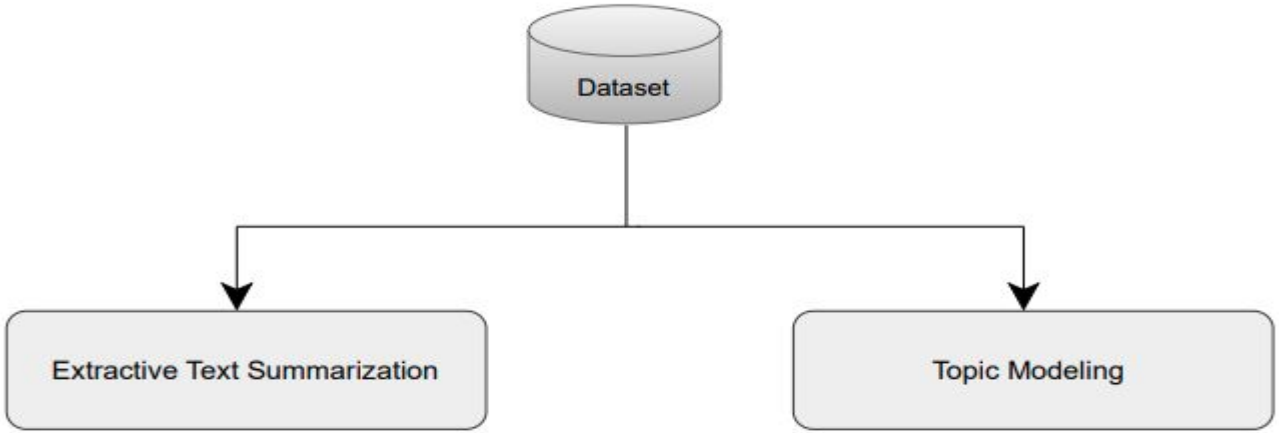
# Extractive Text Summarizer with Topic Identification

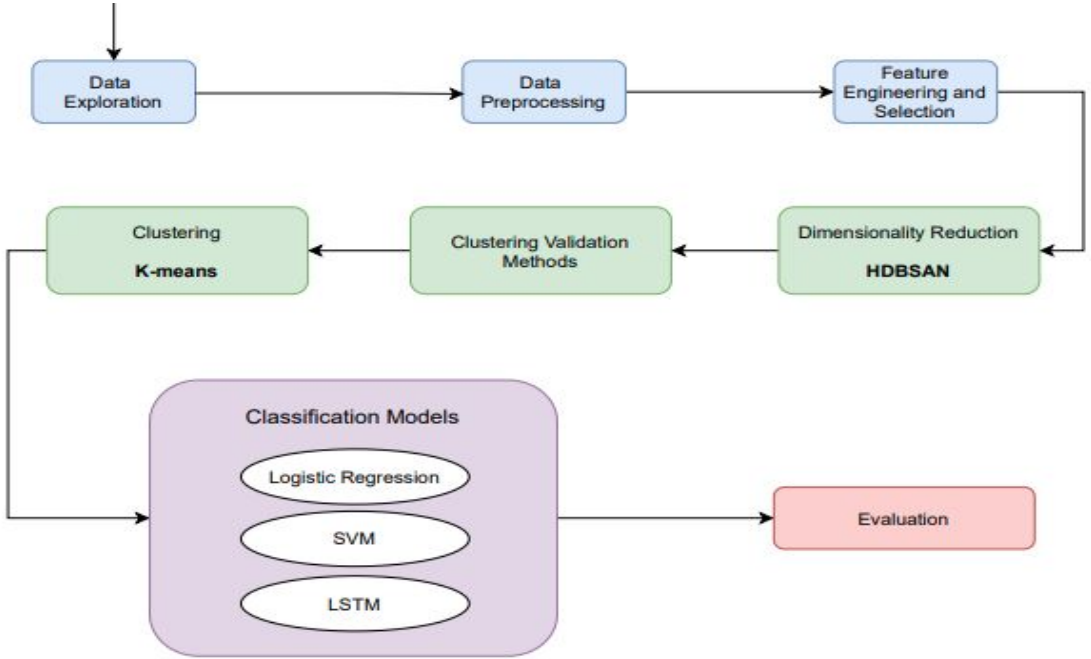
# Overview



# Overview


Text
Sentence 1
Sentence 2
Sentence 3
Sentence 4
Sentence 5
Extractive Summariser
Summary
Sentence 2
Sentence 4
Text
Sentence 1
Sentence 2
Sentence 3
Sentence 4
Sentence 5
Abstractive Summariser
Summary
New Sentences

# Text Summarization



# Dataset

- Cornell University's Newsroom database.
- 1.3 million articles and summaries.
- Each summary is labelled to be abstractive, extractive or mixed.
- 10,000 records from the dataset is used.

| | mean |
|---|---|
| **Article** | 33.195 |
| **Summary** | 3.268 |

| | % Breakdown | Cumulative |
|---|---|---|
| **nytimes** | 0.2383 | 0.2383 |
| **bostonglobe** | 0.1356 | 0.3739 |
| **nydailynews** | 0.0837 | 0.4576 |
| **cnbc** | 0.0828 | 0.5404 |
| **p://fortune** | 0.0512 | 0.5916 |
| **9news** | 0.0510 | 0.6426 |
| **sfgate** | 0.0452 | 0.6878 |
| **tmz** | 0.0426 | 0.7304 |
| **p://nypost** | 0.0386 | 0.7690 |
| **people** | 0.0364 | 0.8054 |

# Methodology

1. Data Preprocessing
2. Article and Summary Sentence Bert Embeddings
3. Document Embeddings
4. Feature Engineering :  Sentence Number - Document Length - Summary Length - Total words count
5. Feature Selection

# Methodology

6 Labels using Cosine Similarity

```
cos_sim_mat
```

```
array([[0.99999976, 0.50047994, 0.5196508 , 0.59983295],
       [0.6196736 , 0.7722818 , 0.7966124 , 0.5275056 ],
       [0.599833  , 0.29468644, 0.52074134, 0.99999976],
       [0.56438607, 0.33038598, 0.55166763, 0.6393672 ],
       [0.4886982 , 0.45817882, 0.51193637, 0.624065  ],
       [0.46047905, 0.35522765, 0.4688924 , 0.48108268],
       [0.28927967, 0.3464931 , 0.5418632 , 0.40989965],
       [0.5305566 , 0.32012153, 0.54381216, 0.5782098 ],
       [0.46696132, 0.32249698, 0.70351344, 0.66038245],
       [0.2646217 , 0.31237996, 0.51453644, 0.3949415 ],
       [0.61532736, 0.36732113, 0.54515684, 0.77773935],
       [0.34573337, 0.33674398, 0.3050996 , 0.35545546],
       [0.48274225, 0.2860468 , 0.5660532 , 0.7047001 ],
       [0.63601404, 0.3220192 , 0.6802253 , 0.57318306],
       [0.27251178, 0.37319538, 0.42750496, 0.35306516]], dtype=float32)
```

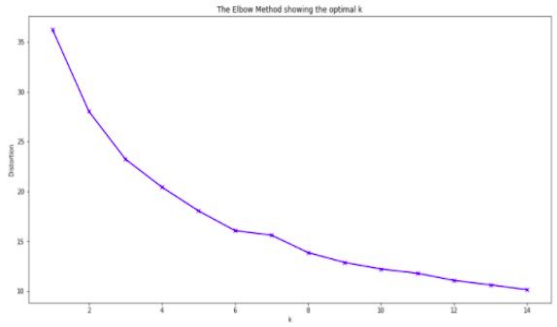
```
idx_arr = np.argmax(cos_sim_mat, axis=0)
idx_arr
```

```
array([0, 1, 1, 2], dtype=int64)
```

# Methodology

7 Dimensionality Reduction: UMAP

8 Clustering Validation Methods: Elbow Method - Silhouette Method




# Methodology

9 Clustering Sentences: K-means

10 Logistic Regression

- Default class weights
- Balanced class weights

11 Support Vector Machine

- Gaussian Kernel

12 Long Short Term Memory

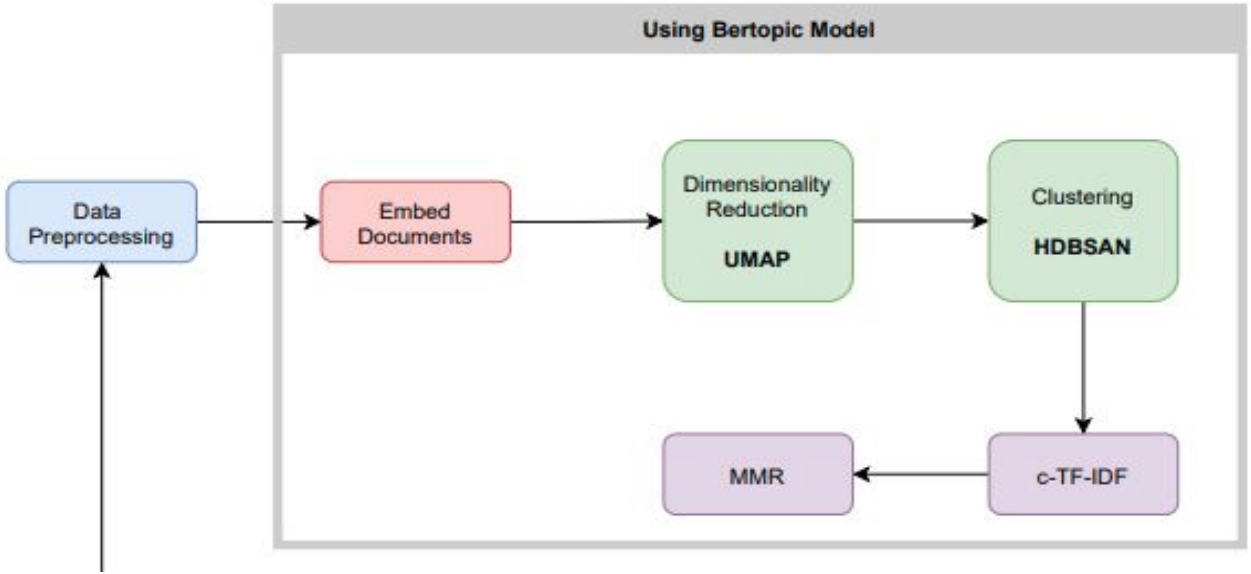
- Unidirectional network with 25 neurons
- Bidirectional network with 25 neurons
- Unidirectional network with 50 neurons
- Bidirectional network with 50 neurons

13 Evaluation

- Rouge-1 : Recall, Precision, F1 score
- Rouge-L : Recall, Precision, F1 score

14 Lead-3 Baseline

# Topic Modeling



# Results - Logistic Regression

K-means Clustering where k = 3

| ROUGE-1 | F1 | Recall | Precision |
|---|---|---|---|
| Default | 0.560 | 0.540 | 0.755 |
| Balanced | 0.544 | 0.520 | 0.744 |

| ROUGE-L | F1 | Recall | Precision |
|---|---|---|---|
| Default | 0.541 | 0.521 | 0.727 |
| Balanced | 0.522 | 0.500 | 0.711 |

K-means Clustering where k = 4

| ROUGE-1 | F1 | Recall | Precision |
|---|---|---|---|
| Default | 0.561 | 0.543 | 0.756 |
| Balanced | 0.544 | 0.521 | 0.743 |

| ROUGE-L | F1 | Recall | Precision |
|---|---|---|---|
| Default | 0.542 | 0.524 | 0.728 |
| Balanced | 0.523 | 0.501 | 0.711 |

# Results - SVM

K-means Clustering where k = 3 & k = 4 had exactly the same accuracies

| ROUGE-1 | F1 | Recall | Precision |
|---|---|---|---|
| rbf | 0.568 | 0.558 | 0.749 |

| ROUGE-L | F1 | Recall | Precision |
|---|---|---|---|
| rbf | 0.552 | 0.541 | 0.725 |

# Results - LSTM

K-means Clustering where k = 3

| ROUGE-1 | F1 | Recall | Precision |
|---|---|---|---|
| Uni 25 | 0.580 | 0.560 | 0.766 |
| Uni 50 | 0.601 | 0.577 | 0.765 |
| Bi 25 | 0.584 | 0.578 | 0.751 |
| Bi 50 | 0.590 | 0.579 | 0.765 |

| ROUGE-L | F1 | Recall | Precision |
|---|---|---|---|
| Uni 25 | 0.561 | 0.541 | 0.739 |
| Uni 50 | 0.572 | 0.559 | 0.739 |
| Bi 25 | 0.566 | 0.560 | 0.725 |
| Bi 50 | 0.572 | 0.561 | 0.739 |

K-means Clustering where k = 4

| ROUGE-1 | F1 | Recall | Precision |
|---|---|---|---|
| Uni 25 | 0.582 | 0.574 | 0.754 |
| Uni 50 | 0.590 | 0.582 | 0.760 |
| Bi 25 | 0.586 | 0.574 | 0.760 |
| Bi 50 | 0.588 | 0.582 | 0.761 |

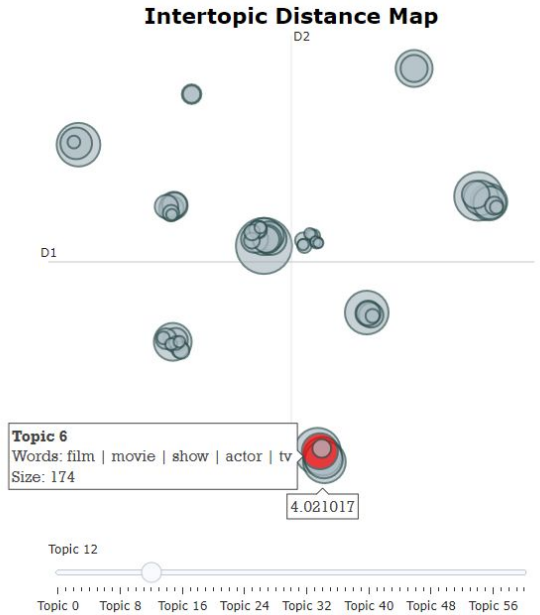
| ROUGE-L | F1 | Recall | Precision |
|---|---|---|---|
| Uni 25 | 0.564 | 0.555 | 0.727 |
| Uni 50 | 0.572 | 0.563 | 0.734 |
| Bi 25 | 0.568 | 0.555 | 0.734 |
| Bi 50 | 0.570 | 0.564 | 0.734 |

# Results

| ROUGE-1 | F1 | Recall | Precision |
|---|---|---|---|
| Lead-3 | 56.1% | 58.4% | 70.6% |
| LR | 56.1% | 54.3% | 75.6% |
| SVM | 56.8% | 55.8% | 74.9% |
| LSTM | 60.1% | 57.7% | 76.5% |

| ROUGE-L | F1 | Recall | Precision |
|---|---|---|---|
| Lead-3 | 54.5% | 56.6% | 68.4% |
| LR | 54.2% | 52.4% | 72.8% |
| SVM | 55.2% | 54.1% | 72.5% |
| LSTM | 57.2% | 55.9% | 73.9% |

# Results


Intertopic Distance Map
D2
D1
Topic 6
Words: film | movie | show | actor | tv
Size: 174
4.021017
Topic 12
Topic 0
Topic 8
Topic 16
Topic 24
Topic 32
Topic 40
Topic 48
Topic 56


Topic 0
republican
candidate
presidential
campaign
clinton
Topic 1
bank
fed
stock
loan
yield
Topic 2
food
restaurant
sauce
cheese
dish
Topic 3
49ers
nfl
kaepernick
quarterback
football
Topic 4
judge
court
attorney
lawyer
trial
Topic 5
art
building
artist
painting
museum
Topic 6
film
movie
show
actor
tv
0
0.02
0.04
Topic 7
book
novel
author
writer
fiction
0
0.01
0.02
Topic 8
child
family
parent
kid
mother
0
0.005
0.01
0.015
0.02

# Results

| Summary | Tags |
| --- | --- |
| All day, every day, Cheryl Bernstein thanks her 16-month-old son. "I gave life to Reid, but he gave me life - a reason to get clean and go on,"she said yesterday after graduating from the Manhattan Family Treatment Court program. Bernstein, 41, and her husband, Doug Flaumenbaum, 33, both recovering crack and heroin addicts, were among three dozen men and women who regained custody of their children. | [ 'drug', 'cancer', ‘medicine', 'addiction', 'pharmaceutical', 'dr', 'biotech', 'technology', 'test', 'scientist' ] |

# Thank you.