Extractive Text Summarisation

Dhruv Kaushik(MT18037) Manshi Goel(MT18039) K Srivatsava(MT18054) P Akhil Kumar(MT18130)

Summarisation: Text Summarisation involves condensing a document to produce a human comprehensible summary. Two kinds of summarisation approaches have been suggested:

- Extractive
- Abstractive

Abstractive summarisation approach typically needs to "understand" the given document and paraphrase the salient concepts across the document. In contrast, Extractive summarisation approach commonly selects sentences that contain the most significant concepts in the document.

Purpose of Summarisation: Document extracts consisting of roughly 20% of the original can be as informative as the full text of a document[1]. Summarisation is to combine 20% of these document extracts so as to convey most information of document to user in short.

Goal of the Project: To develop an automatic text summariser which gives extractive summary of a given news article using machine learning techniques.

Dataset A collection of 2225 BBC news articles along with their respective summaries belonging to different genres like Business, Sports, Entertainment, Politics, and Tech. Each sentence in an article is considered as a data point.
https://www.kaggle.com/pariza/bbc-news-summary/

Extracting features for Classification

There are three categories of features:

- **Surface features:**

Feature Name	Description
Position	1/sentence no.
Doc_First	1 if it is the first sentence of a document 0 otherwise
Para_First	1 if it is the first sentence of a paragraph 0 otherwise
Length	The number of content words in the sentence
Quote	no. of non-quoted words/no. of words in the sentence

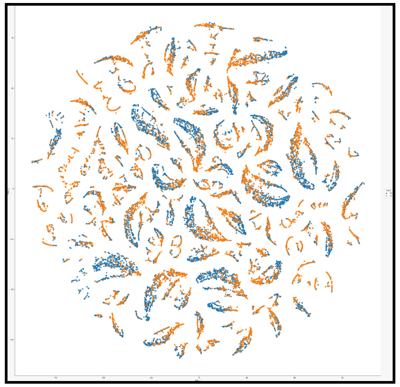
- **Content features:**

Feature Name	Description
CentroidVar_Uni	Average of TFIDF score of all content words considered one at a time(unigram) in a sentence
CentroidVar_Bi	Average of TFIDF score of all content words considered two at a time(bigram) in a sentence
SigTerm_Uni	The sum of signature unigrams in a sentence
FreqWord_Uni	Average of weights of frequent unigrams in a sentence
FreqWord_Bi	Average of weights of frequent bigrams in a sentence

Relevance features:

Feature Name	Description
FirstRel_Doc	Similarity of each sentence in a document with the first sentence of the document
FirstRel_Para	Similarity of each sentence in a paragraph with the first sentence of the paragraph
PageRankRel	PageRank value of each sentence based on the sentence similarity mapping

Methodology



Data Point Scatter plot

Logistic Regression

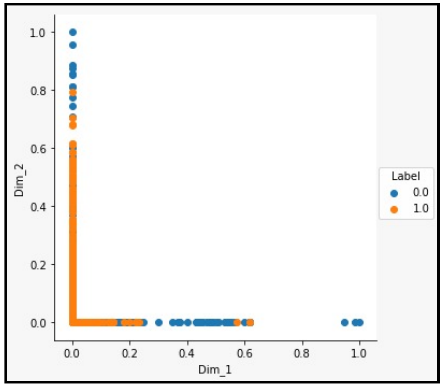


Fig. Logistic Regression - Dimension Altered

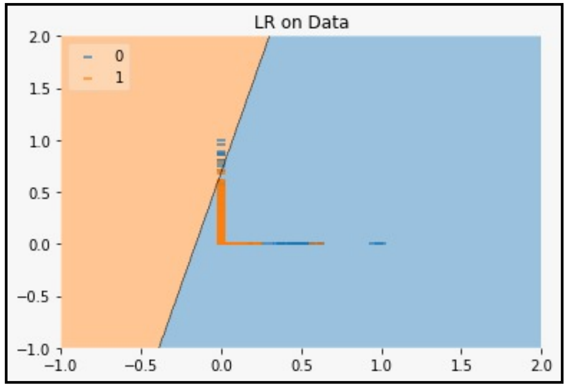


Fig. Decision boundary for Logistic Regression

Gaussian NB

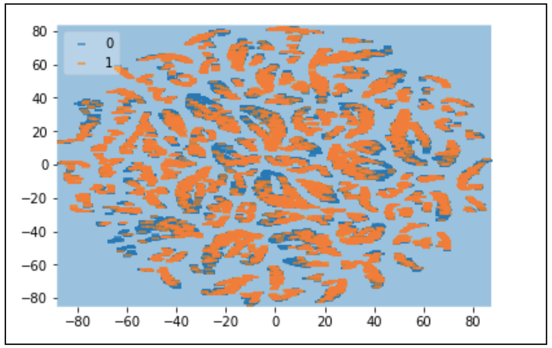


Fig. Decision boundary of Gaussian Naive Bayes

K-NN

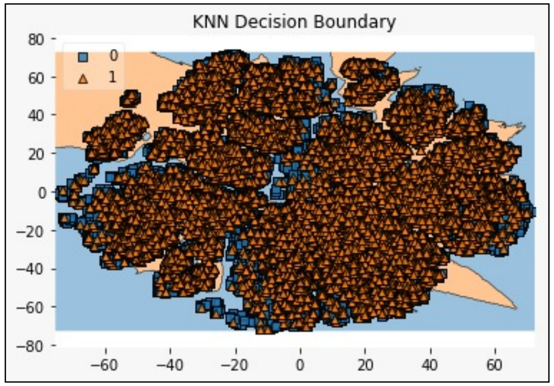


Fig. Decision boundary of K-NN

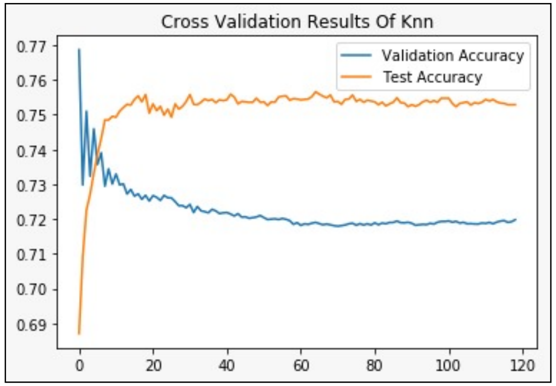


Fig. Feature Selection

Feature Combinations	Accuracy	F1-score
Surface	0.649	0.307
Content	0.710	0.585
Relevance	0.758	0.650
Surface + Content	0.717	0.597
Surface + Relevance	0.763	0.662
Content + Relevance	0.766	0.670
Surface + Content + Relevance	**0.767**	**0.671**

Logistic Regression Results

Feature Combinations	Accuracy	F1-score
Surface	0.633	0.380
Content	0.695	0.627
Relevance	0.737	0.575
Surface + Content	0.692	0.590
Surface + Relevance	0.698	0.507
Content + Relevance	**0.747**	**0.667**
Surface + Content + Relevance	0.725	0.618

Gaussian Naive Bayes Results

Feature Combinations	Accuracy	F1-Score
Surface	0.610	0.418
Content	0.659	0.636
Relevance	0.713	0.684
Surface + Content	0.667	0.642
Surface + Relevance	0.725	0.687
Content+Relevance	0.710	0.683
Surface+Content+Relevance	**0.717**	**0.687**

K-NN Results

Evaluation Metrics

Model	Average F-Measure
K-Nearest Neighbours	0.775951582747121
Gaussian Naive Bayes	0.7406700829597882
Logistic Regression	0.7397993248921582

F1-Scores for all the models