UNIVERSITY OF TARTU
Institute of Computer Science

IntelliBrief News Summarization

MD Murad | Zeshan | Sulaiman | Syed Fakhar

C46989 C47207 C47046 C310761

Introduction

INTERNATIONAL
THE NEWS
Sunday
October 30, 2022
London Edition

Man remanded over lorry deaths of 39 Vietnamese nationals in Essex

Govt rules out talks, as PTI long march moves towards capital

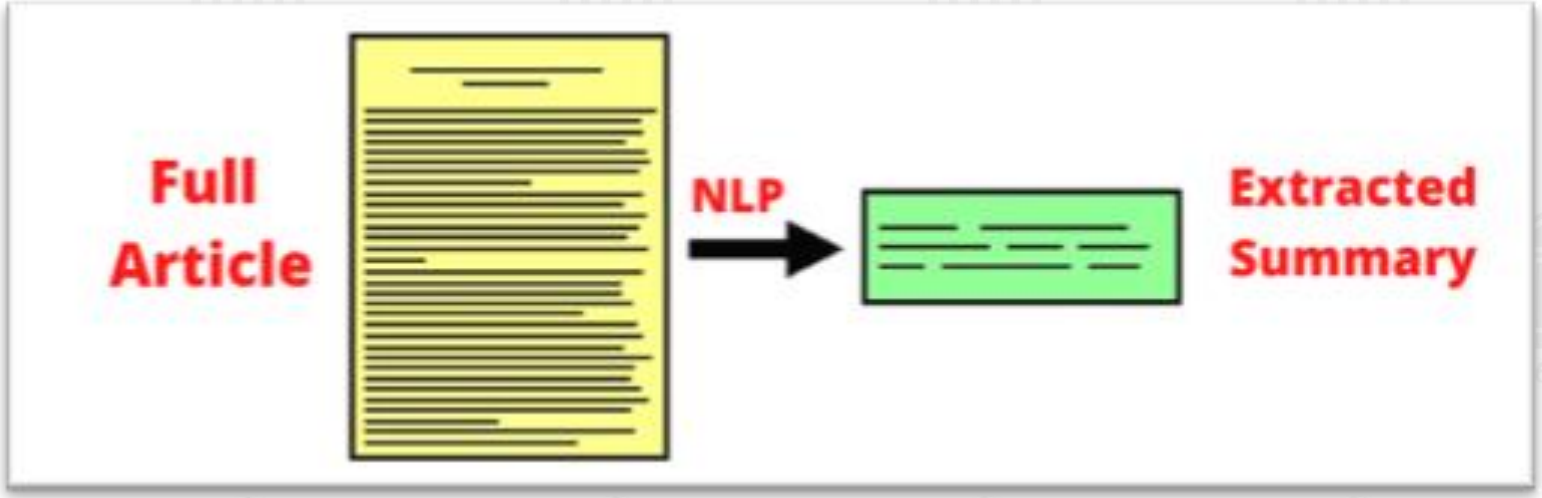
Marriyum says Imran Khan does not want transparent elections, but 'chaos and bloodshed' in the country; ex-PM Khan asks CJP to protect fundamental rights; PTI's long march on its second day reaches Ferozwala on its way to Islamabad, to resume from Muridke this morning

Thousands join March of the Mummies to demand childcare reforms

Octopus Energy takes over collapsed rival Bulb

Octopus announces it's taking on Bulb's 1.5m customers, 'bringing an end to taxpayer losses and uncertainty for Bulb customers and employees'

ML
modularization



Literature Review

A variety of techniques are being used for extractive summarization

- Feature-based models
- Deep Learners
- Unsupervised learning models

Feature Based

Features like word frequency, time words, place words, similarity to title sentence etc. were extracted.

Summary sentences were chosen on the basis of a feature-calculated score.

Deep Learning

Words and sentences are converted into vector form using some word embedding model. They are then fed to a deep learning model.

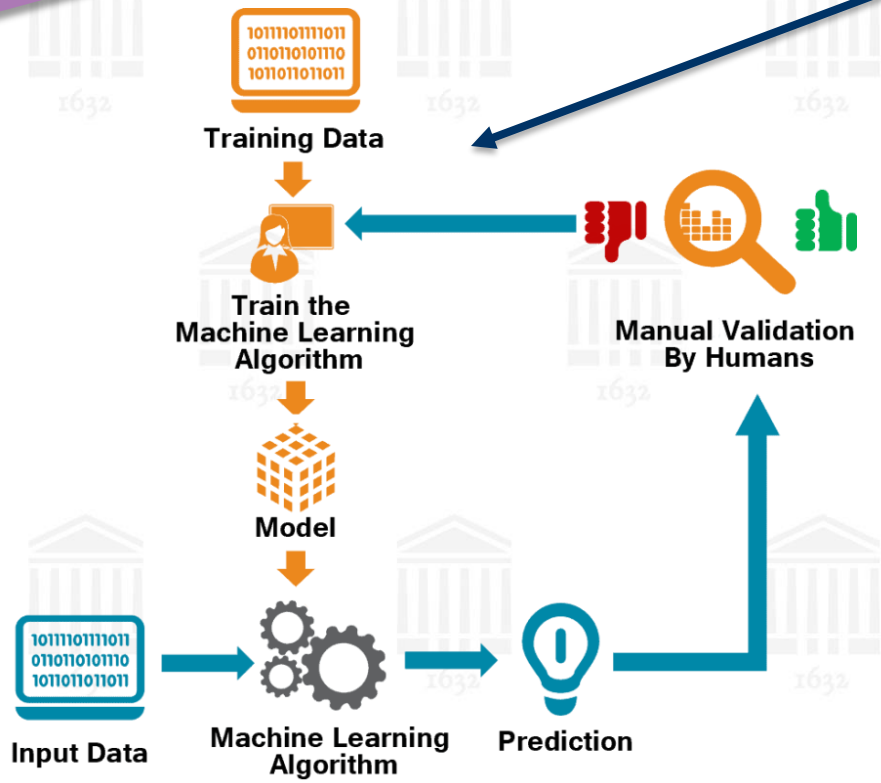
The model predicts yes or no for choosing a sentence for the summary.

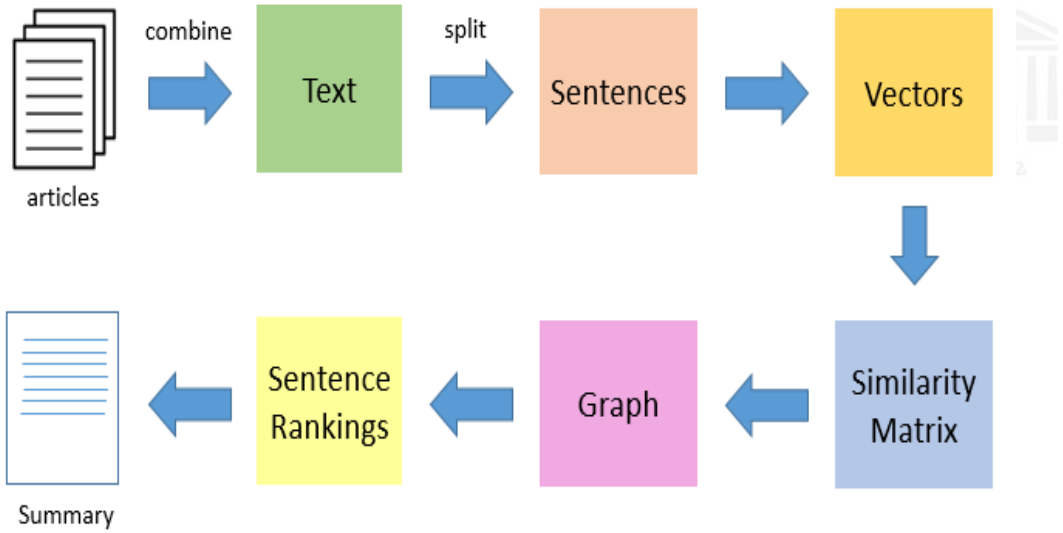
Unsupervised

Sentences are clustered according to some metric like similarity with each other, with title, etc.

N sentences are then chosen from each cluster and displayed in the order they appeared in.

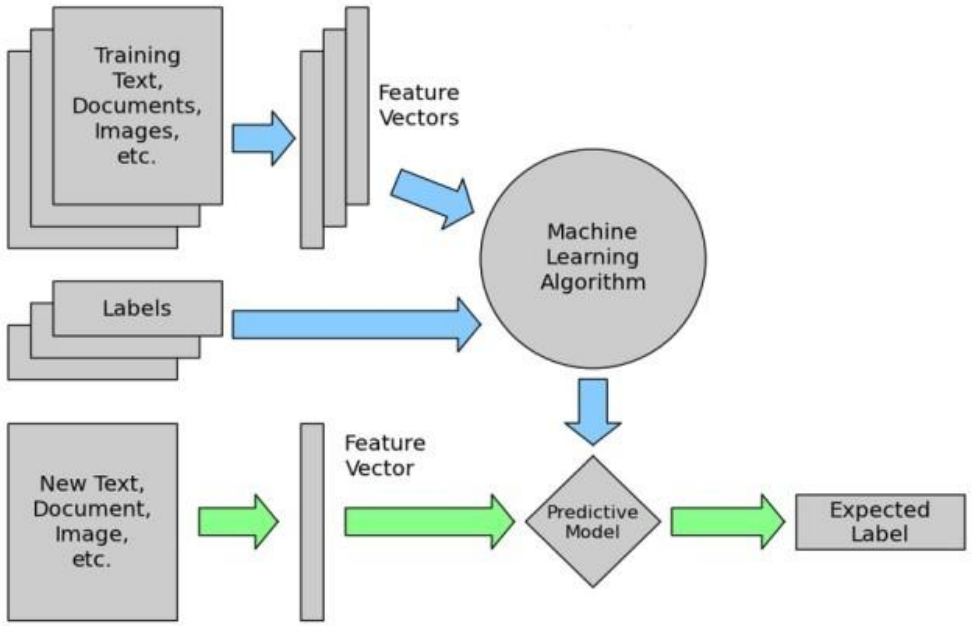
Pipeline





The strategy to adopt in as ML and then implementing NLP model evaluation and the process goes on

Methodology



Pre-Processing Technique:

- Model Vectorization (Skipgram & CBOW)

Classification Analysis:

- Naïve Bayes
- Decision Trees

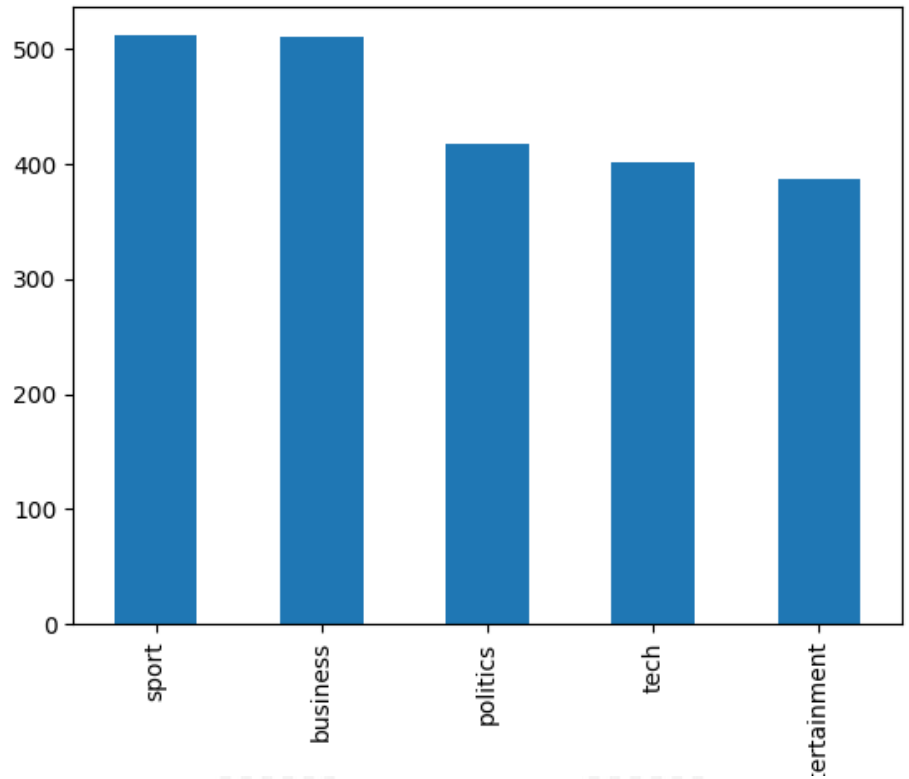
NLP model Evaluator:

- TextRank

Classification Analysis

```
1  docs['Category'].value_counts().plot.bar()
```

```
<Axes: >
```



Decision Trees tend to have lower accuracy compared to Naive Bayes Model when dealing with high dimensional feature space. Based on the experimental results, the Naive Bayes Classifier Model achieved the highest accuracy with **83.82%**. In contrast, the Decision Tree Classifier Model achieved lower accuracy of **70.34%**, respectively.

NLP Model Implementation

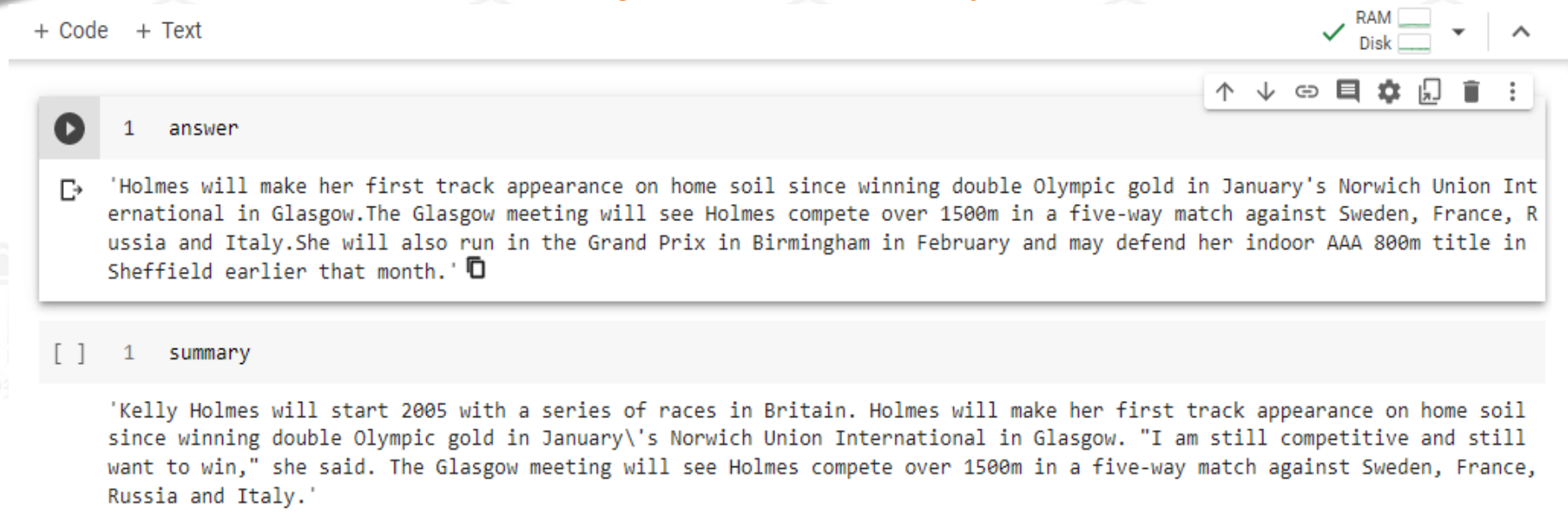
TEXT Rank Algorithm:

Is a graph-based ranking model for text processing which can be used in order to find the most relevant sentences in text and also to find keywords.

For extracting keywords in the extraction, we are providing it a vector of words and a vector of logicals indicating for each word if it is relevant.

Result

Text Rank is an algorithm that creates a summary of a news article by ranking sentences based on their importance in a graph. Naive Bayes is a machine learning algorithm that generates a summary by classifying sentences as either relevant or irrelevant based on probability. Decision trees recursively partition the data into subsets based on sentence features to generate a summary.



Findings

Using Skipgram									
Summary Length	ROUGE-1			ROUGE-2			ROUGE-L		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
40%	0.71	0.65	0.67	0.58	0.53	0.55	0.44	0.41	0.42
45%	0.77	0.63	0.69	0.63	0.53	0.57	0.47	0.39	0.43
50%	0.80	0.61	0.69	0.67	0.52	0.58	0.49	0.38	0.43

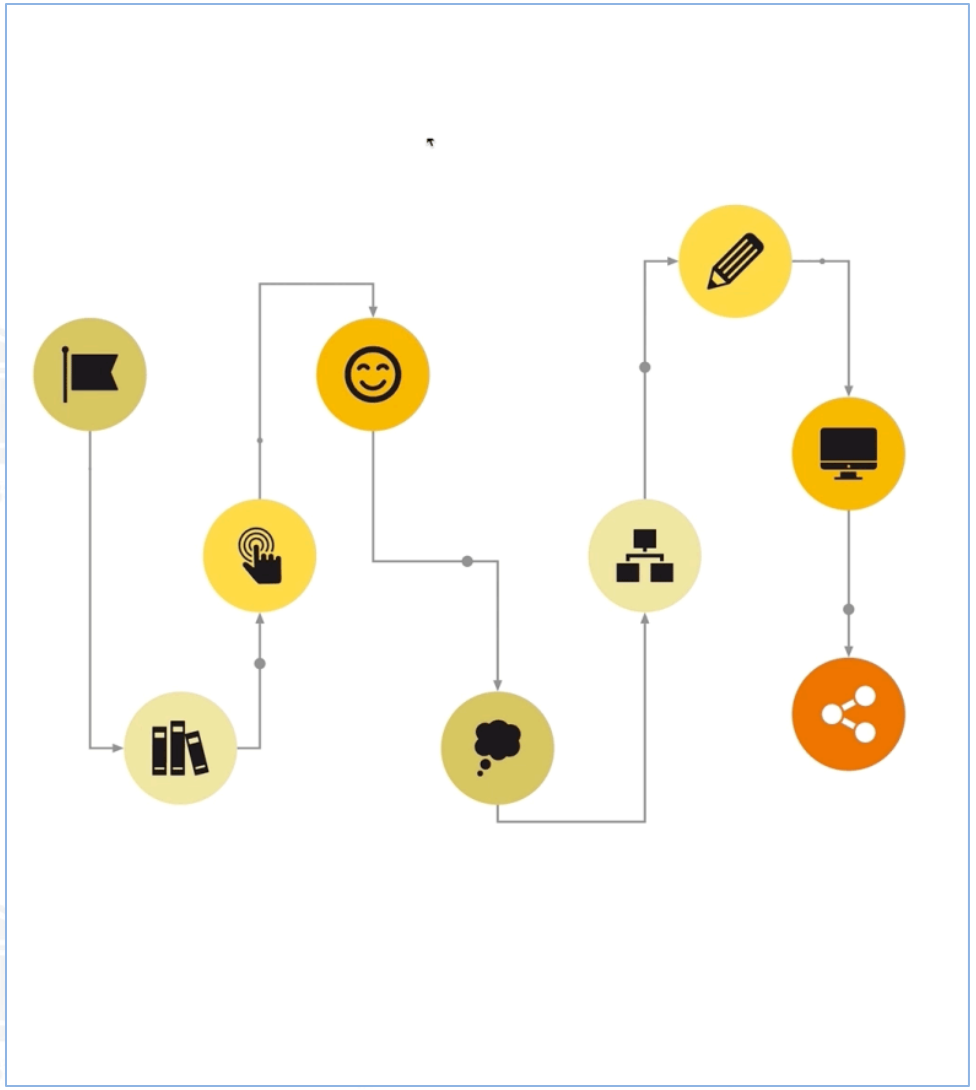
Using CBOW									
Summary Length	ROUGE-1			ROUGE-2			ROUGE-L		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
40%	0.63	0.63	0.63	0.48	0.48	0.48	0.39	0.40	0.39
45%	0.69	0.62	0.65	0.54	0.48	0.51	0.43	0.38	0.40
50%	0.72	0.60	0.65	0.58	0.48	0.52	0.44	0.37	0.40

Findings (by category)

Using SkipGram with 40% length									
News Category	ROUGE-1			ROUGE-2			ROUGE-L		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Business	0.71	0.65	0.67	0.58	0.54	0.56	0.46	0.43	0.44
Entertainment	0.70	0.65	0.67	0.57	0.53	0.55	0.46	0.43	0.44
Politics	0.72	0.66	0.68	0.58	0.54	0.56	0.43	0.40	0.41
Sport	0.69	0.64	0.66	0.57	0.53	0.55	0.45	0.42	0.43
Technology	0.73	0.67	0.70	0.58	0.53	0.56	0.42	0.39	0.40

Conclusion

Intelli-Brief provides a powerful and efficient way to stay up-to-date with the latest news and trends in your field of interest. Its AI-powered summarization and personalized content curation make it a valuable tool for professionals, researchers, and anyone seeking to stay informed in a rapidly evolving world.



Lessons Learned

- Building AI-powered text summarization models that quickly extract key information from large news articles.
- How to work with different libraries
- using natural language processing techniques to personalize and curate content effectively
- Enhanced understanding of how technology can solve real-world problems and improve daily productivity.

THANK YOU