Enhancing News Article Analysis: Integrating Topic Modeling and Summarization Techniques

Vittorio Haardt, Luca Porcelli, Fabio Salerno

Introduction

Goal

Apply and study various techniques from both branches of text mining to assess their effectiveness.

1

2

3

Motivation

A more cohesive text management system designed to handle articles and news texts

Practical Applications

Utilize **Topic Modeling** for a labeling system

Employ **Summarization** to provide meaningful previews

The Dataset

CNN-DailyMail News dataset

- 300.000 articles + summary

20.000 articles subsample

article string · *lengths* 48 15.9k	highlights string · *lengths* 14 7.39k
LONDON, England (Reuters) -- Harry Potter star Daniel Radcliffe gains access to a reported…	Harry Potter star Daniel Radcliffe gets £20M fortune as he turns 18 Monday . Young actor…
Editor's note: In our Behind the Scenes series, CNN correspondents share their…	Mentally ill inmates in Miami are housed on the "forgotten floor" Judge Steven Leifman…
MINNEAPOLIS, Minnesota (CNN) -- Drivers who were on the Minneapolis bridge when it…	NEW: "I thought I was going to die," driver says . Man says pickup truck was folded in…

Topic modeling

Ever noticed how **plane seats** appear to be getting smaller and smaller? With increasing numbers of people taking to the skies, some experts are questioning if having such packed out planes is putting passengers at risk. They say that the shrinking space on aeroplanes is not only uncomfortable - it's putting our health and safety in danger. More than squabbling over the arm rest, shrinking space on planes putting our health and safety in danger? This week, a U.S. **consumer advisory group** set up by the Department of Transportation said at a public hearing that while the government is happy to set standards for animals flying on planes, it doesn't stipulate a minimum amount of space for humans. 'In a world where animals have more rights to space and food than humans,' said Charlie Leocha, consumer representative on the committee. 'It is time that the DOT and FAA take a stand for humane treatment of passengers.' But could crowding on planes lead to more serious issues than fighting for space in the overhead lockers, crashing elbows and seat back kicking? Tests conducted by the FAA use planes with a 31 inch pitch, a standard which on some airlines has decreased . Many **economy seats** on United Airlines have 30 inches of room, while some airlines offer as little as 28 inches . Cynthia Corbertt, a human factors researcher with the Federal Aviation Administration, that it conducts tests on how quickly passengers can leave a plane. But these tests are conducted using planes with 31 inches between each row of seats, a standard which on some airlines has decreased, reported the Detroit News. The distance between two seats from one point on a seat to the same point on the seat behind it is known as the pitch. While most airlines stick to a pitch of 31 inches or above, some fall below this. While United Airlines has 30 inches of space, Gulf Air economy seats have between 29 and 32 inches, Air Asia offers 29 inches and Spirit Airlines offers just 28 inches. British Airways has a seat pitch of 31 inches, while easyJet has 29 inches, Thomson's short haul seat pitch is 28 inches, and Virgin Atlantic's is 30-31.

Airplane Seat Size and Passenger Safety

Regulation and Standards for Airplane Seating

Impact of Decreased Airplane Seat Pitch

Text Summarization

Ever noticed how plane seats appear to be getting smaller and smaller? With increasing numbers of people taking to the skies, some experts are questioning if having such packed out planes is putting passengers at risk. They say that the shrinking space on aeroplanes is not only uncomfortable - it's putting our health and safety in danger. More than squabbling over the arm rest, shrinking space on planes putting our health and safety in danger? This week, a U.S consumer advisory group set up by the Department of Transportation said at a public hearing that while the government is happy to set standards for animals flying on planes, it doesn't stipulate a minimum amount of space for humans. 'In a world where animals have more rights to space and food than humans,' said Charlie Leocha, consumer representative on the committee. 'It is time that the DOT and FAA take a stand for humane treatment of passengers.' But could crowding on planes lead to more serious issues than fighting for space in the overhead lockers, crashing elbows and seat back kicking? Tests conducted by the FAA use planes with a 31 inch pitch, a standard which on some airlines has decreased . Many economy seats on United Airlines have 30 inches of room, while some airlines offer as little as 28 inches . Cynthia Corbertt, a human factors researcher with the Federal Aviation Administration, that it conducts tests on how quickly passengers can leave a plane. But these tests are conducted using planes with 31 inches between each row of seats, a standard which on some airlines has decreased, reported the Detroit News. The distance between two seats from one point on a seat to the same point on the seat behind it is known as the pitch. While most airlines stick to a pitch of 31 inches or above, some fall below this. While United Airlines has 30 inches of space, Gulf Air economy seats have between 29 and 32 inches, Air Asia offers 29 inches and Spirit Airlines offers just 28 inches. British Airways has a seat pitch of 31 inches, while easyJet has 29 inches, Thomson's short haul seat pitch is 28 inches, and Virgin Atlantic's is 30-31.

The shrinking space in airplane seats has raised concerns among experts who argue that it not only causes discomfort but also poses risks to passengers' health and safety. A U.S consumer advisory group criticized the lack of government standards for minimum seat space for humans, contrasting it with regulations for animals on planes. Tests conducted by the FAA revealed that many economy seats, such as those on United Airlines, offer less than the standard 31-inch pitch. Some airlines, like Spirit Airlines, provide as little as 28 inches, potentially impacting passengers' ability to evacuate quickly in emergencies. The debate suggests a need for regulatory action to ensure humane treatment and address safety concerns in increasingly crowded airplanes.

Task 1

Topic Modeling

Topic modeling

1

Preprocessing

Pre-process of text to properly apply LDA model

2

LDA Model Optimization

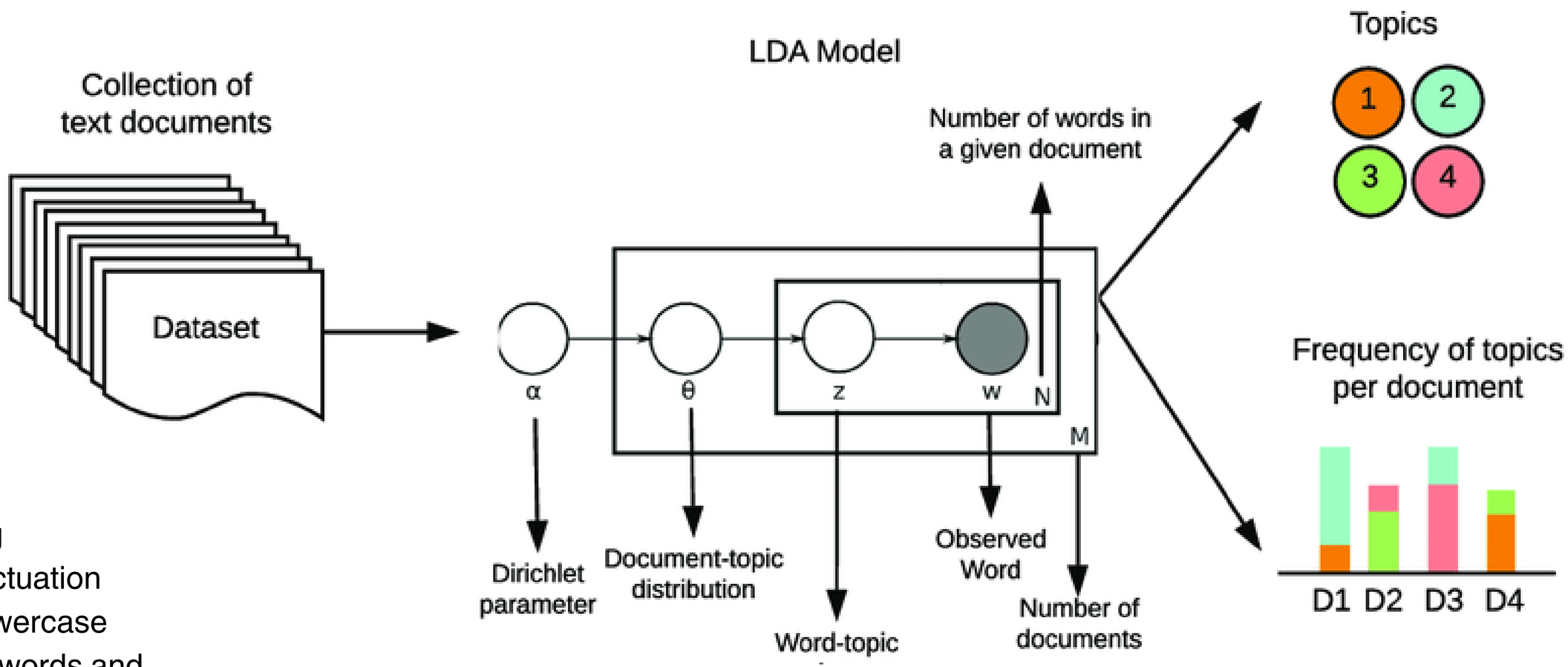
Optimisation of the hyperparemeters of the models under evaluation metrics (precision and choerence)

3

Evaluation and BerTopic comparison

Quantitative and qualitative comparison of performance with the state-of-art deep topic model BerTopic

LDA Model



Text processing

- Remove punctuation
- Convert to lowercase
- Remove stopwords and frequent words
- N-gram
- Lemmatization

LDA Oprimization

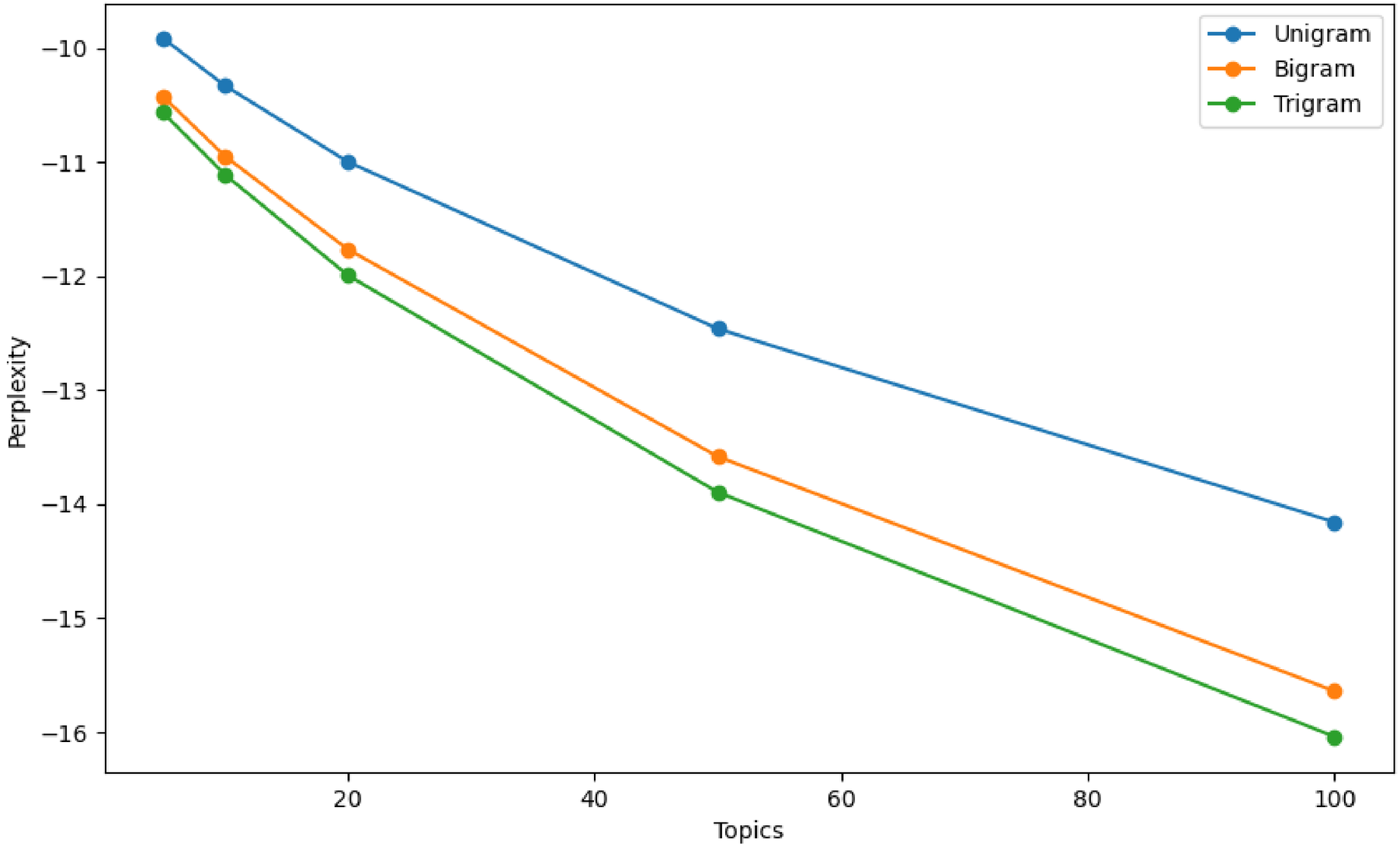
Hyperparameter

- Best Alpha and Beta
- Best n-gram

Evaluated using

Perplexity

- Alpha: 0.91
- Beta: 0.1
- Trigrams



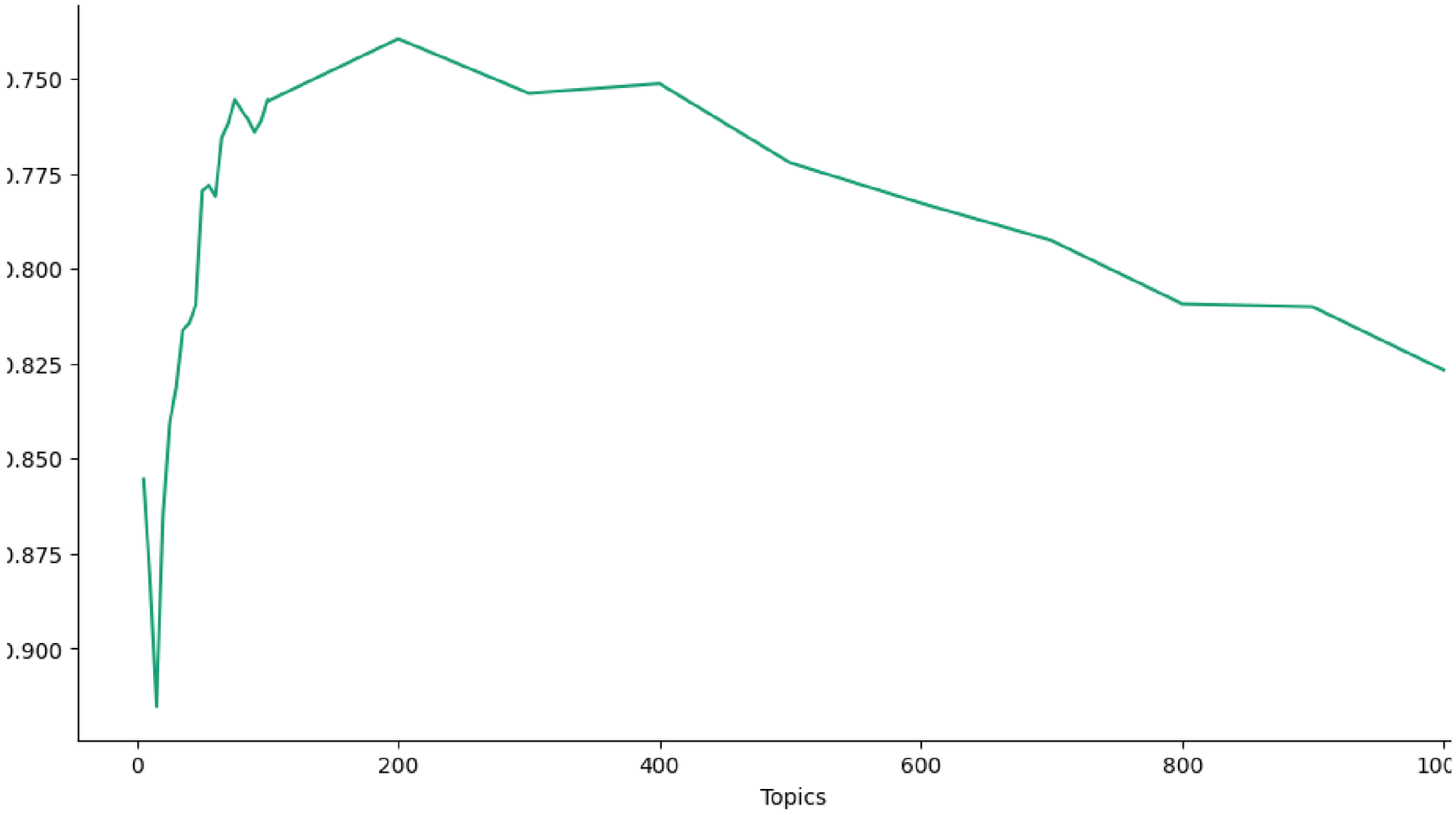
LDA Optimization

Number of topics

Evaluated using the maximum **U-mass** choerence score

↓

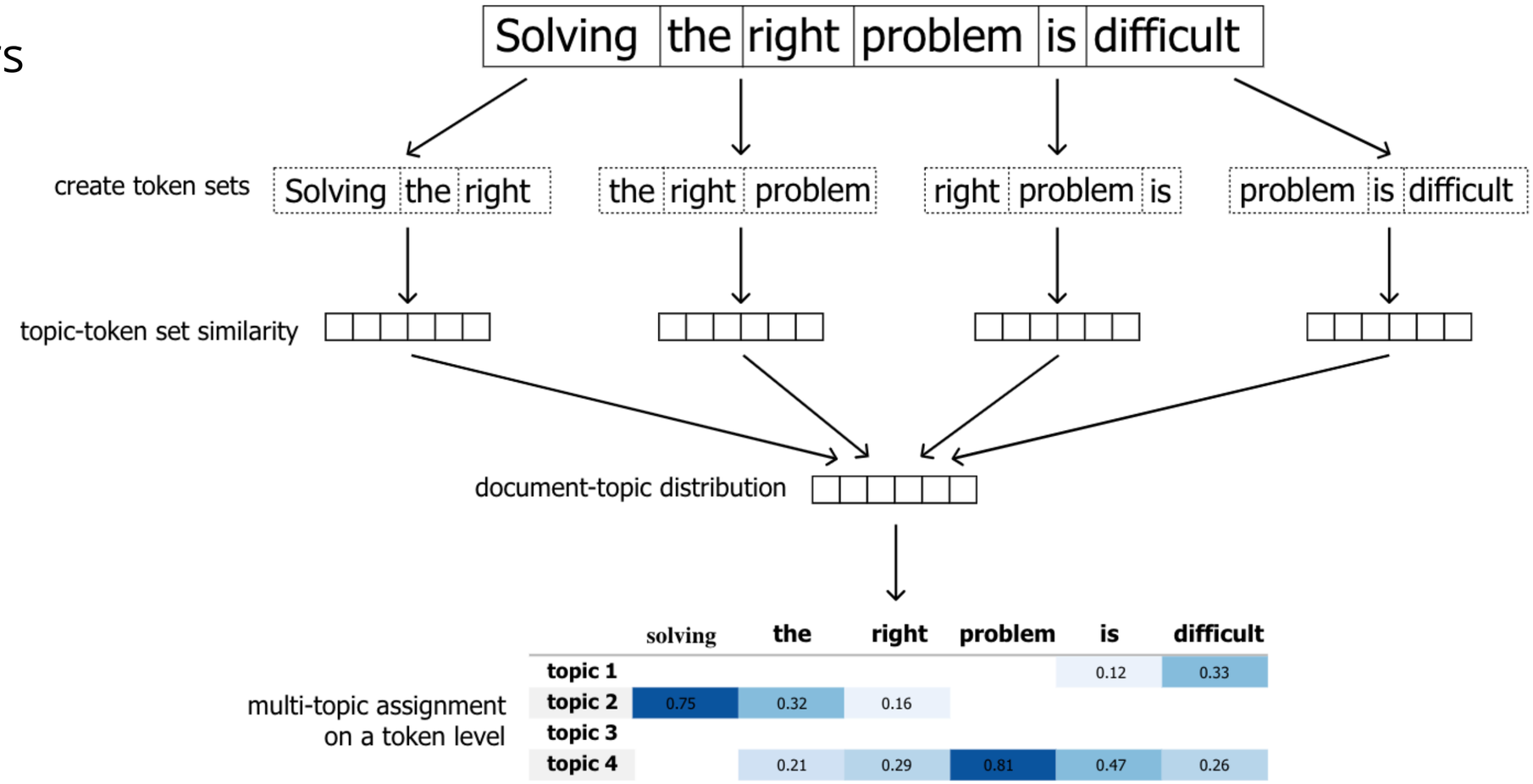
200 Topics



BerTopic

Deep Learning Approach

- Leverages transformers and c-TF-IDF
- Create dense clusters allowing for easily interpretable topics
- Keep important words in the topic descriptions.

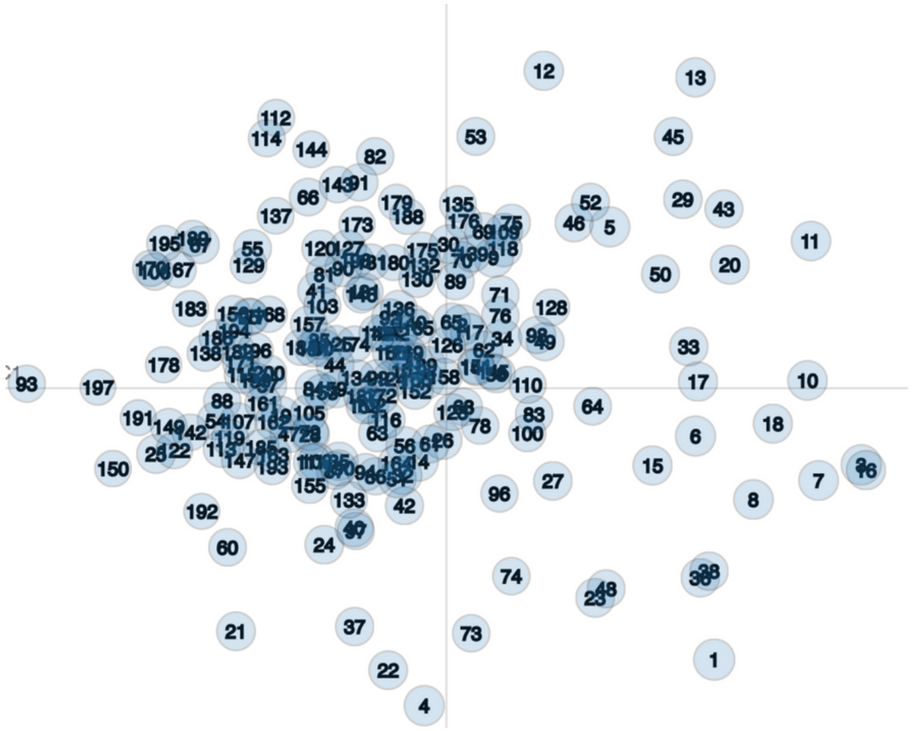


Evaluation

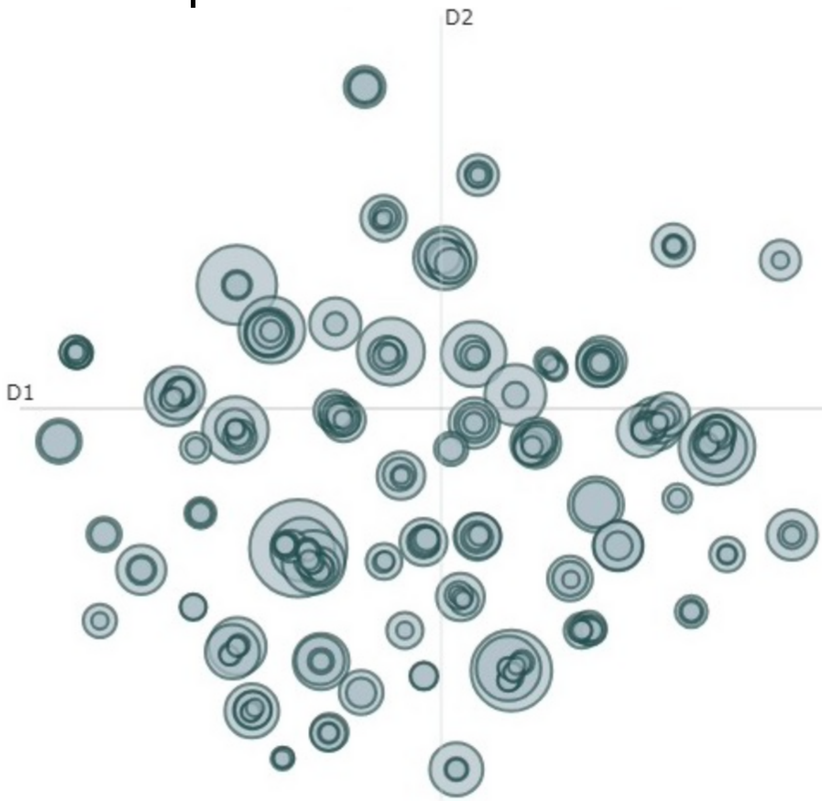
Bert offer a better spatial distribution and higher c_v

	c_v	# topic
LDA	0.26	200
BERT	**0.74**	259

LDA

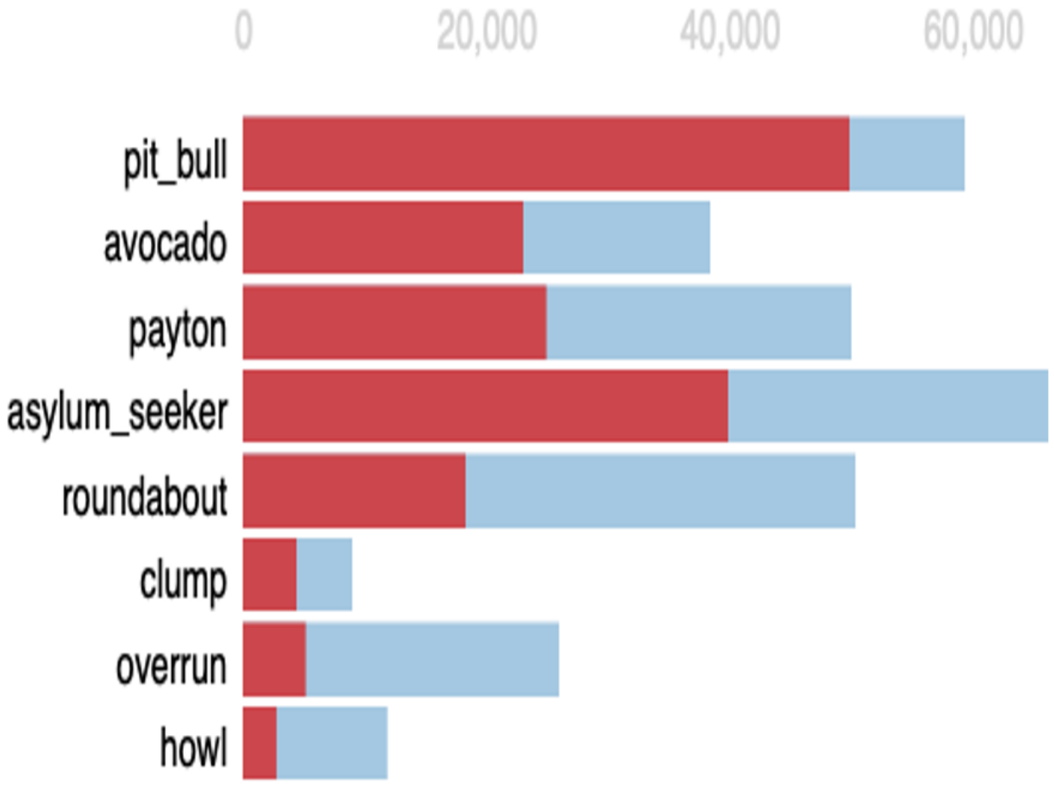


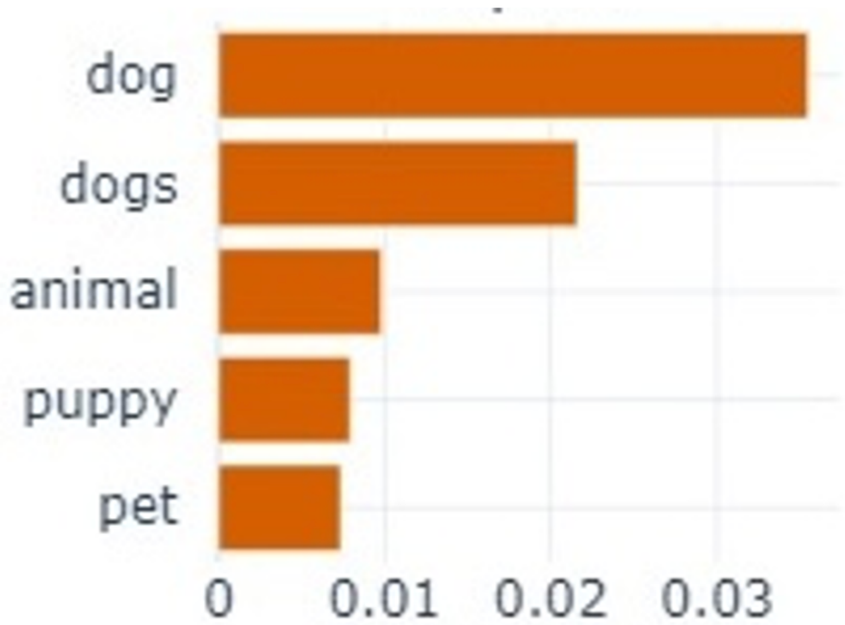
BerTopic



Evaluation

Bert Topics are more interpretable then LDA ones

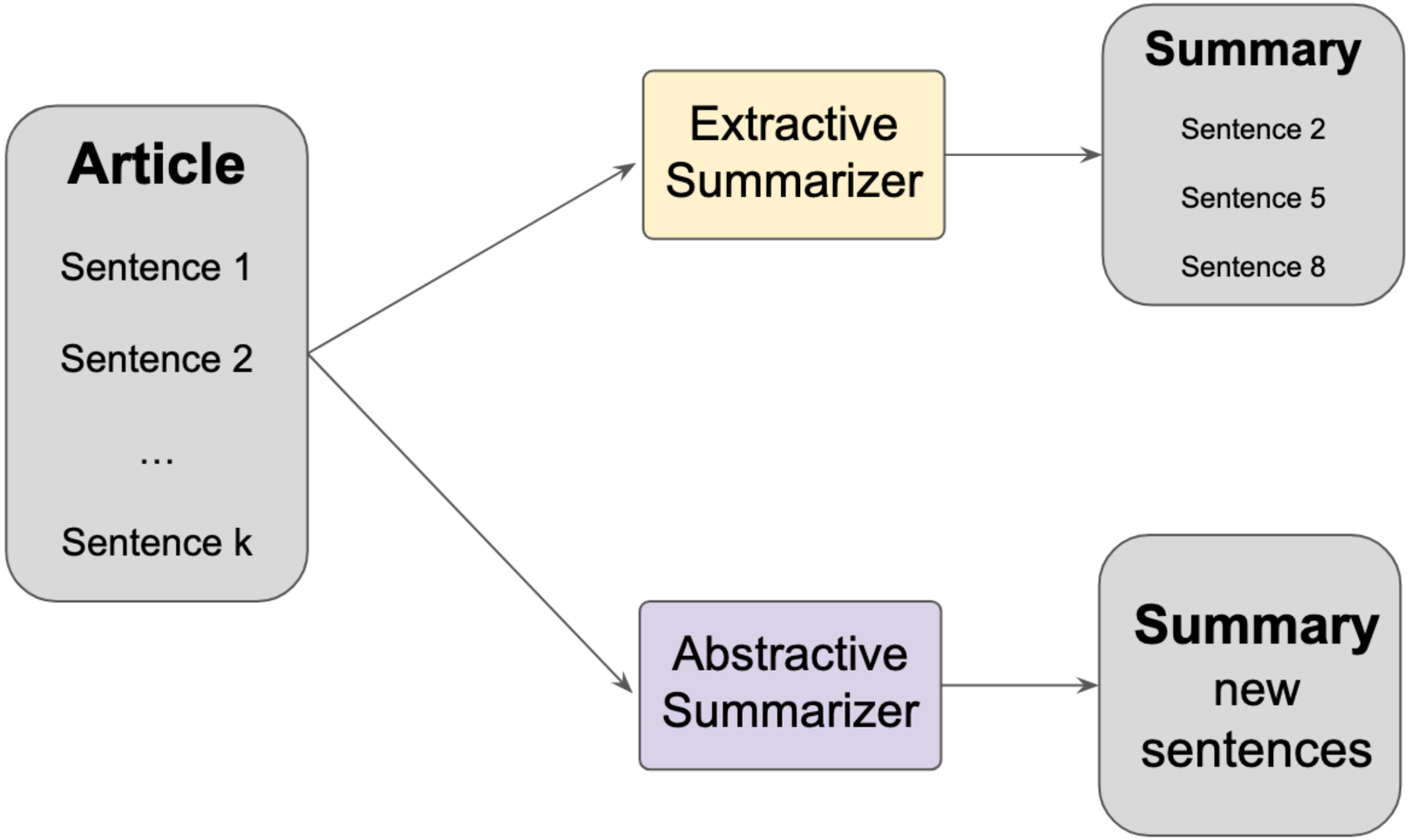




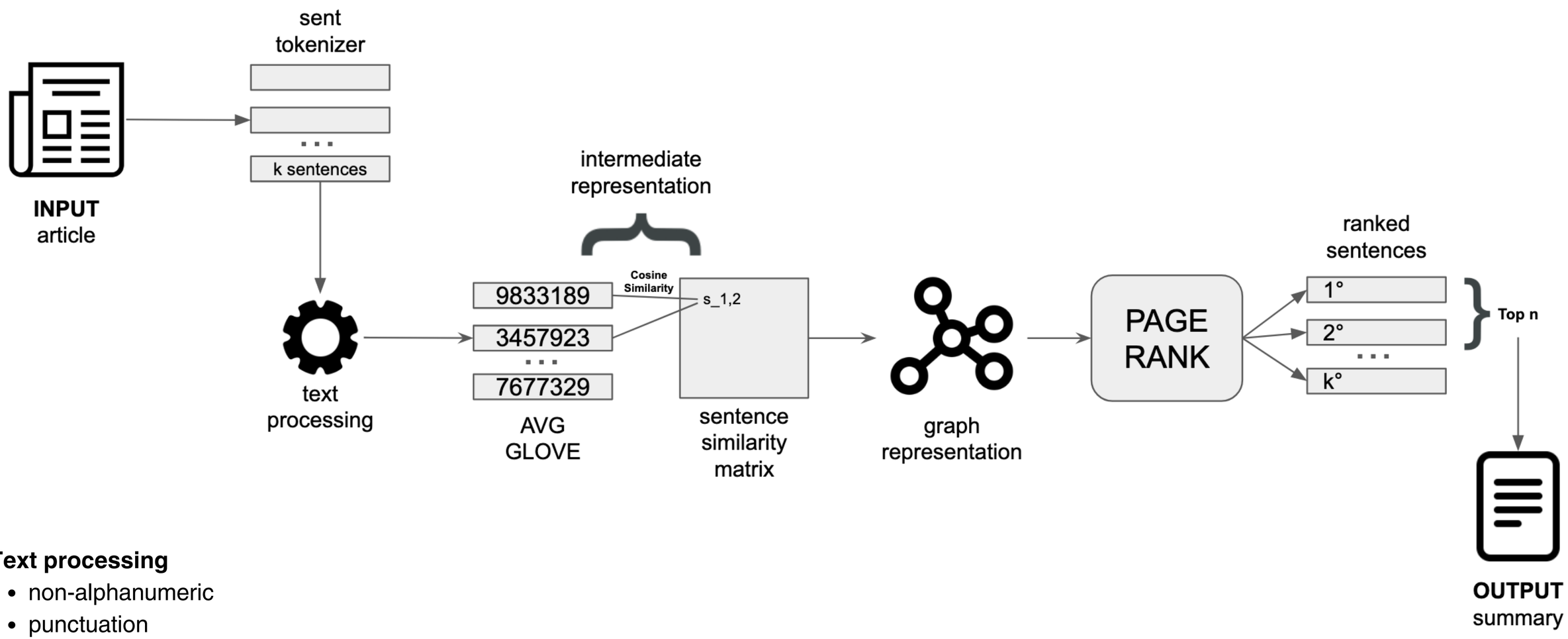
Task 2

Summarization

Summarization



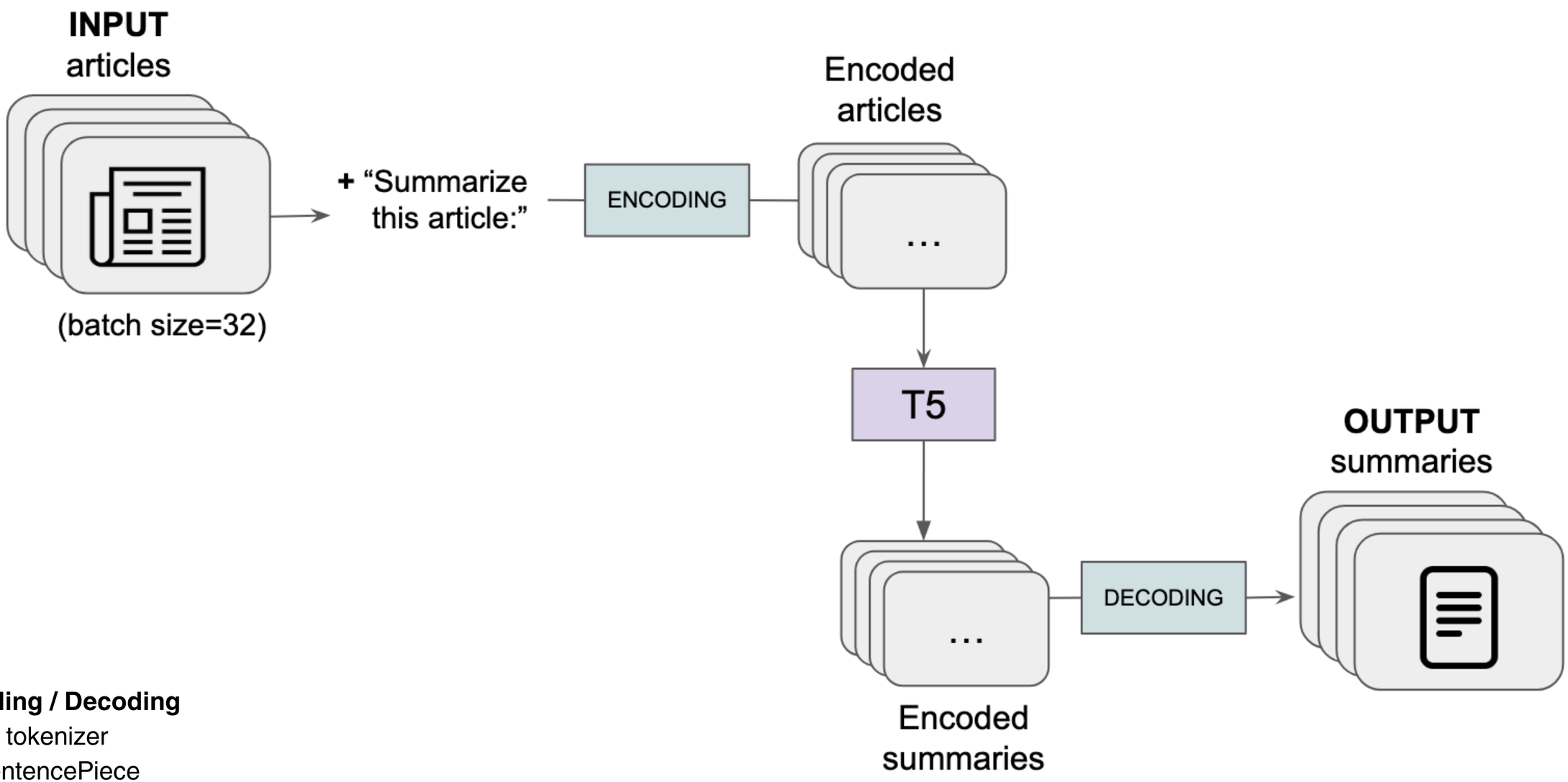
Text Rank summarization



Text processing

- non-alphanumeric
- punctuation
- case folding
- stopwords

T5 summarization



Encoding / Decoding

- T5 tokenizer
- SentencePiece

Evaluation

A 2.000 sample was used

- Computational costs
- Time costs

	ROUGE Lsum	BLEU	METEOR	BERT score
Extractive TextRank	21.38	4.97	**30.35**	77.51
Abstractive T5	**27.93**	**9.46**	24.5	**78.55**

Rouge and BLEU scores on the T5 summaries outperform TextRank summaries

METEOR score on the TextRank summaries is greater than the one on theT5 summaries

BERT score is almost the same. So both models possess strong semantic and contextual alignment capabilities

Conclusion

Conclusion

Topic Modeling

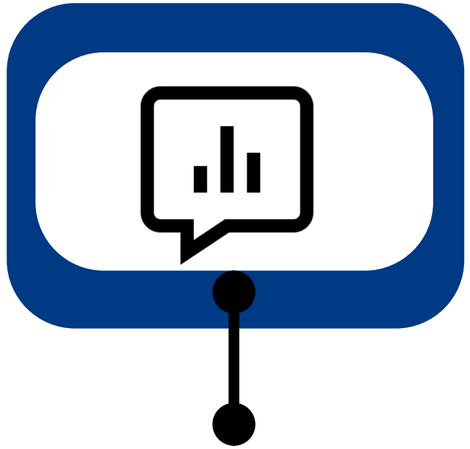
Better results with BerTopic approch then with LDA, highlighting the effectiveness of a deeper approach

Summarization

Abstractive fine tuned T5-small model performed better then extractive one, in therms of BLEU and ROUGE scores

Possible Development

Possible development of a system for an informed approach to accessing news and articles.



Topic Modeling

Labeling system, enabling users to apply filters

Summarization

Offer a concise overview of the articles

The End

Thanks for the attention!

- Vittorio Haardt
- Luca Porcelli
- Fabio Salerno