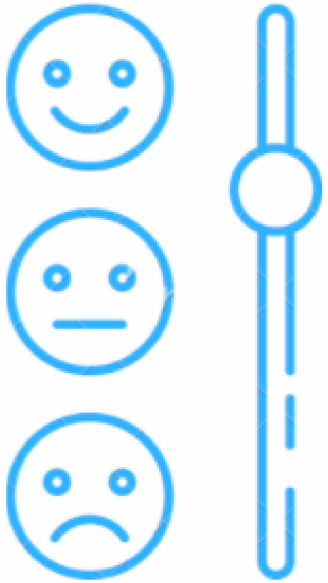
T5 DATASCIENCE BOOTCAMP 24 OCT 2021

# Twitter Sentiment analysis

## Natural language processing

Raghad Aloraini - Raghad Alawad

# Dataset



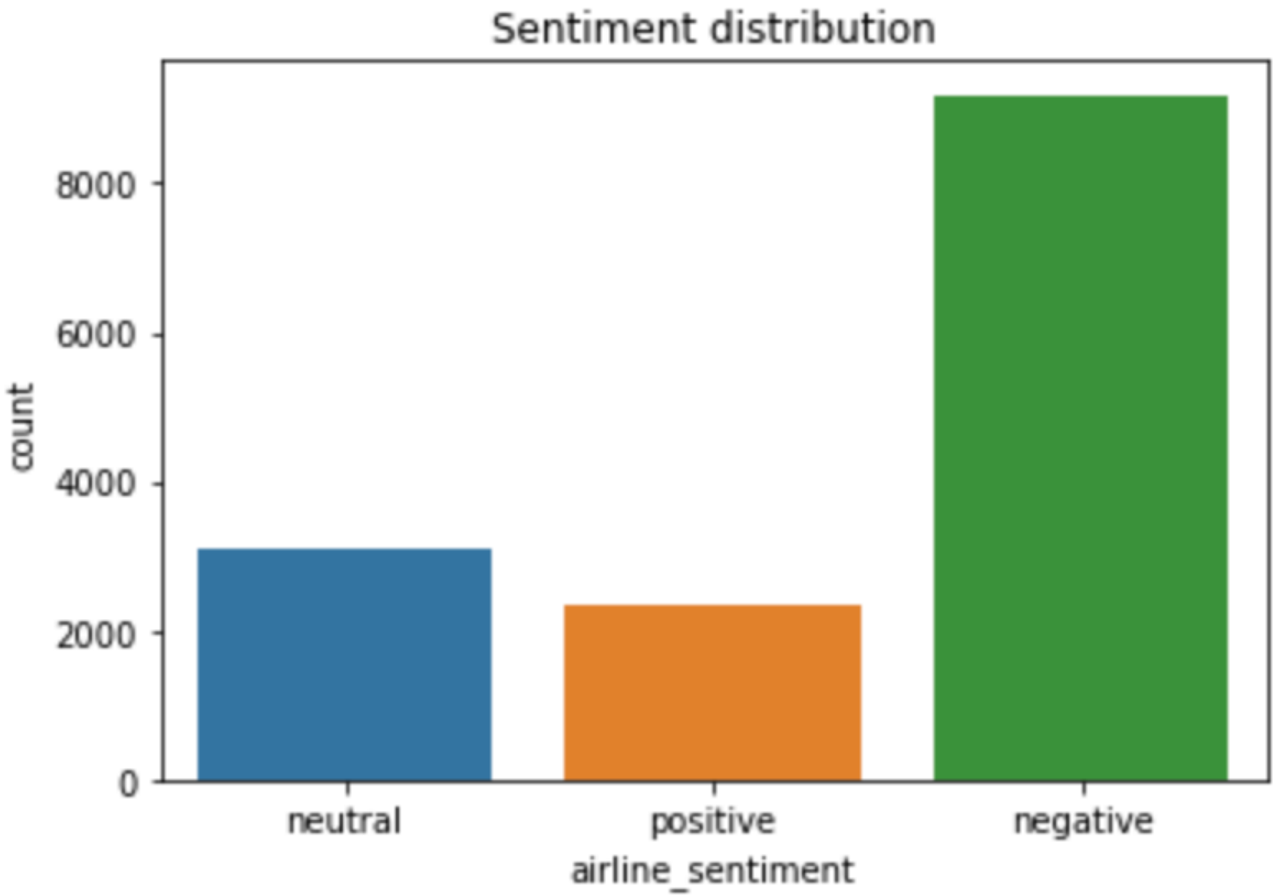
Our data set was taken from Kaggle website to analyze how travelers in 2015 expressed their feelings on Twitter

It contains whether the sentiment of the tweet is positive, neutral or negative for 6 US airlines.

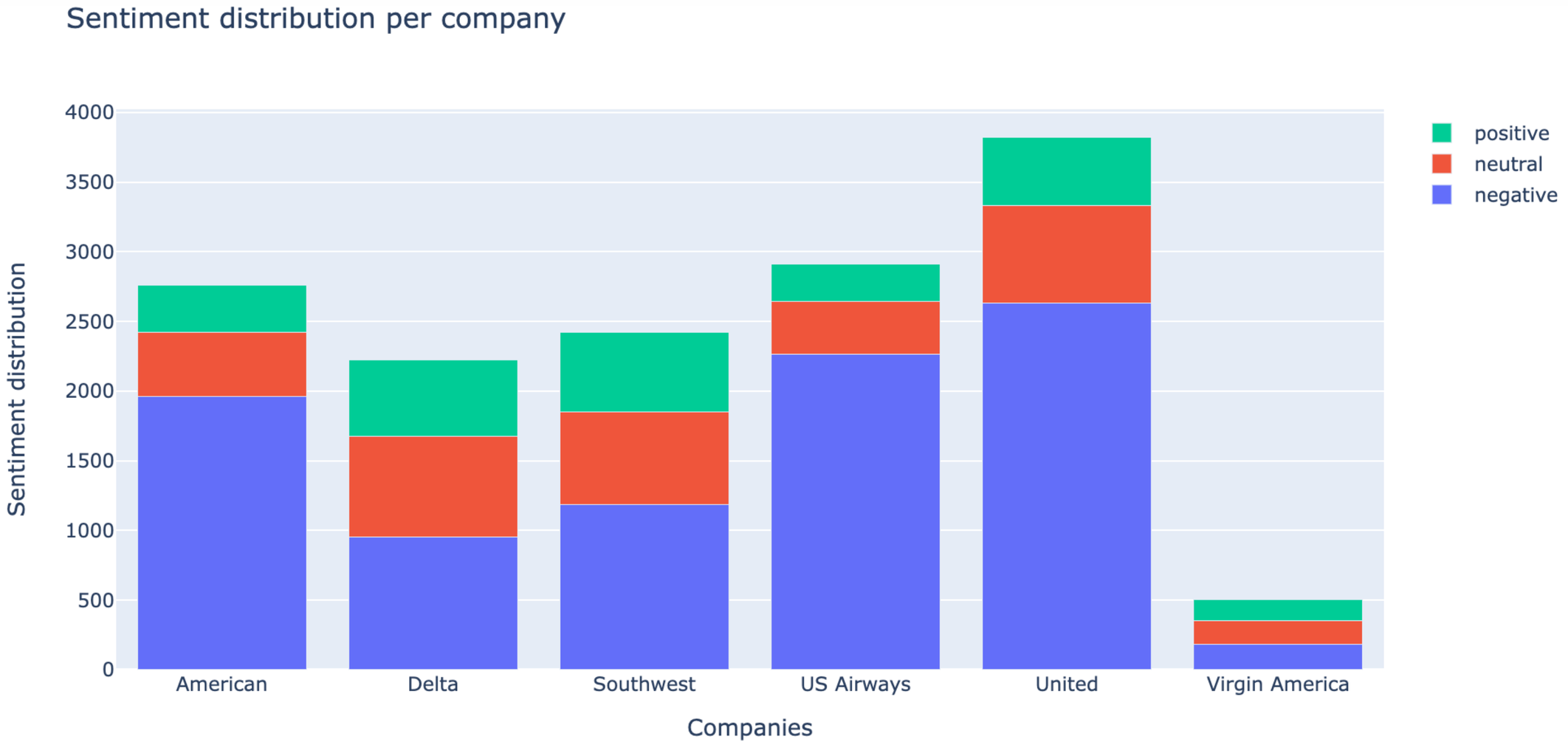
# Project stages:

1. EDA
2. Cleaning, Preprocessing
3. Baseline model
4. Parameter tuning
5. Models evaluation

# 1 Exploratory Data Analysis



# 1 Exploratory Data Analysis



# 2 Data Cleaning + Preprocessing

## Clean the tweet

Remove words with special characters.
Lower the words.

OR

## Process the tweet ★

Remove punctuations only.
Join words.
Lower the words.

```
Original tweet -> @VirginAmerica What @dhepburn said.

Cleaned tweet -> dhepburn said

Processed tweet -> virginamerica dhepburn said
```

3

# Baseline model

Support Vector Classifier

| Support Vector Classifier | |
|---|---|
| **Accuracy** | 0.90 |
| **Recall** | 0.79 |
| **Precision** | 0.89 |
| **F1-score** | 0.82 |

4

# Parameter tuning

Trying to improve the performance

## Cross Vallidation

Search for an optimal value

## Grid Search

Find the best parameters

5

# Models evaluation

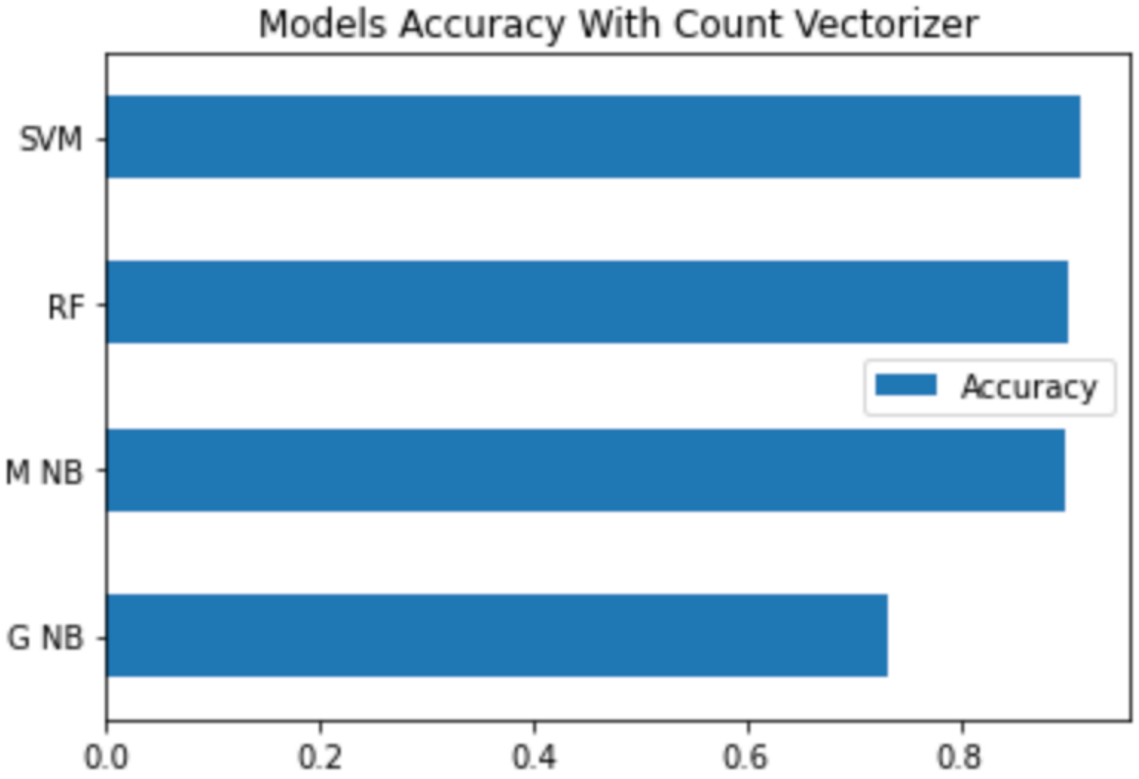
With count vectorizer and TF-IDF

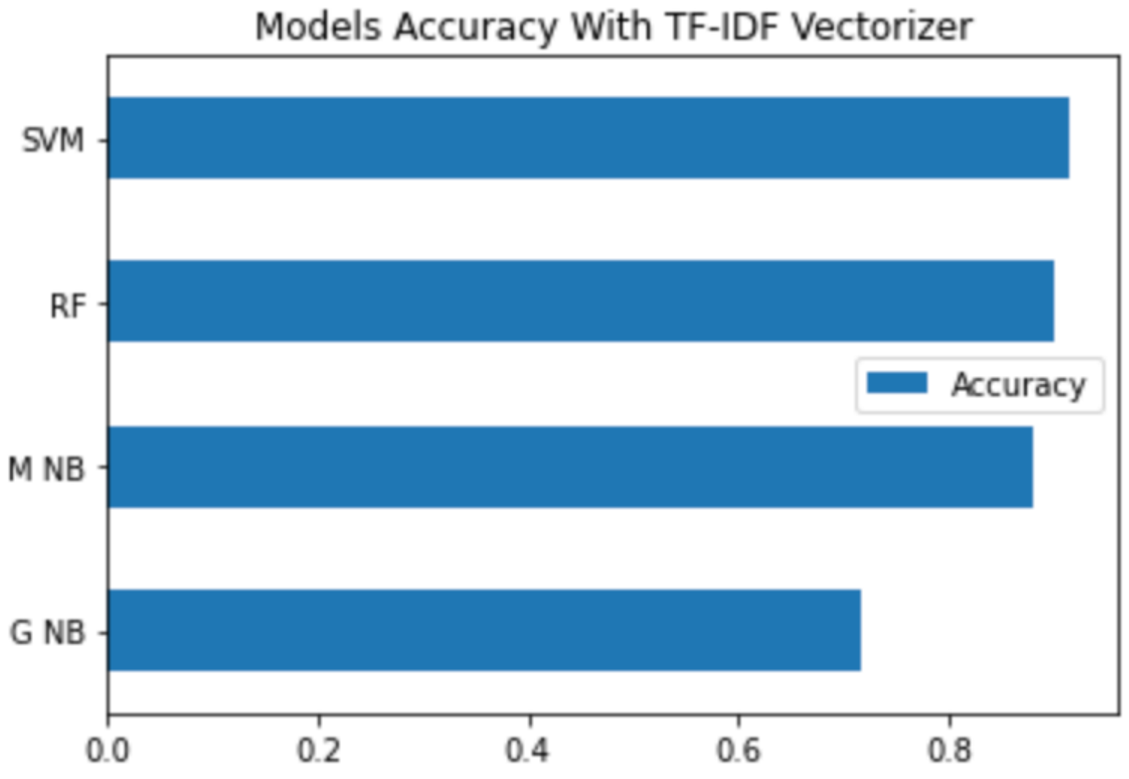
| Model | Accuracy with Count Vectorizer | Accuracy with TF-IDF Vectorizer |
|---|---|---|
| SVM | 0.913 | 0.915 |
| Multi-nominal NV | 0.89 | 0.87 |
| Gaussian NV | 0.73 | 0.71 |
| Random Forest | 0.902 | 0.901 |

5

# Models evaluation

Models Accuracy scores visualisation



# Conclusion

- The winner was SVM with 91% accuracy.
- SVM better with TF-IDF vectorizer.
- Multi-nominal NB, Gaussian NB and Random Forest better with count vectorizer.


# Future work

- Train unsupervised models
- Use more parameter tuning techniques

# Thank you!