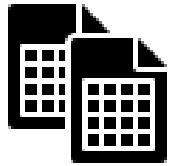
University of Milano-Bicocca
Master Degree in Data Science
Text Mining & Search
Accademic Year 23/24

Text Classification and Text Summarization of BBC NEWS

Authors:

JULIA TSYMBAL, PAOLA MARIA CAVANA

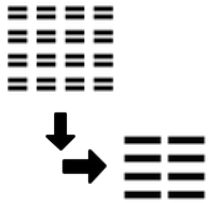
Outline



1. Dataset

2. Preprocessing

3. Text Classification



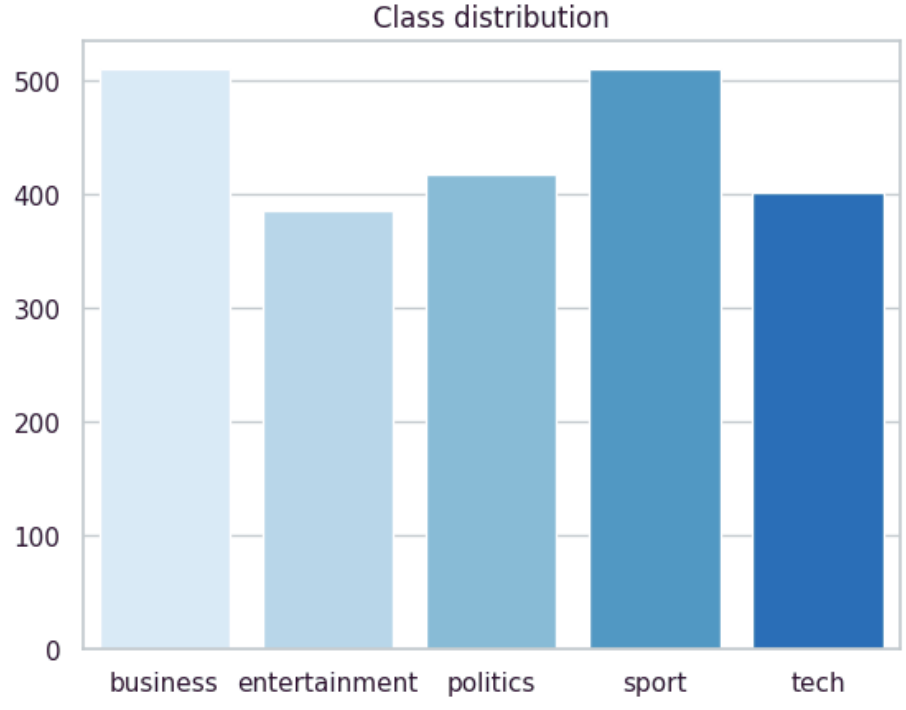
4. Text Summarization

Dataset: BBC News

The dataset used for this project was found on the **Kaggle** platform. It consists of **2225** articles from the BBC news website corresponding to stories in five topical areas from 2004-2005.

The news are divided into **5** classes:

- Business
- Entertainment
- Politics
- Sport
- Tech



Dataset: BBC News

Features presented in the dataset are:

- category of article
- article id
- text of article
- summary of article

category	article_id	text	summary
business	73	German growth goes into reverse Germany's economy shrank 0.2% in the last three months of 2004, upsetting hopes of a sustained recovery. The figures confounded hopes of a 0.2% expansion in the fourth quarter in Europe's biggest economy. The Federal Statistics Office said growth for the whole of 2004 was 1.6%, after a year of contraction in 2003, down from an earlier estimate of 1.7%. It said growth in the third quarter had been zero, putting the economy at a standstill from July onward. Germany has been reliant on exports to get its economy back on track, as unemployment of more than five million and impending cuts to welfare mean German consumers have kept their money to themselves. Major companies including Volkswagen, DaimlerChrysler and Siemens have spent much of 2004 in tough talks with unions about trimming jobs and costs. According to the statistics office, Destatis, rising exports were outweighed in the fourth quarter by the continuing weakness of domestic demand. But the relentless rise in the value of the euro last year has also hit the competitiveness of German products overseas. The effect has been to depress prospects for the 12-nation eurozone as a whole, as well as Germany. Eurozone interest rates are at 2%, but senior officials at the rate-setting European Central Bank are beginning to talk about the threat of inflation, prompting fears that interest rates may rise. The ECB's mandate is to fight rising prices by boosting interest rates - and that could further threaten Germany's hopes of recovery.	The figures confounded hopes of a 0.2% expansion in the fourth quarter in Europe's biggest economy.Germany's economy shrank 0.2% in the last three months of 2004, upsetting hopes of a sustained recovery.The ECB's mandate is to fight rising prices by boosting interest rates - and that could further threaten Germany's hopes of recovery.It said growth in the third quarter had been zero, putting the economy at a standstill from July onward.Germany has been reliant on exports to get its economy back on track, as unemployment of more than five million and impending cuts to welfare mean German consumers have kept their money to themselves.

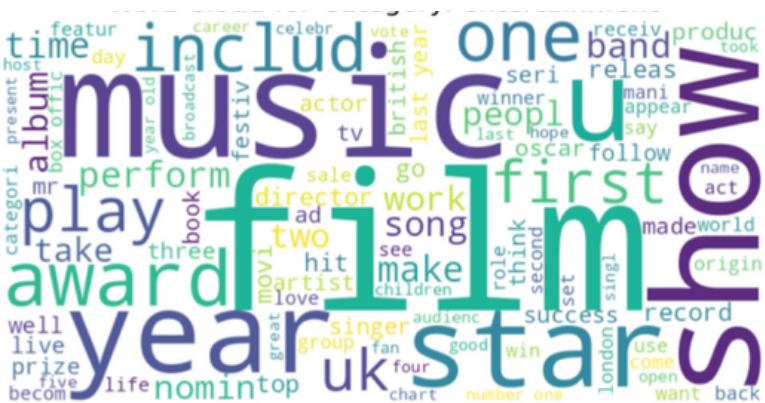
Dataset

Business

Politics

Tech

Entertainment



Sport

PreProcessing

- **Text Normalization**
  - Lowercasing
  - Removing punctuation, numbers, and special characters
  - Eliminating Accents
  - Handling Contractions and Abbreviations
  - Handling Whitespace
- **Stop-Words removal**
- **Text Tokenization**
  - Word tokenization
- **Stemming**

PreProcessing Result

category	article_id	text	processed_text
business	73	Germany's economy shrank 0.2% in the last three months of 2004, upsetting hopes of a sustained recovery.The figures confounded hopes of a 0.2% expansion in the fourth quarter in Europe's biggest economy. The Federal Statistics Office said growth for the whole of 2004 was 1.6%, after a year of contraction in 2003, down from an earlier estimate of 1.7%. It said growth in the third quarter had been zero, putting the economy at a standstill from July onward.	"german,growth,goe,revers,germani,economi,shrank,last,three,month,upset,hope,sustain,recoveri,figur,confound,hope,expans,fourth,quarter,europ,biggest,economi,feder,statist,offic,growth,whole,year,contract,earlier,estim,growth,third,quarter,zero,put,economi,standstil,juli,onward”

Text Classification

Text representation

- TF-IDF
- Word2Vec

Modeling algorithms

- LogisticRegression
- Decision tree
- Random forest
- SVM
- XGBoost

Size of vocabulary is 2116 unique words; number of tokens is 2224

Dataset size

Split 80%-20%

- Training set: 1779
- Test set: 445

Vectors dimension

IT-IDF

- train vector: (1779, 17443)
- test vector: (445, 17443)

Word2Vec

- train vector: (1779, 100)
- test vector: (445, 100)

Evaluation

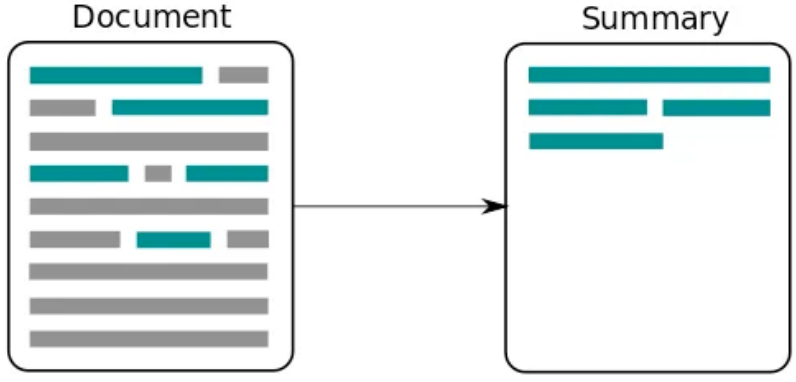
	Model	f1-score_tfidf	f1-score_w2v
0	Logistic Regression	**0.973**	0.887
1	SVM	**0.980**	0.873
2	XGBoost	0.953	**0.918**
3	DecisionTree	0.841	0.885
4	Random Forest	0.953	0.898

Text Summarization

Text summarization is the process of generating a **concise** and **coherent** summary of a given text while retaining its **essential information** and **meaning**.

In our project we performed:

- **Extractive summarization**:
  - TextRank algorithm
- **Abstractive summarization**:
  - BART

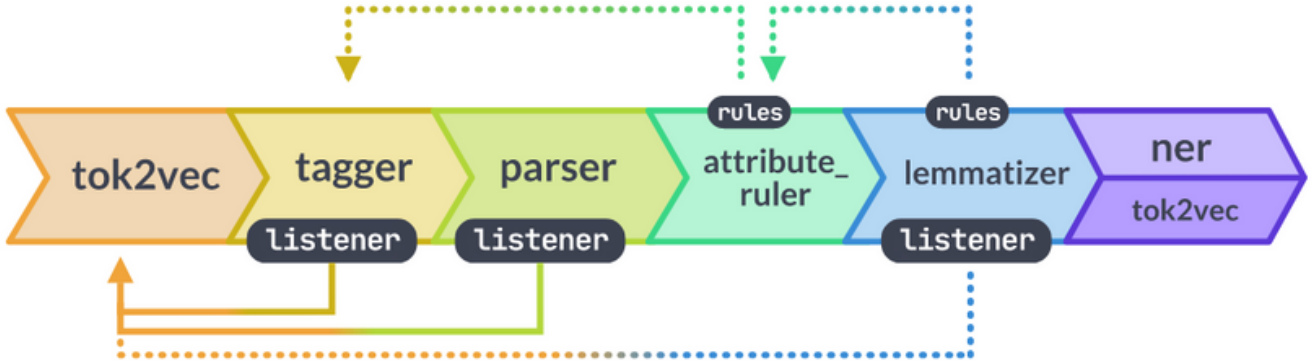


Extractive

For this part, we employed **TextRank.**

TextRank is an algorithm based on graph theory, It represents sentences as nodes in a graph, with edges indicating the similarity between sentences. By ranking sentences using graph algorithms, TextRank identifies the most important sentences, forming a concise summary by selecting top-ranked sentences.

In our project we used **pytextrank** library, that is a Python implementation of TextRank as a spaCy pipeline extension, which provides fast, effective phrase extraction from texts.



Original pipeline design provided spaCy

Extractive Results

Text	Extracted summary	Reference summary
German growth goes into reverse Germany's economy shrank 0.2% in the last three months of 2004, upsetting hopes of a sustained recovery. The figures confounded hopes of a 0.2% expansion in the fourth quarter in Europe's biggest economy. The Federal Statistics Office said growth for the whole of 2004 was 1.6%, after a year of contraction in 2003, down from an earlier estimate of 1.7%. It said growth in the third quarter had been zero, putting the economy at a standstill from July onward. Germany has been reliant on exports to get its economy back on track, as unemployment of more than five million and impending cuts to welfare mean German consumers have kept their money to themselves. Major companies including Volkswagen, DaimlerChrysler and Siemens have spent much of 2004 in tough talks with unions about trimming jobs and costs. According to the statistics office, Destatis, rising exports were outweighed in the fourth quarter by the continuing weakness of domestic demand. But the relentless rise in the value of the euro last year has also hit the competitiveness of German products overseas. The effect has been to depress prospects for the 12-nation eurozone as a whole, as well as Germany. Eurozone interest rates are at 2%, but senior officials at the rate-setting European Central Bank are beginning to talk about the threat of inflation, prompting fears that interest rates may rise. The ECB's mandate is to fight rising prices by boosting interest rates - and that could further threaten Germany's hopes of recovery.	Eurozone interest rates are at 2%, but senior officials at the rate-setting European Central Bank are beginning to talk about the threat of inflation, prompting fears that interest rates may rise. The ECB's mandate is to fight rising prices by boosting interest rates - and that could further threaten Germany's hopes of recovery. German growth goes into reverse Germany's economy shrank 0.2% in the last three months of 2004, upsetting hopes of a sustained recovery. But the relentless rise in the value of the euro last year has also hit the competitiveness of German products overseas. Germany has been reliant on exports to get its economy back on track, as unemployment of more than five million and impending cuts to welfare mean German consumers have kept their money to themselves.	The figures confounded hopes of a 0.2% expansion in the fourth quarter in Europe's biggest economy.Germany's economy shrank 0.2% in the last three months of 2004, upsetting hopes of a sustained recovery.The ECB's mandate is to fight rising prices by boosting interest rates - and that could further threaten Germany's hopes of recovery.It said growth in the third quarter had been zero, putting the economy at a standstill from July onward.Germany has been reliant on exports to get its economy back on track, as unemployment of more than five million and impending cuts to welfare mean German consumers have kept their money to themselves.

Extractive Results

Text	Extracted summary	Reference summary
German bidder in talks with LSE Deutsche Boerse bosses have held "constructive, professional and friendly" talks with the London Stock Exchange (LSE), its chief has said. Werner Seifert met LSE chief executive Clara Furse amid rumours the German group may raise its bid to £1.5bn ($2.9bn) from its initial £1.3bn offer. However, rival suitor Euronext also upped the ante in the bid battle. Ahead of talks with the LSE on Friday, the pan-European bourse said it may be prepared to make its offer in cash. The Paris-based exchange, owner of Liffe in London, is reported to be ready to raise £1.4bn to fund a bid. The news came as Deutsche Boerse held its third meeting with the LSE since its bid approach in December which was turned down by the London exchange for undervaluing the business. However, the LSE did agree to leave the door open for talks to find out whether a "significantly improved proposal" would be in the interests of LSE's shareholders and customers. In the meantime, Euronext, which combines the Paris, Amsterdam and Lisbon stock exchanges, also began talks with the LSE. In a statement on Thursday, Euronext said any offer was likely to be solely in cash, but added that: "There can be no assurances at this stage that any offer will be made." A deal with either bidder would create the biggest stock market operator in Europe and the second biggest in the world after the New York Stock Exchange. However, neither side has made a formal offer for the LSE, with sources claiming such a step may still be weeks away. Deutsche Boerse could also face mounting opposition to a bid at home. Among sweeteners reported to have been discussed by Mr Seifert with Ms Furse were plans to move the management of its cash and Eurex derivatives market to London, as well as two members of its executive board. But, Hans Reckers, a board member of Germany's central bank, the Bundesbank, said that cash trading should also remain in Frankfurt, something Deutsche Boerse could move to the UK. "It is not just the headquarters of the Boerse but also important market segments that must stay permanently in Frankfurt. This has special importance for the business activities of the banks and the consultants," he said. Local government officials in Frankfurt's state of Hessen have also spoken out against the move. "It is our wish that the headquarters stay here to maintain Frankfurt's standing as the number one financial centre in continental Europe," Alois Rhiel, its minister for economic affairs added.	But, Hans Reckers, a board member of Germany's central bank, the Bundesbank, said that cash trading should also remain in Frankfurt, something Deutsche Boerse could move to the UK. The news came as Deutsche Boerse held its third meeting with the LSE since its bid approach in December which was turned down by the London exchange for undervaluing the business. Werner Seifert met LSE chief executive Clara Furse amid rumours the German group may raise its bid to £1.5bn ($2.9bn) from its initial £1.3bn offer. Ahead of talks with the LSE on Friday, the pan-European bourse said it may be prepared to make its offer in cash. "It is our wish that the headquarters stay here to maintain Frankfurt's standing as the number one financial centre in continental Europe," Alois Rhiel, its minister for economic affairs added.	Deutsche Boerse bosses have held "constructive, professional and friendly" talks with the London Stock Exchange (LSE), its chief has said. But, Hans Reckers, a board member of Germany's central bank, the Bundesbank, said that cash trading should also remain in Frankfurt, something Deutsche Boerse could move to the UK. The news came as Deutsche Boerse held its third meeting with the LSE since its bid approach in December which was turned down by the London exchange for undervaluing the business. Ahead of talks with the LSE on Friday, the pan-European bourse said it may be prepared to make its offer in cash. Deutsche Boerse could also face mounting opposition to a bid at home. In the meantime, Euronext, which combines the Paris, Amsterdam and Lisbon stock exchanges, also began talks with the LSE. Werner Seifert met LSE chief executive Clara Furse amid rumours the German group may raise its bid to £1.5bn ($2.9bn) from its initial £1.3bn offer."It is not just the headquarters of the Boerse

Extractive Evaluation

For evaluating the performance of TextRank in extractive summarization, we employed the **ROUGE** score, specifically focusing on **ROUGE-1** and **ROUGE-L,** then **BLEU** score

- ROUGE-1: **0.63**
- ROUGE-L: **0.42**
- BLEU: **0.40**

Abstractive

For this part, we employed **BART** from Hugging Face.

BART is a transformer **encoder-decoder** (seq2seq) model with a bidirectional (BERT-like) encoder and an autoregressive (GPT-like) decoder. BART is pre-trained by corrupting text with an arbitrary noising function, and learning a model to reconstruct the original text.

BART is particularly effective when fine-tuned for text generation.

Abstractive

1. Tokenize and process data
2. Apply train test split. (80%-20%)
3. Train BART-base model
4. Evaluation

During training we have implemented **Gradient Accumulation.** With gradient accumulation, instead of updating the model parameters after each small batch, the gradients are accumulated over several batches, and the update is performed less frequently.

Abstractive Results

Text	Actual summary	Predicted summary
summarize: Jansen suffers a further setback Blackburn striker Matt Jansen faces three weeks out after surgery to treat a cartilage problem. But central defender Lorenzo Amoruso is moving closer to fitness following a knee operation. Rovers' assistant manager Mark Bowen said: "Matt had a small operation to trim knee cartilage. "It's a tiny piece of work, which should be a fairly quick recovery. Lorenzo is also jogging for the first time, along with kicking a ball." Jansen's career has been dogged by injury since a freak scooter accident two years ago. He returned to first-team action soon after Mark Hughes' appointment as Blackburn boss and marked it with a goal against Portsmouth in his first appearance of the season. Bowen added: "I'm guessing, but I reckon maybe two to three weeks before he is back in action completely." The Rovers assistant boss forecast a longer time spell for Amoruso's availability for first-team duties. Bowen said: "There's still some scar tissue present so it will be some weeks. "It's a case of see how he goes. You can't put a real-time on a comeback, we'll see how he progresses."	Rovers' assistant manager Mark Bowen said: "Matt had a small operation to trim knee cartilage. Bowen added: "I'm guessing, but I reckon maybe two to three weeks before he is back in action completely."The Rovers assistant boss forecast a longer time spell for Amoruso's availability for first-team duties. You can't put a real-time on a comeback, we'll see how he progresses."He returned to first-team action soon after Mark Hughes' appointment as Blackburn boss and marked it with a goal against Portsmouth in his first appearance of the season.	Jansen suffers a further setback, along with kicking a ball, as he is wearing a jersey wearing a shirt wearing a football jersey. Blackburn striker Matt Jansen faces three weeks out after surgery to treat a cartilage problem. But central defender Lorenzo Amoruso is moving closer to fitness following a knee operation. Rovers' assistant manager Mark Bowen said: "Matt had a small operation to trim knee cartilage. "It's a tiny piece of work, which should be a fairly quick recovery. Lorenzo is also jogging for the first time, along with a rugby ball." Jansen's career has been dogged by injury since a freak scooter accident two years ago.

Abstractive Evaluation

For evaluating the performance of BART in abstractive summarization, we employed the **ROUGE** score, specifically focusing on **ROUGE-1**, **ROUGE-2, ROUGE-L,** and then **BLEU** score

- ROUGE-1: **0.52**
- ROUGE-2: **0.37**
- ROUGE-L: **0.33**
- BLEU: **0.26**

THANK YOU FOR YOUR ATTENTION!