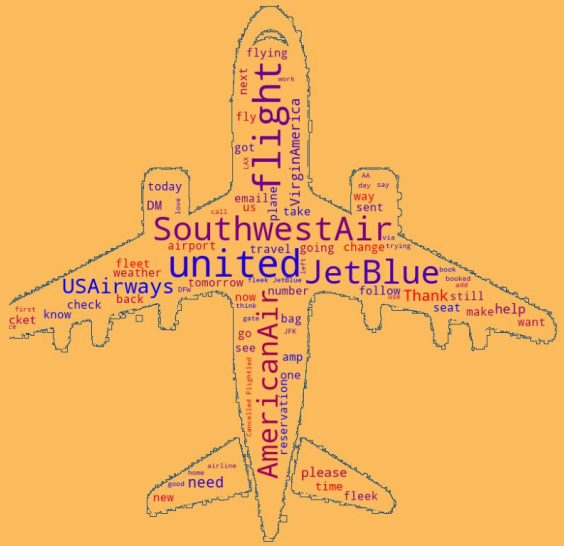


Frequency of Words

Positive



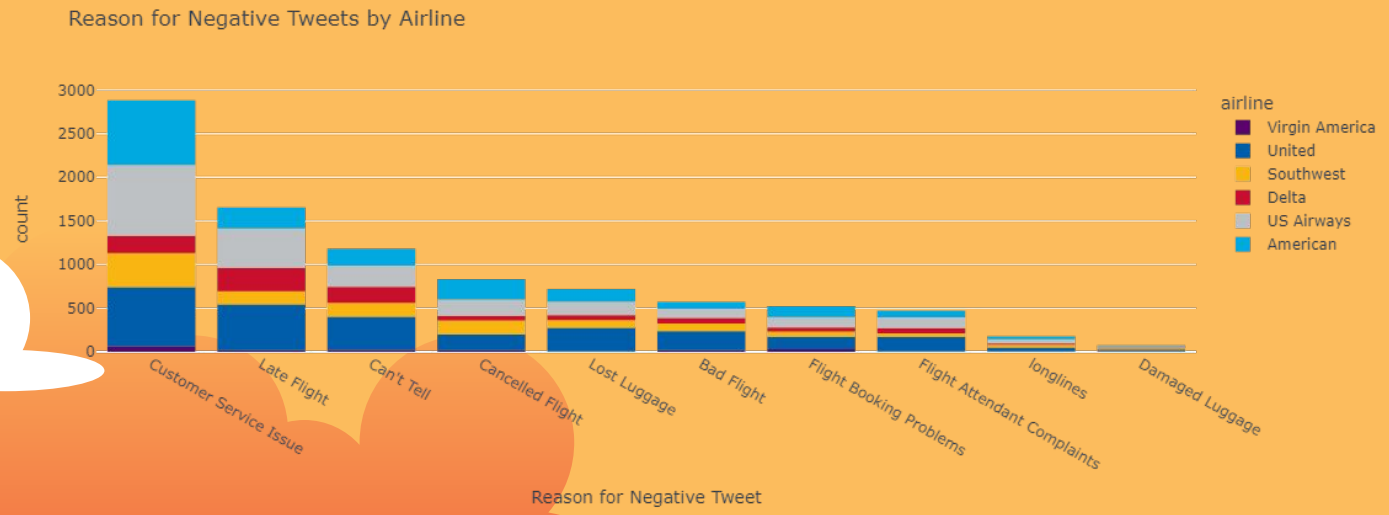
Neutral



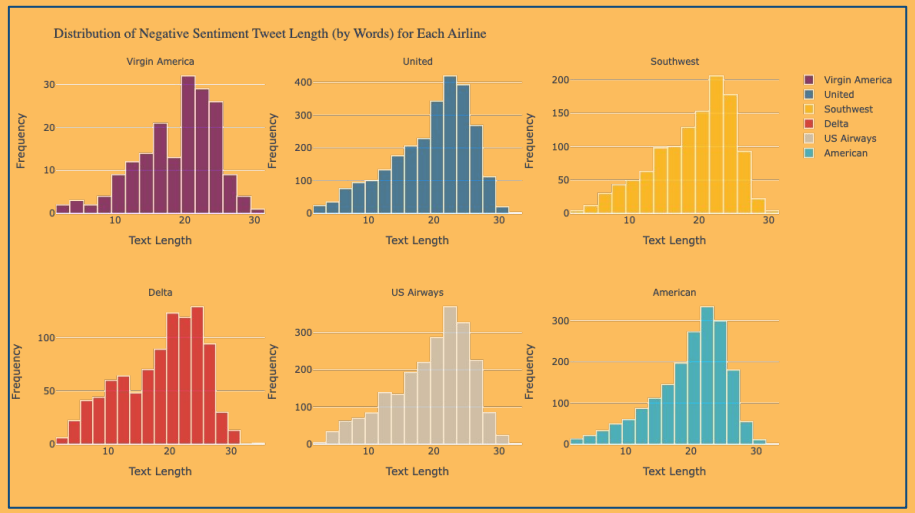
Negative

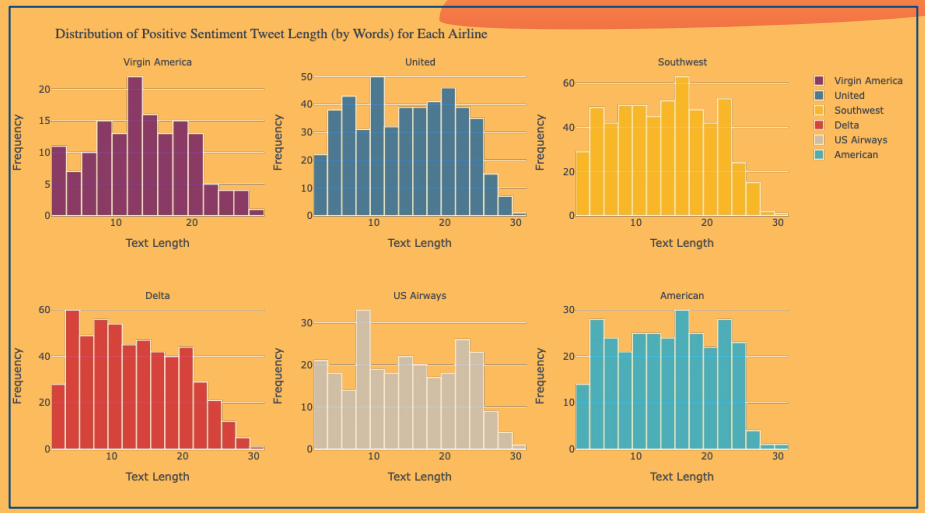
Reasons for Negative Tweets

- From the histogram, we can see the most frequent reason for negative tweets are Customer Service Issues, led by US Airways.
- A little over 1000 negative tweets could not have a reason labeled (“Can’t Tell”).



Text Length By Airline





Kruskal Wallis:

*H*o: The median text length is the same across all airlines. Vs. *H*α: The median text length is different for some airline(s)

03

Sentiment Analysis and Modeling

Exploration of different models for classification

Classification Models

SVM	Hyperplane Classification
Multinomial	Multi-class classification with and without penalization
RNN/CNN	Neural Network Classification
Transfer Learning	BERT

SVM

01

Hyperparameter Tuning

Choosing the best settings to classify. Not learned from the data.

02

N-Grams

Do phrases help improve accuracy in classification?

03

Alternative Models

Nu-Support, LinearSVC, Penalty. Kernel experimentation.

SVM Methodology

- SVM is a popular classification technique that is focused on finding a hyperplane boundary between different classes. This technique is suitable for our data, given the dimensionality of text.
- Split into training and test set. Use TF-IDF to vectorize our text into numerical units.
- Train data based on whichever method we are using.
    - Hyperparameterization - Finding best parameters of SVM model based on the training data.
    - SVC - The general definition of SVM modelling. Defining a linear usage will change the kernel type. Linear is specialized and works for high-dimensional data.
    - N-Gram - Concatenate text into multiple clumps, feeded to model. I.e sets of two, three
    - Nu-Support - Introduces “Nu” parameter that controls the number of support vectors and training error

Classification Report

Method	Accuracy	Recall	F1 Score (Weighted Avg)
Poly SVC	78.37%	0.79	0.78
Linear SVC	78.78%	0.78	0.78
Linear SVC With Penalty	78.92%	.79	0.78

Classification Report (cont.)

Method	Accuracy	Recall	F1 Score (Weighted Avg)
Base SVC - N Gram	78.47%	0.78	0.77
Hyperparameter	78.61%	0.79	0.78
Nu Support	79.06%	0.79	0.78

Multinomial Regression

- Type of logistic regression for multi-class classification problems.
- Suitable when the dependent variable has more than two categories, such as our labels.
- Predicts probabilities of each class, based on the input features, in our case the TF-IDF matrix.
- Assumes independence among classes.
- For the multinomial model, we can measure the performance with four metrics
    - Accuracy: ratio of correctly predicted observations to total obs.
    - F1 Score: a useful metric when dealing with imbalanced datasets such as ours, balancing Precision and Recall (higher is better).
    - ROC-AUC: the model's ability to distinguish between classes (higher the better).
    - Log-Loss: a measure of accuracy that considers the uncertainty of the predictions (lower is better).

Penalization Methods

LASSO

Performs variable selection and regularization to enhance prediction accuracy.

Elastic Net

Combines the penalties of Lasso and Ridge regression to handle collinearity and select variables.

Multinomial Models' Results w/o N-Grams

	Accuracy	F1 Score	ROC-AUC	LOG-LOSS
Standard	79.90%	0.785	0.915	0.511
LASSO*	79.06%	0.782	0.905	0.510
Elastic Net*	80.31%	0.791	0.925	0.500

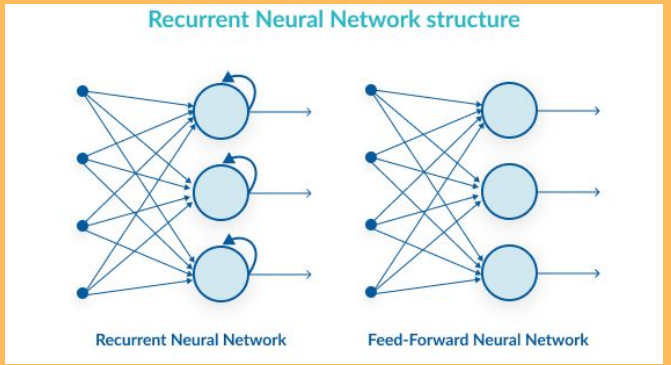
*Optimal values tuned with Grid-Search Cross Validation

Multinomial Models' Results w/ N-Grams

	Accuracy	F1 Score	ROC-AUC	LOG-LOSS
Standard	79.90%	0.788	0.915	0.511
LASSO	78.78%	0.778	0.904	0.512
Elastic Net	80.17%	0.794	0.911	0.492

RNN

- Introduction to RNNs
  - Neural network designed to recognize patterns in sequences of data
- How it Works:
  - RNNs have a "memory" that captures information about what has been calculated so far.
  - Process sequences by iterating through the elements and maintaining a state that contains information
- Applications:
  - Natural Language Processing (NLP) tasks
  - Time Series Analysis
  - Speech Recognition, Music Generation, and other sequence-based tasks.
- Implementation
  - Tensorflow



CNN

- Introduction to CNNs
  - Processes data with a grid-like topology.
  - More effective for image and video recognition, image classification, and similar tasks.
- How it Works:
  - Utilize convolutional layers that capture spatial features.
  - In text -> capture local contextual patterns, such as specific word combinations or phrases
- Applications:
  - Image and Video Analysis: Object detection, image segmentation, face recognition, MRI scans
- Implemented by Deep Learning Frameworks such as Tensorflow

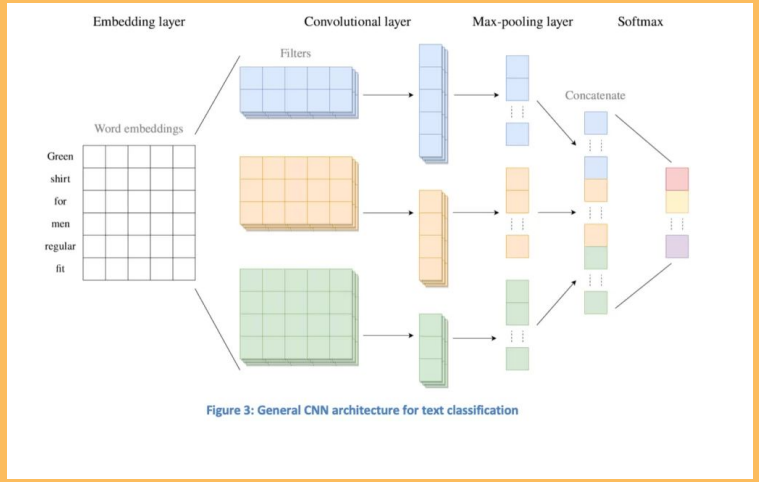
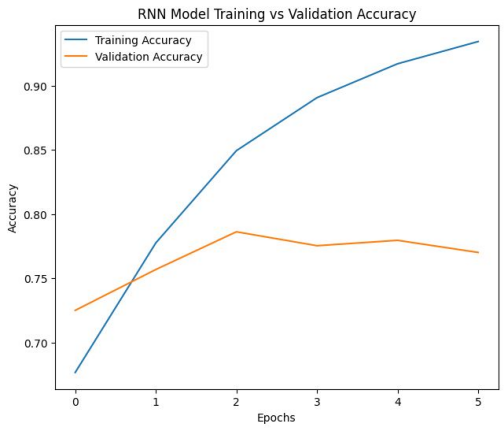
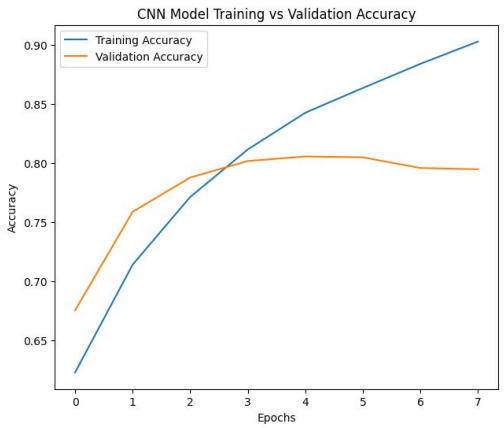


Figure 3: General CNN architecture for text classification

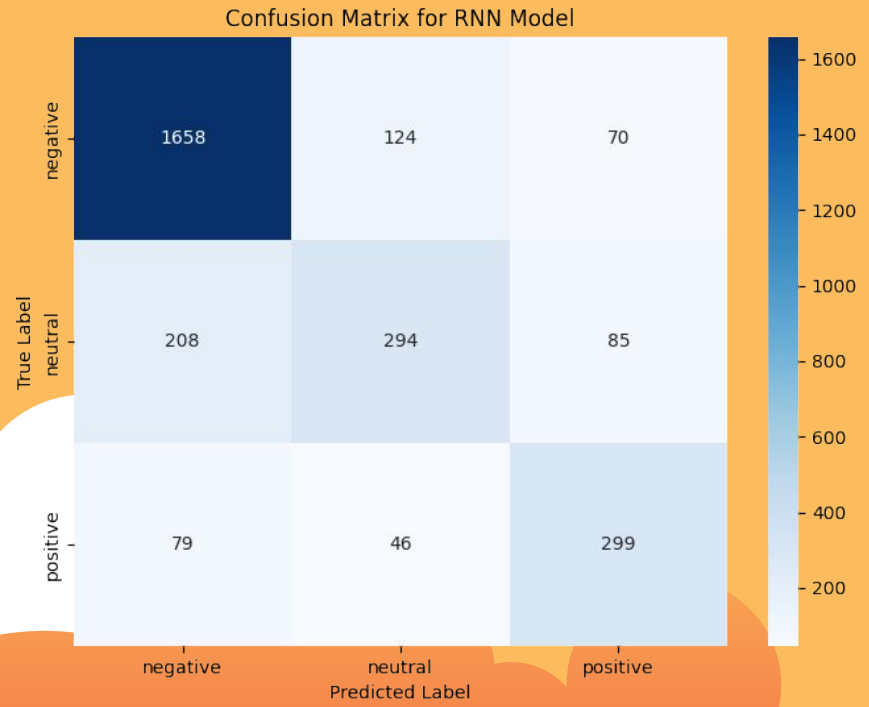
Results

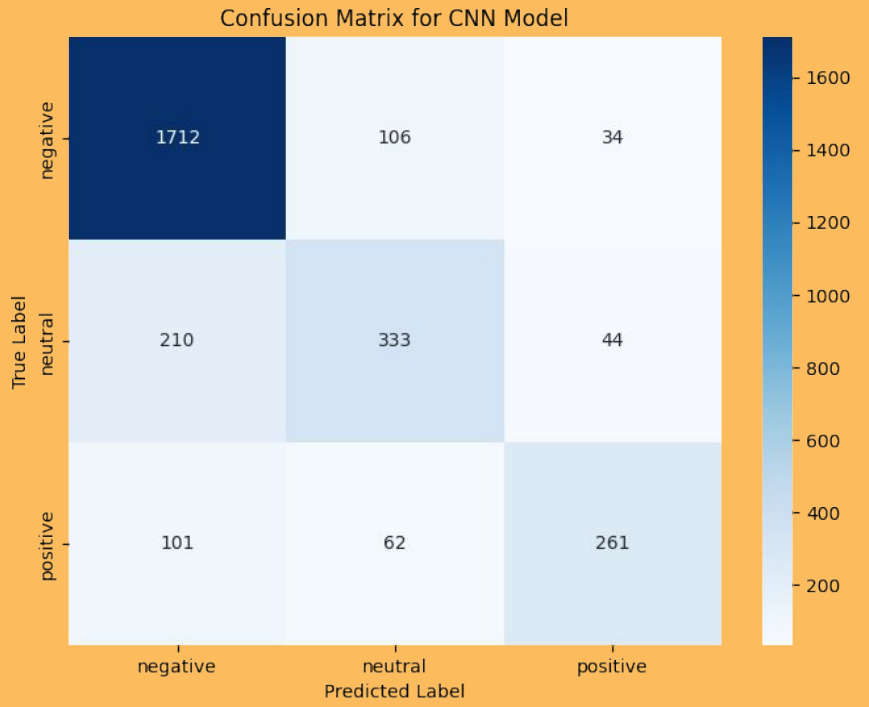
	Accuracy	F1 Score	ROC-AUC	LOG-LOSS
RNN	78.8%	0.78	0.89	0.56
CNN	80.5%	0.79	0.89	0.53



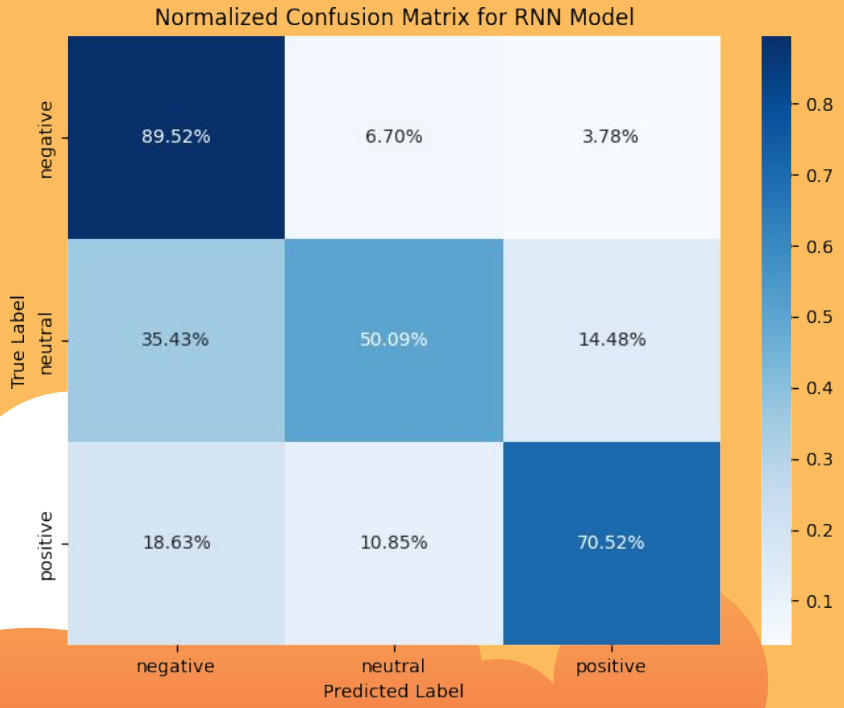


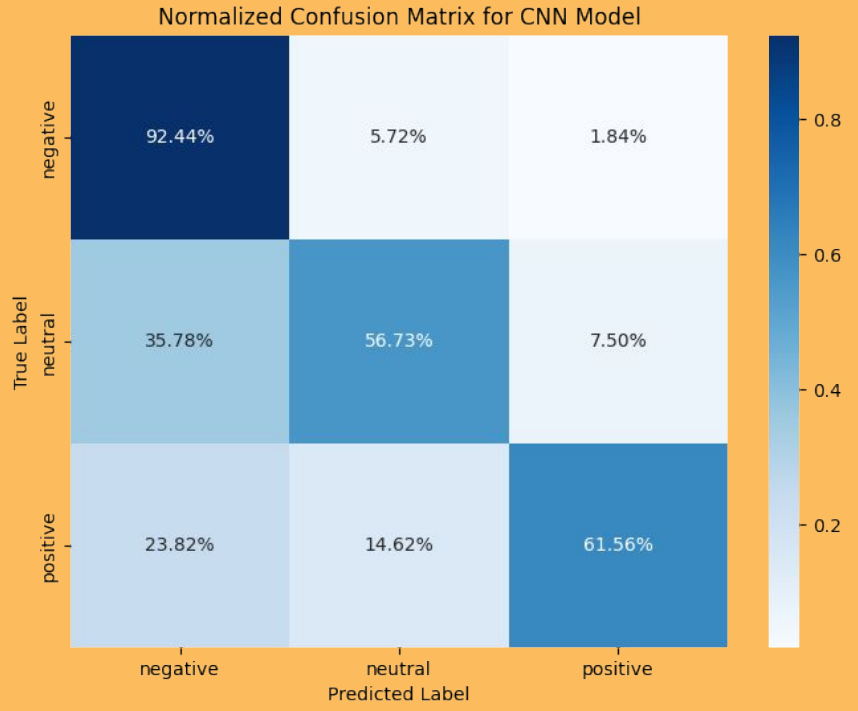
Results


Confusion Matrix for RNN Model
True Label
negative
neutral
positive
1658
124
70
208
294
85
79
46
299
negative
neutral
positive
Predicted Label
1600
1400
1200
1000
800
600
400
200


Confusion Matrix for CNN Model
True Label
negative
neutral
positive
1712
106
34
210
333
44
101
62
261
negative
neutral
positive
Predicted Label
1600
1400
1200
1000
800
600
400
200

Results


Normalized Confusion Matrix for RNN Model
True Label
negative
neutral
positive
89.52%
6.70%
3.78%
35.43%
50.09%
14.48%
18.63%
10.85%
70.52%
negative
neutral
positive
Predicted Label
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1


Normalized Confusion Matrix for CNN Model
True Label
negative
neutral
positive
92.44%
5.72%
1.84%
35.78%
56.73%
7.50%
23.82%
14.62%
61.56%
negative
neutral
positive
Predicted Label
0.8
0.6
0.4
0.2

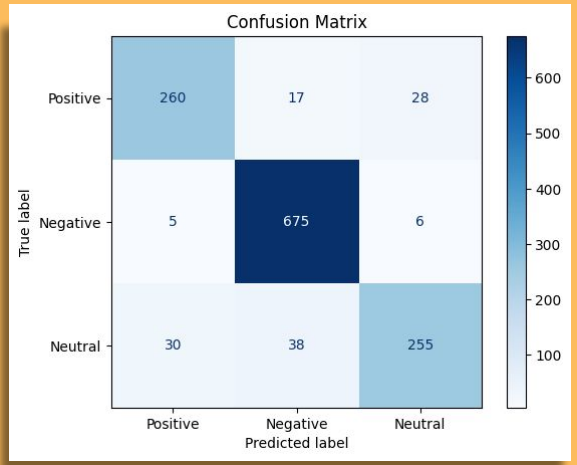
How was BERT Trained?

- BERT was pretrained with two Objectives in mind

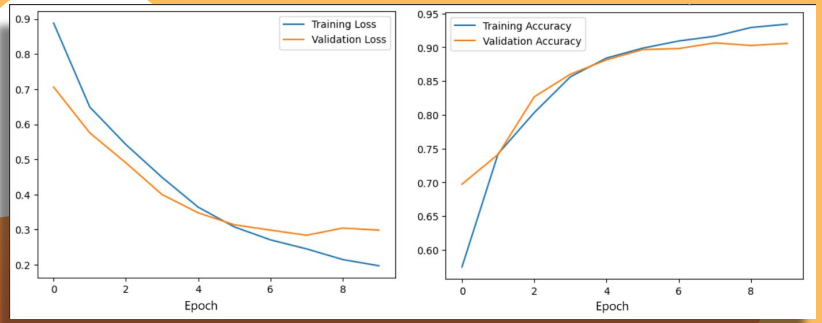
- **Masked Language Modeling**
  - 15% of words in a sentence are masked (hidden). Next, this masked sentence is ran though the model which strives to predict the masked word.
  - Allows model to learn bidirectional representation of the sentence
  - Example: The boy is [blank] milk.

- **Next Sentence Prediction**
  - Training set is created uses 50% Correct sentence pairs and 50% random sentence pair. The model is predicts if a given sentence pair is correct or random
  - Examples:
    - Correct sentence pair: Matt went fishing. He caught a massive trout
    - Random sentence pair: Matt went fishing. Time is more valuable than money

BERT Results



Method	Precision	Recall	F1-score
Positive	0.88	0.87	0.87
Negative	0.88	0.96	0.95
Neutral	0.86	0.82	0.84



Model Accuracy = 90%

04

Results and Conclusions

Presenting our final results and discussing the next steps.

Our Best Models: Accuracy

80.31%
Elastic Net

80.54%
CNN

90.00%
Transfer Learning

What's Down the Road?

- Exploring Ridge Penalty on Multinomial Regression
  - Given the lesser performance of LASSO compared to Elastic Net and Standard model, we are curious to see how Ridge penalty alone would perform in our predictive ability.
- Creating a predictive model using both text and other included features of the data
  - We can account for things such as time of day Tweet was made, retweet count, etc.
- Refine hypertuning parameters further with more testing.
- Improve overfitting in NN Models

THANK YOU

Do you have any questions?

Sources

Sentiment Analysis of Twitter Data: A Survey of Techniques [1]

Emotion Detection using Natural Language Processing [2]

Sentiment Analysis of Political Tweets for Israel using Machine Learning [3]

Working With RNNS [4]

Working With CNNS [5]

BERT 101 🤗 State Of The Art NLP Model Explained [6]

12 / 07/ 2022

Saturn is a gas giant and has several rings. It's composed mostly of hydrogen and helium. It was named after the Roman god of wealth and agriculture

Multinomial Regression

Our Methods of Expansion

08:00 am	**LASSO**	Using the L1 penalty
12:00 pm	**Elastic Net**	Combining L1 and L2
6:00 pm	**N-Grams**	Increasing Predictors

01

NAME OF THE SECTION

You can enter a subtitle here in case you need it

TABLE OF CONTENTS

WHOA!

So many airplanes and flights!

MANUFACTURERS

JUPITER

It’s the biggest planet in the Solar System

MARS

Despite being red, Mars is a cold place

SATURN

Saturn is a gas giant and has several rings

THE SLIDE TITLE GOES HERE!

Do you know what helps you make your point clear? Lists like this one:

- They’re simple
- You can organize your ideas clearly
- You’ll never forget to buy milk!

And the most important thing: the audience won’t miss the point of your presentation

A PICTURE REINFORCES THE CONCEPT

Images reveal large amounts of data, so remember: use an image instead of a long text. Your audience will appreciate it

MAYBE YOU NEED FOUR COLUMNS

JUPITER

It's the biggest planet in the Solar System

SATURN

Saturn is a gas giant and has several rings

MARS

Despite being red, Mars is a cold place

VENUS

Venus is the second planet from the Sun

17

Jupiter is the biggest planet of them all

25

Saturn is composed of hydrogen and helium

33

Mars is actually a very cold place

PRINCIPALS ASPECTS

SATURN

Saturn is a gas giant and has several rings

VENUS

Venus is the second planet from the Sun

JUPITER

It’s the biggest planet in the Solar System

MARS

Despite being red, Mars is a cold place

POPULAR MODELS

01

VENUS

Venus is the second planet from the Sun

02

JUPITER

It's the biggest planet in the Solar System

03

MARS

Despite being red, Mars is a cold place

02

NAME OF THE SECTION

You can enter a subtitle here in case you need it

CELEBRATE THE DAY

SOME CHARACTERISTICS

MERCURY

It’s the closest planet to the Sun

VENUS

Venus is the second planet from the Sun

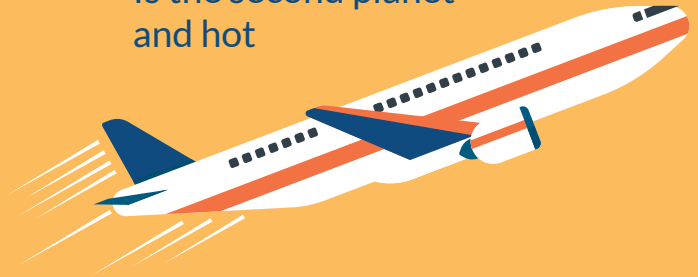
MARS

Mars is actually a very cold place

AVIATION PROCESS

VENUS

Is the second planet and hot

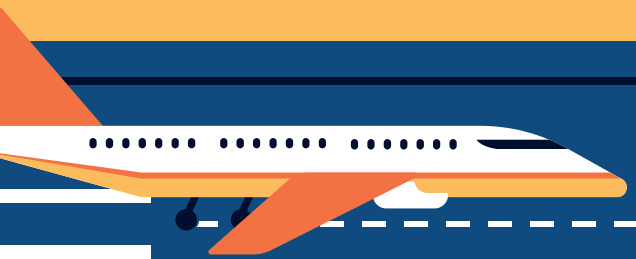


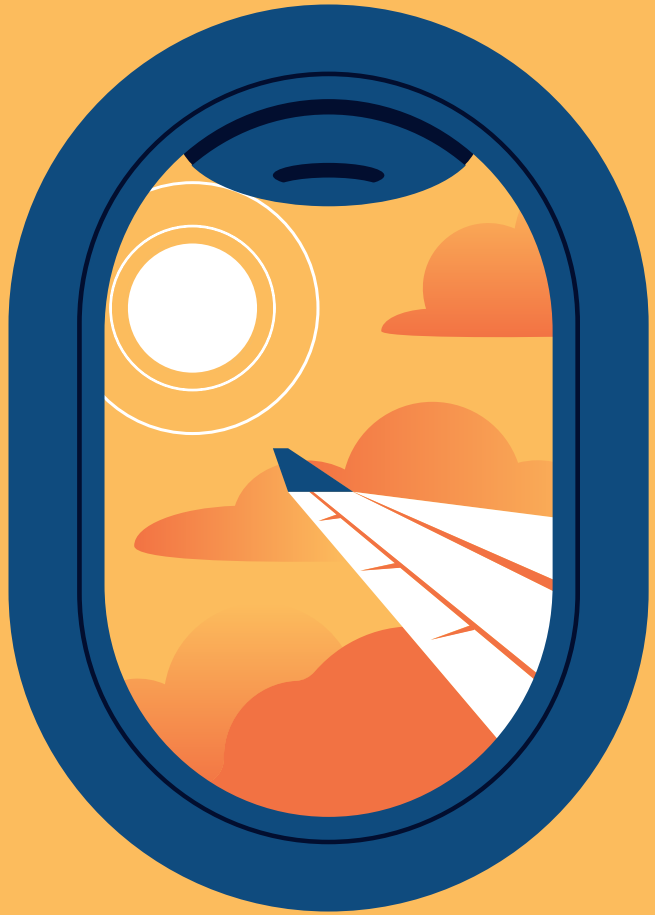
MARS

It’s composed of hydrogen and helium

JUPITER

It’s a gas giant and the biggest planet





WHAT SETS US APART?

MERCURY

It’s the closest planet to the Sun and the smallest in the Solar System

VENUS

Venus has a beautiful name and is the second planet from the Sun

OUR TEAM

JENNA DOE

You can talk a bit about this person here

DAN JONES

You can talk a bit about this person here

JOHN JAMES

You can talk a bit about this person here

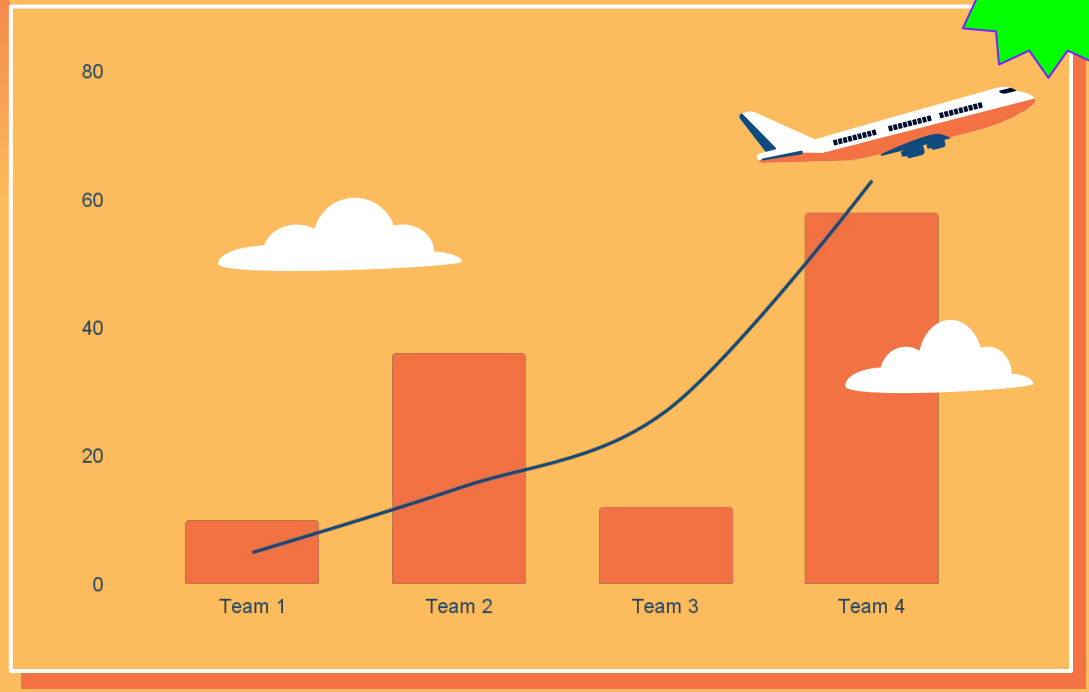
THIS IS A GRAPH

SATURN

It’s composed of hydrogen and helium

NEPTUNE

It’s the farthest planet from the Sun



Follow the link in the graph to modify its data and then paste the new one here. For more info, **click here**

STATISTICS THIS YEAR

XL

62+

TRILLION

Passengers

3+

BILLION

Dollars in global GDP

AWESOME WORDS

MAYBE YOU NEED SIX COLUMNS

MERCURY
It's the closest planet to the Sun

VENUS
Venus is the second planet from the Sun

MARS
Mars is actually a very cold place

JUPITER
Jupiter is the biggest planet of them all

SATURN
It's composed of hydrogen and helium

NEPTUNE
It's the farthest planet from the Sun

CALENDAR ACTIVITIES

XL

ACTIVITIES OF THE DAY

08:00 am	**MARS**	Is a cold place
12:00 pm	**MERCURY**	Is the closest planet
6:00 pm	**EARTH**	Is the third planet

150,000

Big numbers catch your audience’s attention

FOUR COLUMNS

MARS

Mars is actually a very cold place

JUPITER

Jupiter is the biggest planet of them all

SATURN

It’s composed of hydrogen and helium

NEPTUNE

It’s the farthest planet from the Sun

A PICTURE ALWAYS REINFORCES THE CONCEPT

Images reveal large amounts of data, so remember: use an image instead of a long text. Your audience will appreciate it

CIVIL AVIATION AUTHORITIES

PERSONNEL LICENSING

Despite being red, Mars is a cold place

FLIGHT OPERATIONS

Saturn is a gas giant and has several rings

AIR TRAFFIC SERVICES

Jupiter is the biggest planet of them all

AERODROME SERVICE

Venus is the second planet from the Sun

DESKTOP SOFTWARE

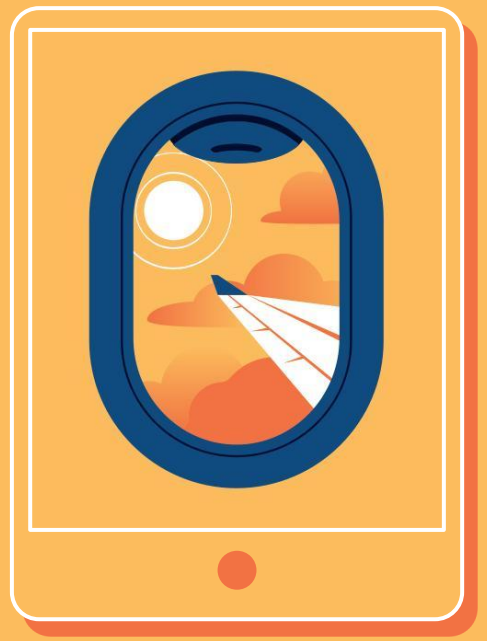
You can replace the image on the screen with your own work. Just right-click on it and select “Replace image”

WEBSITE FOR PHONES

You can replace the image on the screen with your own work. Just right-click on it and select “Replace image”

TABLET APP

You can replace the image on the screen with your own work. Just right-click on it and select “Replace image”




A PICTURE IS WORTH A THOUSAND WORDS

It's the closest planet to the Sun

MERCURY

Venus is the second planet from the Sun

VENUS

Mars is actually a very cold place

MARS

HOW TO CELEBRATE?

JUPITER

Jupiter is the biggest planet of them all

SATURN

It's composed of hydrogen and helium

NEPTUNE

It's the farthest planet from the Sun

AIR TRAFFIC

8%
Recreational flights

26%
Freight-only flights

66%
Passenger & freight flights

THIS IS A MAP

SATURN

It’s composed of hydrogen and helium

MARS

Mars is actually a very cold place

JUPITER

Jupiter is the biggest planet of them all

“This is a quote, words full of wisdom that someone important said and can make the reader get inspired.”

—SOMEONE FAMOUS

THIS IS A TABLE

	MERCURY	VENUS	EARTH	MARS
JUPITER	1.25	2.33	5.4	0.75
SATURN	721	274	640	88
NEPTUNE	121	726	853	500

7,809

Big numbers catch your audience’s attention

PURPOSES OF THE DAY

MARS

Despite being red, Mars is a cold place

SATURN

Saturn is a gas giant and has several rings

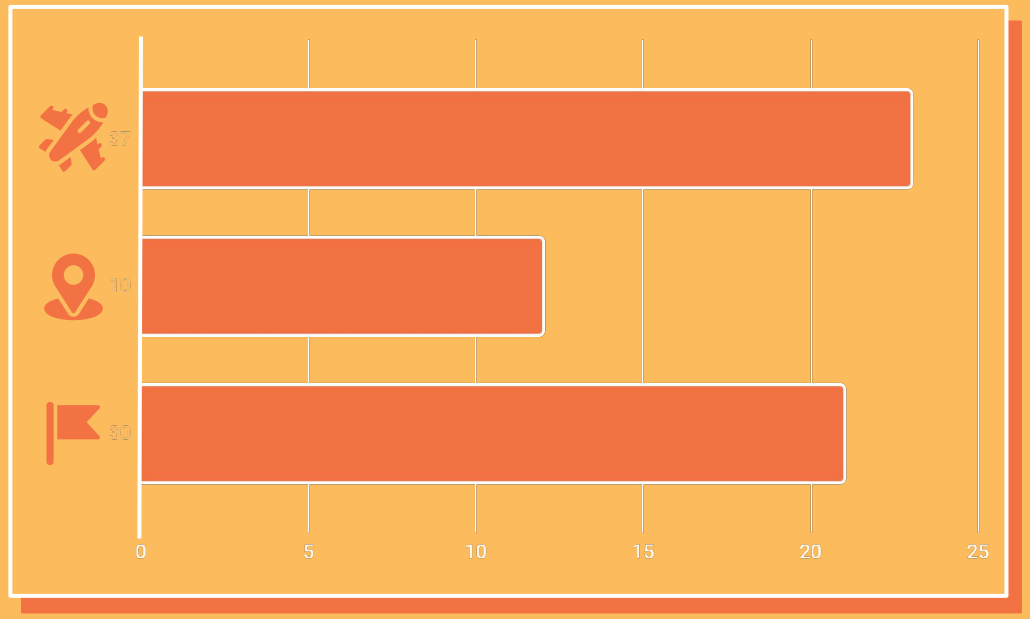
JUPITER

It's the biggest planet in the Solar System

VENUS

Venus is the second planet from the Sun

THIS IS A GRAPH



VENUS

Venus is the second planet from the Sun

JUPITER

Jupiter is the biggest planet of them all

SATURN

Saturn is a gas giant and has several rings

Follow the link in the graph to modify its data and then paste the new one here. For more info, click here


AWESOME
WORDS

SIGNIFICANCE

Saturn is a gas giant and has several rings. It's composed mostly of hydrogen and helium. It was named after the Roman god of wealth

HISTORY OF AVIATION

MERCURY

It’s the closest planet to the Sun

1993

MARS

Despite being red, Mars is a cold place

2004

JUPITER

It’s the biggest planet in the Solar System

2015

SATURN

Saturn is a gas giant and has several rings

2021

03

NAME OF THE SECTION

You can enter a subtitle here in case you need it

WHAT DO THEY SAY ABOUT US?

"Venus is the second planet from the Sun"

—BRADLEY DAWSON

"Despite being red, Mars is actually a cold place"

—MARIA LEBLANC

THIS IS A GRAPH

65%
Mercury is the smallest planet

35%
Venus has a beautiful name

Follow the link in the graph to modify its data and then paste the new one here. For more information, please click here

COMMERCIAL AND PRIVATE AVIATION

COMMERCIAL

Mercury is the closest planet to the Sun and the smallest one in the Solar System it's only a bit larger

PRIVATE

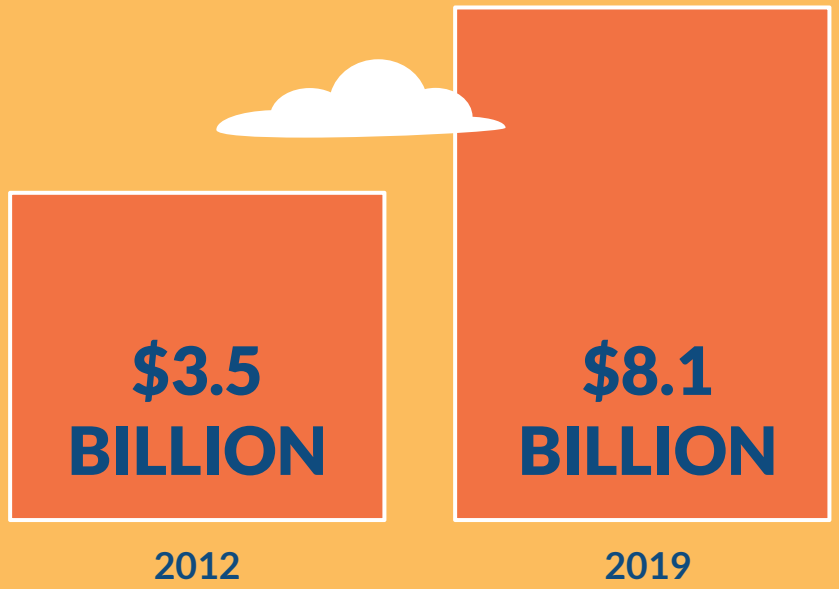
Venus has a beautiful name and is the second planet from the Sun. It's hot and has a poisonous atmosphere

04

NAME OF THE SECTION

You can enter a subtitle here in case you need it

ECONOMIC IMPACT



IMPACT INCLUDES:

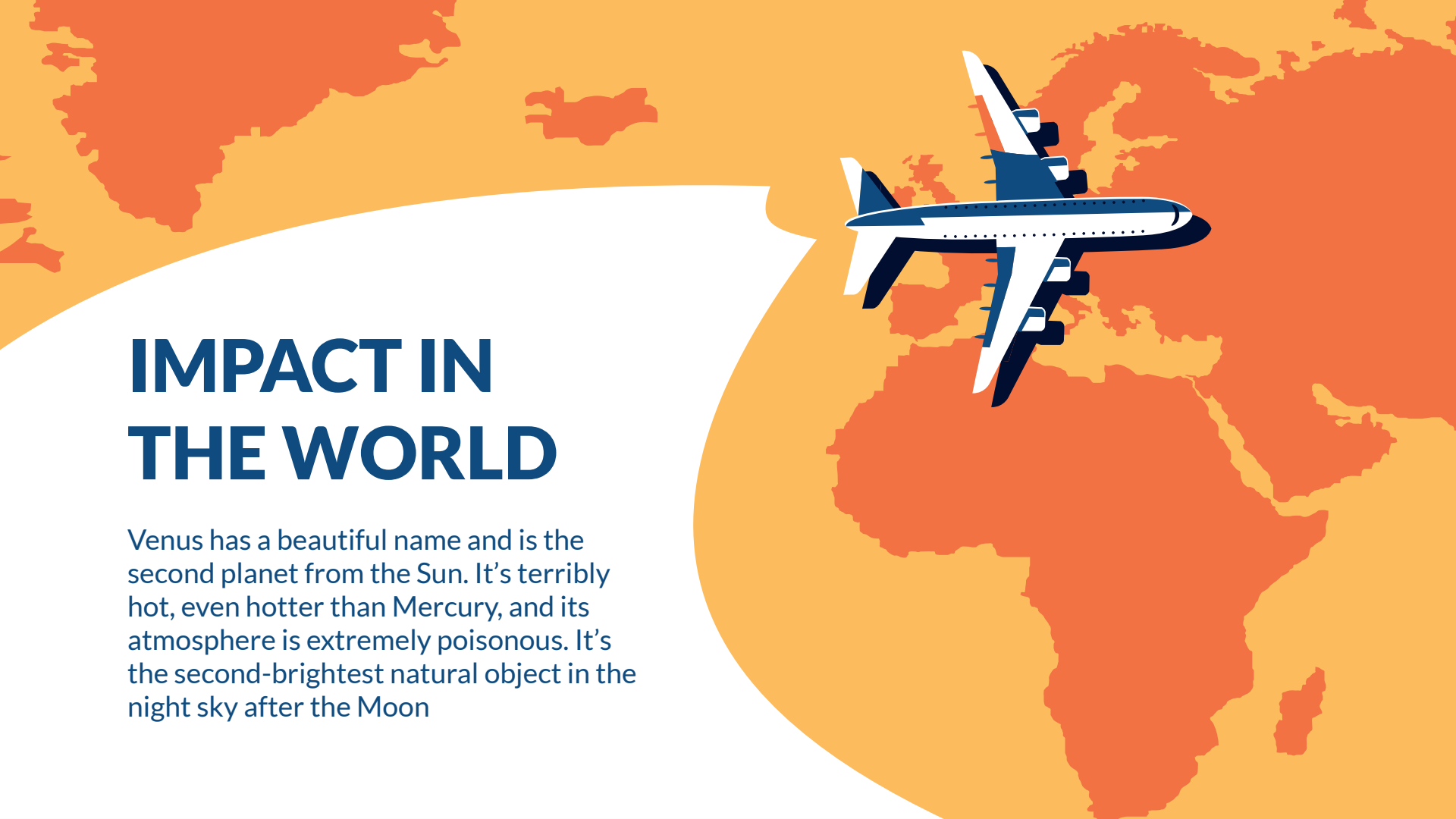
- Venus has a beautiful name, but it's terribly hot
- Mercury is the closest planet to the Sun
- Despite being red, Mars is a very cold place

THIS IS A CALENDAR

30	31	1	2	3	4	5
		MERCURY				
6	7	8	9	10	11	12
			VENUS			
30	13	14	15	16	17	18
						MARS
19	20	21	22	23	24	25
	JUPITER					
26	27	28	29	30	1	2
		SATURN				

IMPACT IN THE WORLD

Venus has a beautiful name and is the second planet from the Sun. It’s terribly hot, even hotter than Mercury, and its atmosphere is extremely poisonous. It’s the second-brightest natural object in the night sky after the Moon



PRINCIPAL AUTHORITIES

TIMMY JIMMY

You can speak a bit about this person here

JENNA DOE

You can speak a bit about this person here

ALBERT BONES

You can speak a bit about this person here

CONCLUSIONS

- Write the conclusions of your presentation here
- Write the conclusions of your presentation here
- Write the conclusions of your presentation here
- Write the conclusions of your presentation here
- Write the conclusions of your presentation here
- Write the conclusions of your presentation here
- Write the conclusions of your presentation here
- Write the conclusions of your presentation here

THANKS

Do you have any questions?
youremail@freepik.com
+91 620 421 88
yourcompany.com

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**

Please keep this slide for attribution

Sources

Sentiment Analysis of Twitter Data: A Survey of Techniques [1]

Emotion Detection using Natural Language Processing [2]

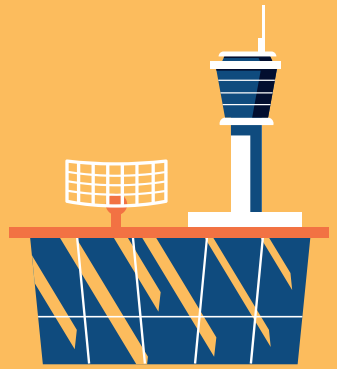
Sentiment Analysis of Political Tweets for Israel using Machine Learning [3]

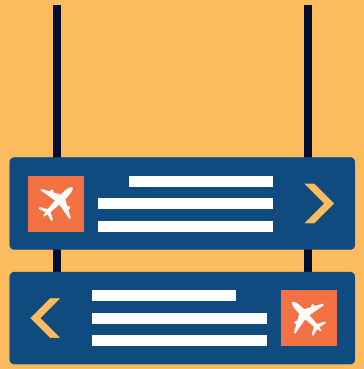
Working With RNNS [4]

Working With CNNS [5]

ALTERNATIVE RESOURCES

ALTERNATIVE RESOURCES

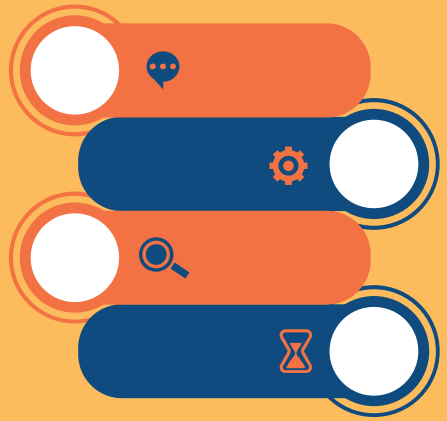




ALTERNATIVE RESOURCES

AVIATION ICON PACK

ALTERNATIVE RESOURCES



RESOURCES

Did you like the resources on this template? Get them for free at our other websites:

VECTORS

- Sky wallpaper for video conferencing
- Retro travel poster flat design
- Closed airport and rescheduling the vacation
- Travel landing page isometric style
- Sky - background for video conferencing
- Flat vintage promotional travel poster
- Travel banners with plane and train
- Travelling poster window of a plane

PHOTOS

- Close up hand holding small plane
- Hand holding small plane close up

PREMIUM RESOURCES

Did you like the resources on this template? Get them for free at our other websites:

VECTORS

- Flat design vintage travel label
- Flat vintage promotional travel poster
- Travel landing page I
- Travel landing page II
- Travel landing page III
- Flat travel background I
- Flat travel background II
- Flat travel background III
- Flat travel background IV
- Flat infographic element collection
- Blue sky background
- Travel banners with destination
- Travel background in flat style