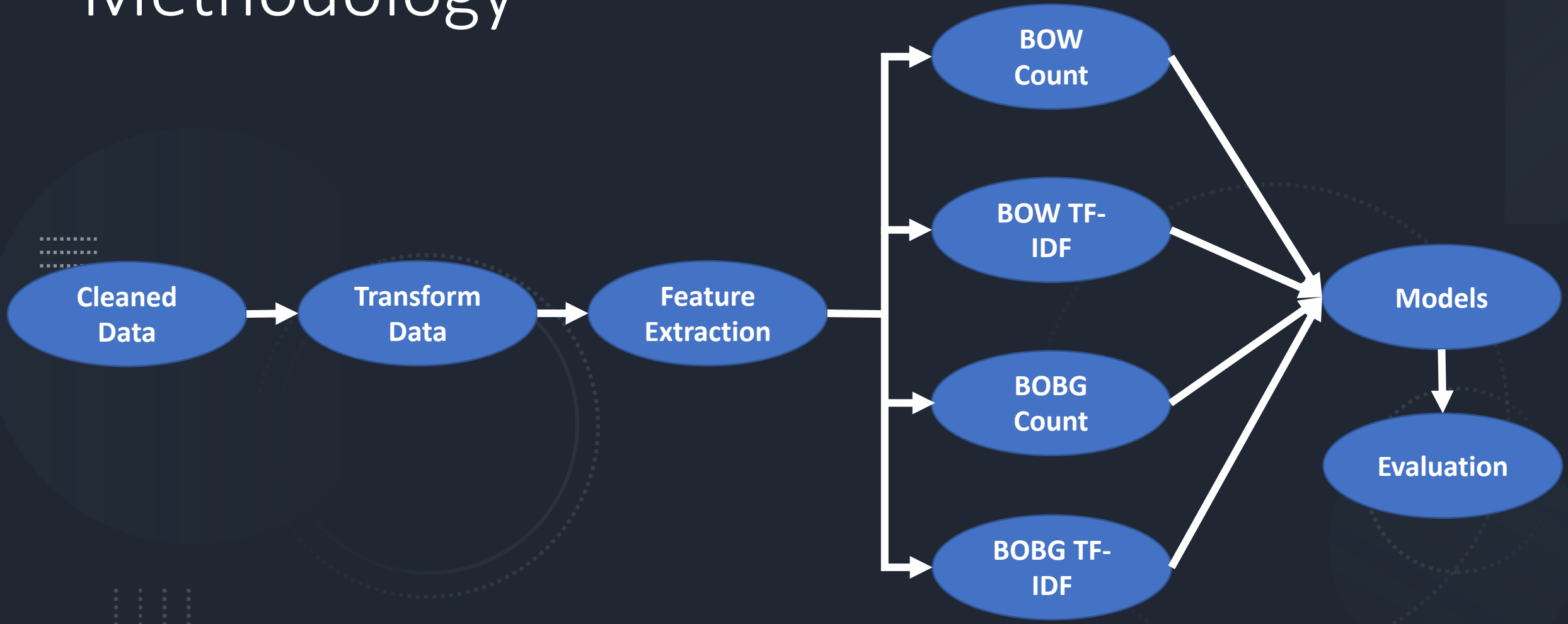
Milestone 2: Feature Engineering, Baseline Model & Interpretability

Albina Cako and Joshua Dalphy

Methodology



Data Transformation

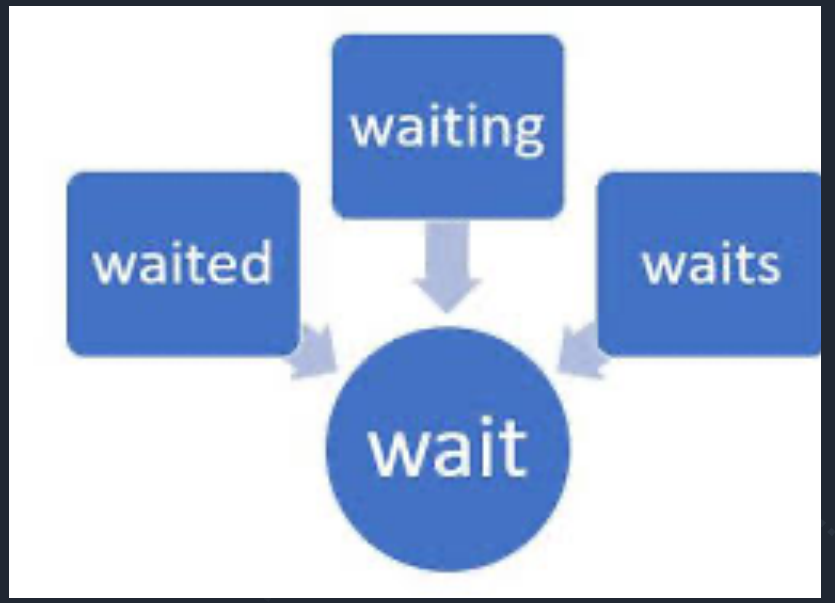
- Changed sentiment labels:
  - Positive sentiments = 1
  - Negative sentiments = 0
- Separated Sentiment & Reviews from dataframe
- Split data into train and test sets (70/30)

https://medium.com/@datamonsters/sentiment-analysis-tools-overview-part-1-positive-and-negative-words-databases-ae35431a470c

	Reviews	Sentiments
2932	not understand comment focus mcconaughey never...	1
5537	let us say simple word even maker film may cha...	0
1414	year lose gorgeous jane parker maureen osulliv...	1

Data Transformation

- The movie reviews were stemmed
- Digits (0-9) were removed as well



https://medium.com/@datamonsters/sentiment-analysis-tools-overview-part-1-positive-and-negative-words-databases-ae35431a470c

```
 'ye southern star featur pretti forgett titl tune sing heavi set crooner
matt monro pretti much establish tone bloat rather dull featur stunningli
miscast georg segal ursula andress adventur coupl search larg diamond add
harri andrew strang accent no less chase ostrich ton stock footag wildlif
poorli compos dull photographi raoul coutard end thoroughli unexcit romp
jungl seneg ',
```

Feature Extraction & Baseline Modelling
1) Bag of Words - Count

- Performed BOW on training/testing sets
- Created 24959 features

	bernard	bernhard	berni	bernic	bernier	bernsen	bernstein	berri	berrisford	berryman
0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0

Logistic Regression Performance

```
Model Performance metrics:
------------------------------
Accuracy: 0.8743
Precision: 0.8747
Recall: 0.8743
F1 Score: 0.8743
```

```
Prediction Confusion Matrix:
------------------------------
           Predicted:
                     1     0
Actual: 1          922   116
        0          148   914
```

Feature Extraction & Baseline Modelling
2) Bag of Words – TF-IDF

- Performed TF-IDF on BOW 24959 features

	bernard	bernhard	berni	bernic	bernier	bernsen	bernstein	berri	berrisford	berryman
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Logistic Regression Performance

```
Model Performance metrics:
------------------------------
Accuracy: 0.8833
Precision: 0.8839
Recall: 0.8833
F1 Score: 0.8833
```

```
Prediction Confusion Matrix:
------------------------------
           Predicted:
                    1     0
Actual: 1         935   103
        0         142   920
```

Feature Extraction & Baseline Modelling
3) Bag of Bi-Grams - Count

- Performed BOBG on training/testing sets
- Created 385 586 features

	acid lay	acid morph	acid movi	acid mr	acid mushroom	acid never	acid poptart	acid quickli	acid seem	acid start
0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0

Logistic Regression Performance

```
Model Performance metrics:
------------------------------
Accuracy: 0.8295
Precision: 0.833
Recall: 0.8295
F1 Score: 0.8292
```

```
Prediction Confusion Matrix:
------------------------------
           Predicted:
                      1     0
Actual: 1           911   127
        0           231   831
```

Feature Extraction & Baseline Modelling
4) Bag of Bi-Grams – TF-IDF

- Performed TF-IDF on BOBG 385 586 features

	acid lay	acid morph	acid movi	acid mr	acid mushroom	acid never	acid poptart	acid quickli	acid seem
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Logistic Regression Performance

```
Model Performance metrics:
------------------------------
Accuracy: 0.8295
Precision: 0.833
Recall: 0.8295
F1 Score: 0.8292
```

```
Prediction Confusion Matrix:
------------------------------
           Predicted:
                     1     0
Actual: 1          911   127
        0          231   831
```

Model Interpretation

1) Bag of Words - Count vs. TF-IDF

Summary of Results:

Extraction	Accuracy	False Positives	False Negatives
Count	0.8743	148	116
TF-IDF	0.8833	142	103

- Model produced using TF-IDF had better accuracy (approx. 1%)
- The TF-IDF model has a lower false positive and false negative rate then the count method.
- Both models have higher false positive rates than false negative rates.
- Runtime Performance:
  - Count: 851 ms
  - TF-IDF: 293 ms

The TF-IDF trained model outperformed the count model

Model Interpretation

1) Bag of Bi-Grams Count vs. TF-IDF

Summary of Results:

Extraction	Accuracy	False Positives	False Negatives
Count	0.8295	231	127
TF-IDF	0.8295	231	127

- The results were the same for both the count and TF-IDF models.
- Applying TF-IDF did not make a difference in terms of accuracy or false negatives/positives
- Runtime Performance:
  - Count: 3.08 s
  - TF-IDF: 2.61 s

Given that both models yielded same statistical results, based on runtime the TF-IDF trained model outperformed the count model

Feature Importance Using Random Forest

Bag of Words

Count

bad	0.022325
wast	0.009658
great	0.009370
not	0.007198
love	0.006977
beauti	0.005598
bore	0.005547
excel	0.005399
aw	0.004676
movi	0.004617

TF-IDF

bad	0.022325
wast	0.009658
great	0.009370
not	0.007198
love	0.006977
beauti	0.005598
bore	0.005547
excel	0.005399
aw	0.004676
movi	0.004617

Bag of Bi-Grams

Count

bad movi	0.009489
not even	0.006314
wast time	0.005406
one bad	0.004259
bad film	0.003904
bad act	0.003404
not wast	0.003240
look like	0.002857
movi bad	0.002468
must see	0.002430

TF-IDF

bad movi	0.009668
not even	0.006154
wast time	0.006069
bad film	0.005183
one bad	0.004875
not wast	0.004077
look like	0.003653
movi bad	0.003521
bad act	0.003065
highli recommend	0.002675

Feature Importance Using Random Forest

Bag of Words - Count

Number of unimportant features:

12955 of total 24959

```
['aavjo',
 'gangstermovi',
 'aaja',
 'gam',
 'frutti',
 'abet',
 'entwistl',
 'gard',
 'frenchwoman',
 'enrol']
```

Bag of Bi-Grams – TF-IDF

Number of unimportant features:

300998 of total 385586

```
['franc dalen',
 'armi get',
 'franci matthew',
 'art master',
 'fortun get',
 'franci may',
 'franci drake',
 'fulci bava',
 'armi garrison',
 'frasier make']
```

Conclusions & Future Work

Conclusions:

- 4 models were produced using BOW (Count and TF-IDF) and BOBG (Count and TF-IDF) for feature extraction.
- Logistic regression was chosen for baseline modelling.
- Model performance was evaluated based on accuracy and runtime.
- Random forest was used assess feature importance.
- **The best performing model was using the features extracted through Bag of Words - TF-IDF.**

Future Work:

- Hyperparameter tuning on the baseline model obtained using BOW TF-IDF.
- Model evaluation and comparison.