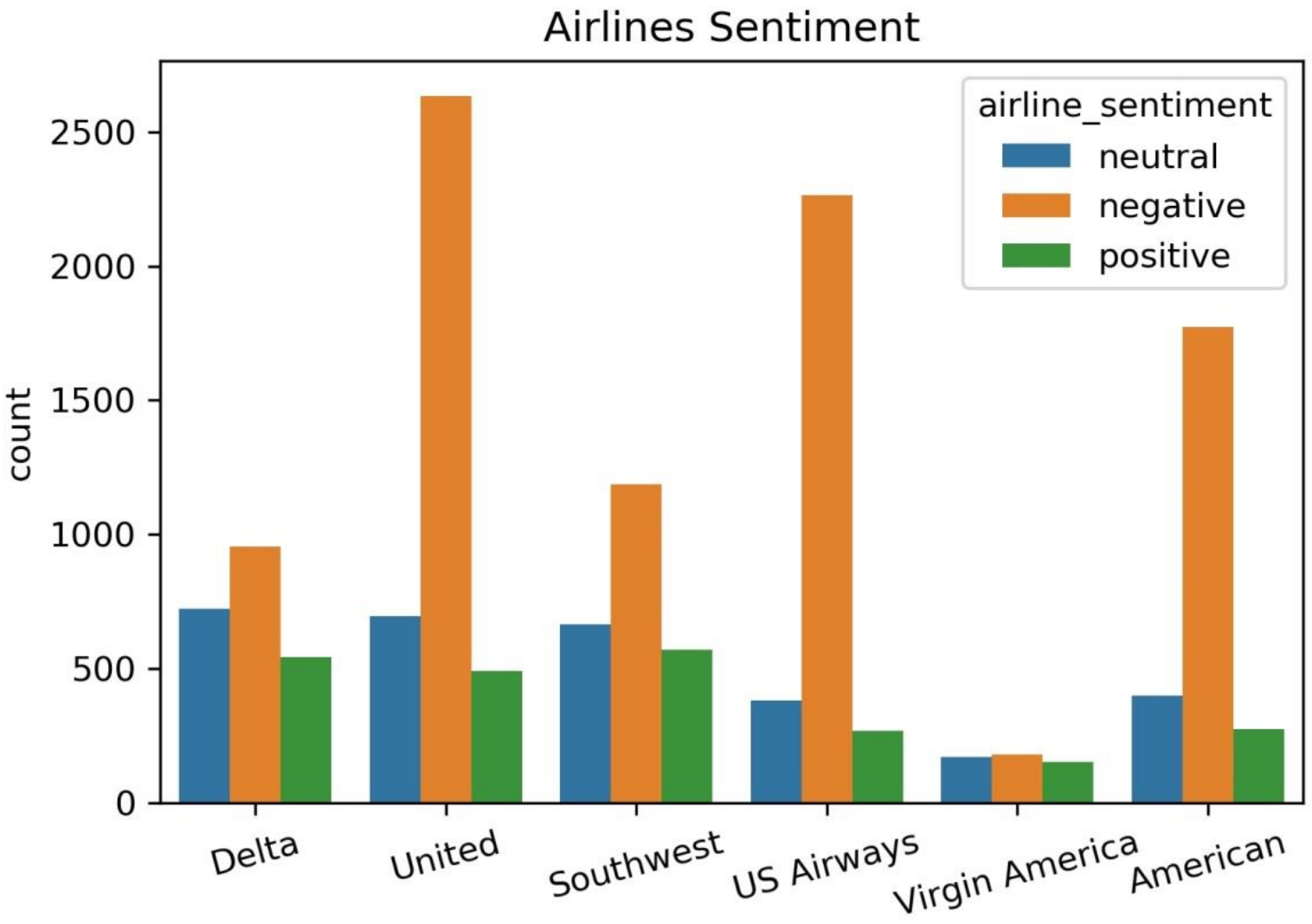
TWITTER US AIRLINE SENTIMENT ANALYSIS

BY WILLIAM DEW

DATA


Airlines Sentiment
airline_sentiment
neutral
negative
positive
count
2500
2000
1500
1000
500
0
Delta
United
Southwest
US Airways
Virgin America
American

DATA WRANGLING

Tweet cleaning steps

1. Convert text to lower-case
2. Remove URLs
3. Remove usernames
4. Remove punctuation
5. Remove stopwords

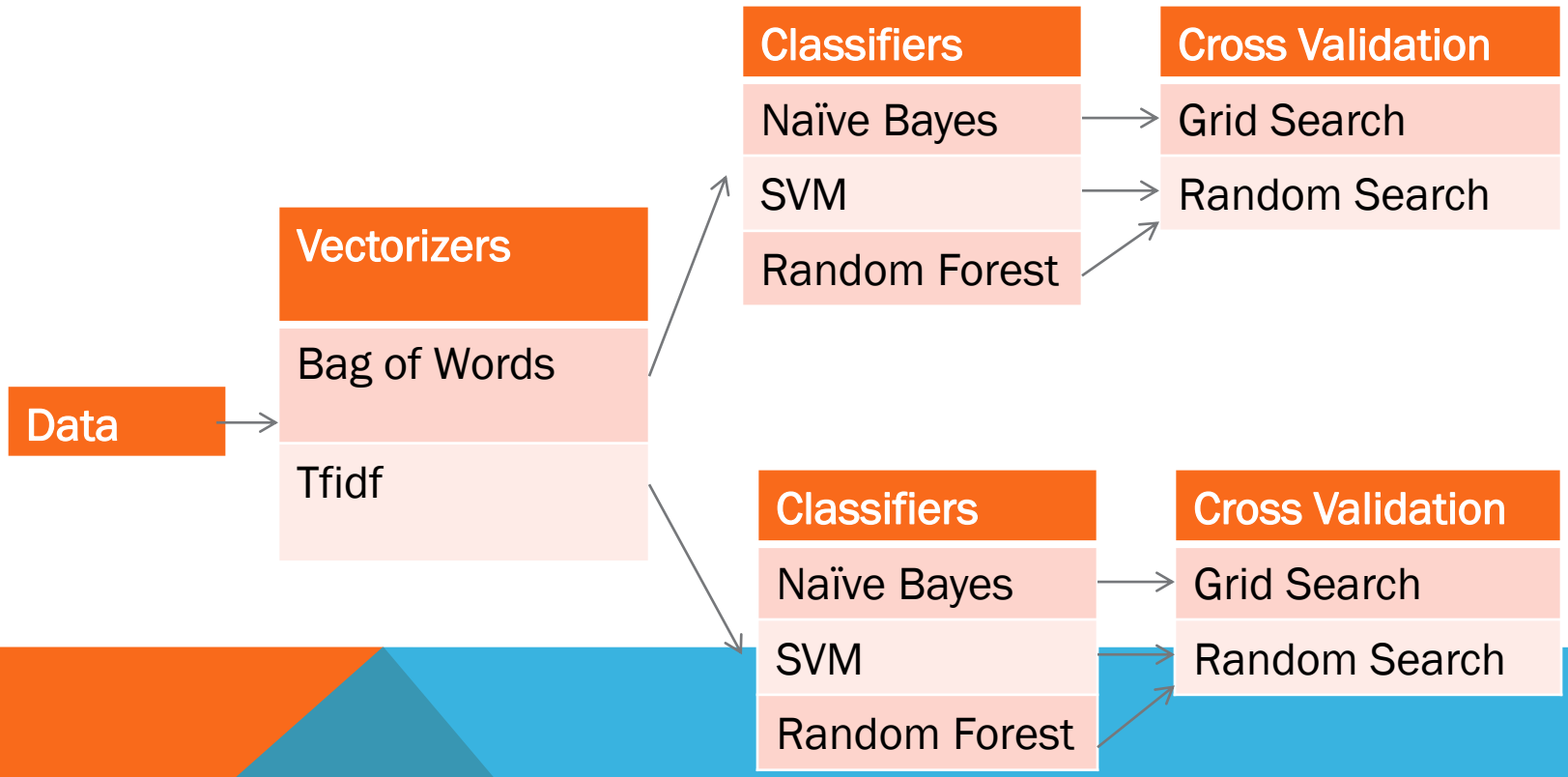
WORD BALLOONS

Negative

Neutral

Positive

MODELS


Data
Vectorizers
Bag of Words
Tfidf
Classifiers
Naïve Bayes
SVM
Random Forest
Cross Validation
Grid Search
Random Search
Classifiers
Naïve Bayes
SVM
Random Forest
Cross Validation
Grid Search
Random Search

CROSS VALIDATION

Vectorizer	Classifier	Training Set F1	Validation Set F1
Bag of Words	Naïve Bayes	0.969 (0.002)	0.634 (0.009)
Bag of Words	*SVM*	*0.968 (0.001)*	*0.707 (0.010)*
Bag of Words	Random Forest	0.991 (0.001)	0.701 (0.012)
Tfidf	Naïve Bayes	0.985 (0.001)	0.639 (0.009)
Tfidf	SVM	0.880 (0.001)	0.695 (0.005)
Tfidf	Random Forest	0.990 (0.001)	0.698 (0.011)

TEST SET RESULTS

Vectorizer	Classifier	Precision	Recall	F1
Bag of Words	SVM	0.731	0.699	0.713

- Test Set F1 score (0.713) is higher than Train (0.707) and Validation (0.707) F1 scores
- Precision and Recall given to show model performance

FURTHER INQUIRIES

1. Explore URLs and usernames
2. Validate more hyper parameters for Bag of Words and Tfidf
3. Look into oversampling and downsampling data
4. Build model with non-text features