Making Text Analytics More Approachable: From Traditional to Generative AI for Business Applications

Paper # AL96

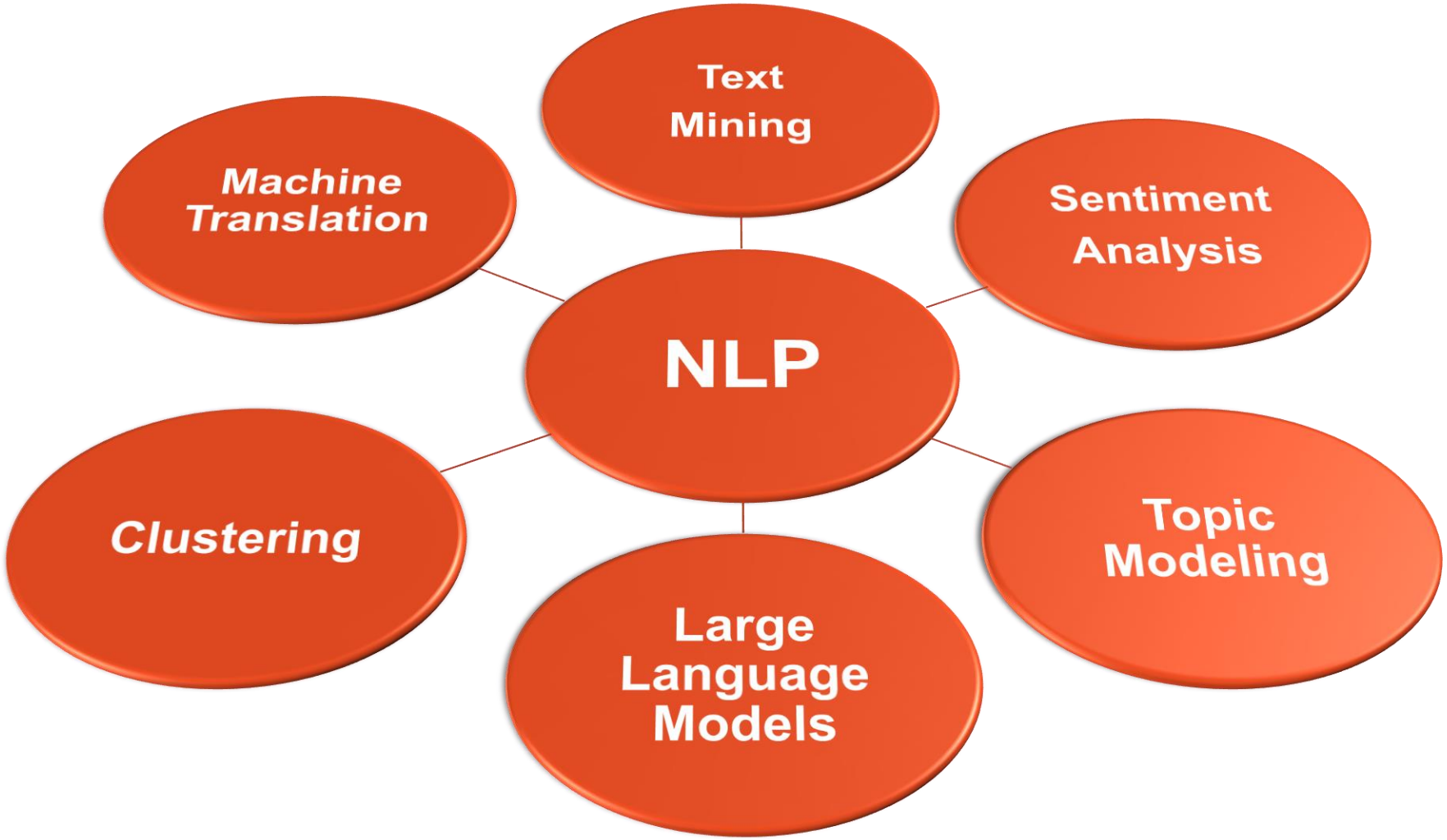
Vikranth Reddy Chapaala is a graduate student in the Business Analytics and Data Science program at Oklahoma State University, Stillwater, Oklahoma, with an expected graduation in May 2025. He holds a bachelor's degree in Computer Science and developed a passion for solving meaningful problems through data. Between his bachelor's and master's journey, Vikranth worked for approximately 2.5 years in Data Engineering and Analytics teams. During the summer, he worked as a Data Science Intern for a leading payroll and human resources company and is continuing in the same role alongside pursuing his master’s degree.

Making Text Analytics More Approachable: From Traditional to Generative AI for Business Applications

Paper # AL96

Vikranth Reddy Chapaala, Graduate Student
Oklahoma State University
Stillwater, Oklahoma

Background


Text Mining
Machine Translation
Sentiment Analysis
NLP
Clustering
Topic Modeling
Large Language Models

Focus Area

To conduct a comparative study on performance of unsupervised topic modeling techniques:

- Spanning from traditional to generative models
- Devise a new approach for combining topic modeling and text classification.

Techniques Used

- SAS Visual Text Analytics
- Text Analytics with Python

Image Credit: DALL-E

Topic Modeling

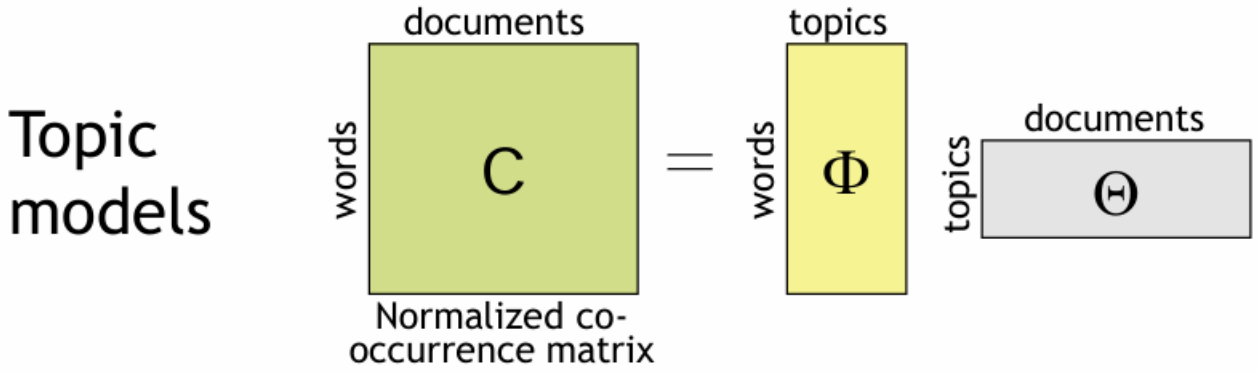
Key Idea: Documents are mixtures of latent topics, where a topic is a probability distribution over a word.

- The main way of automatically capturing the meaning of documents.
- The topic of an image: a cat, a dog..

Politics, President, Law, Policy….
Space, Planet, Astronaut, Mission….
Sports, Team, Player, Coach, Stadium….

Topic Modeling

- Hidden variables, generative processes, and statistical inference are the foundation of probabilistic modeling of topics



Reference: https://viasm.edu.vn/Cms_Data/Contents/Viasm-EN

Dataset

- 2,225 articles published online
- Each article is **labeled** under one of 5 categories:
  - Business
  - Entertainment
  - Politics
  - Sport
  - Tech

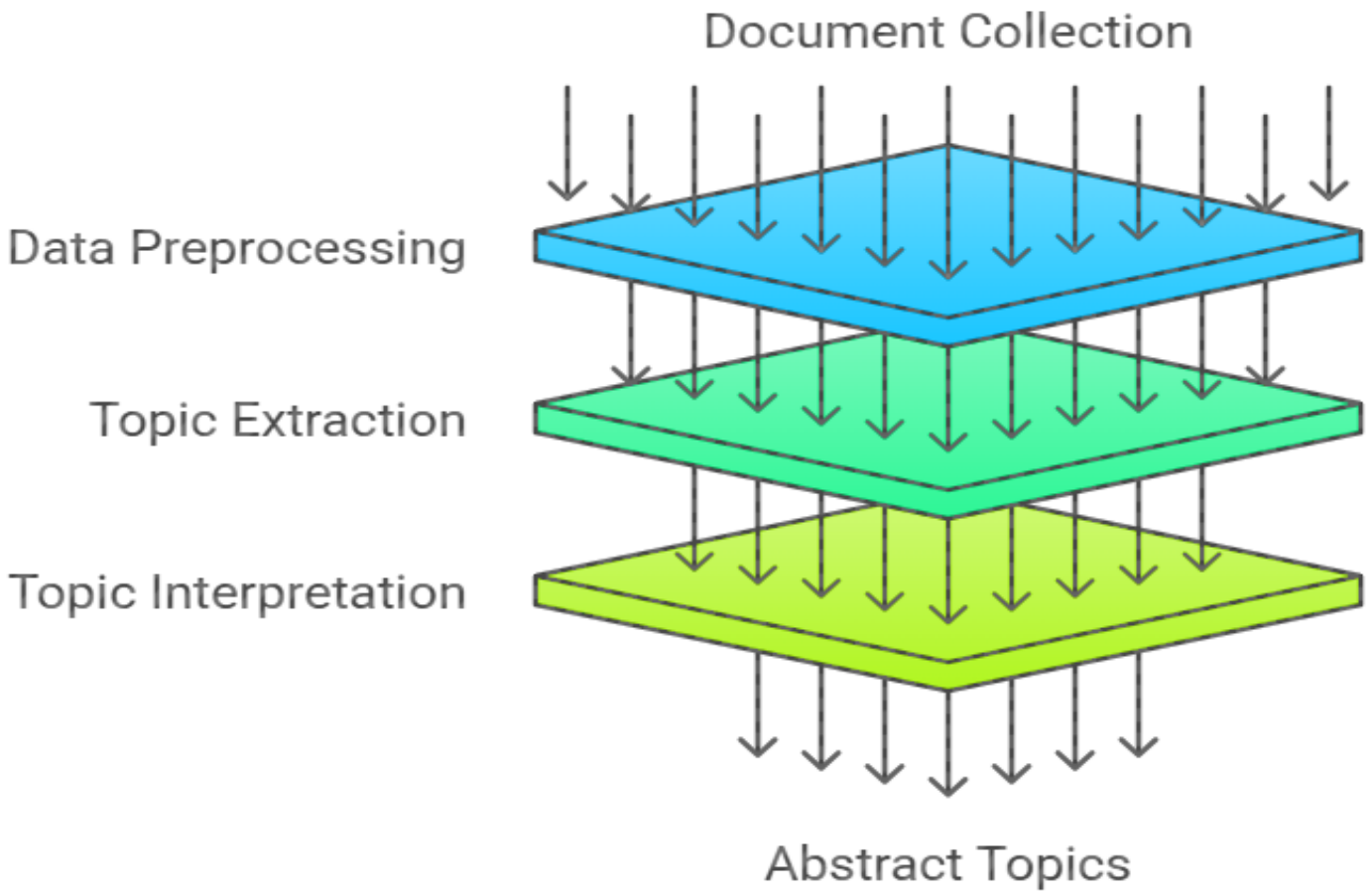
Sample Data

text	label_text
wales want rugby league training…..	Sport
new harry potter tops book……	Entertainment

https://huggingface.co/datasets/SetFit/bbc-news

http://mlg.ucd.ie/datasets/bbc.html

Traditional Architecture

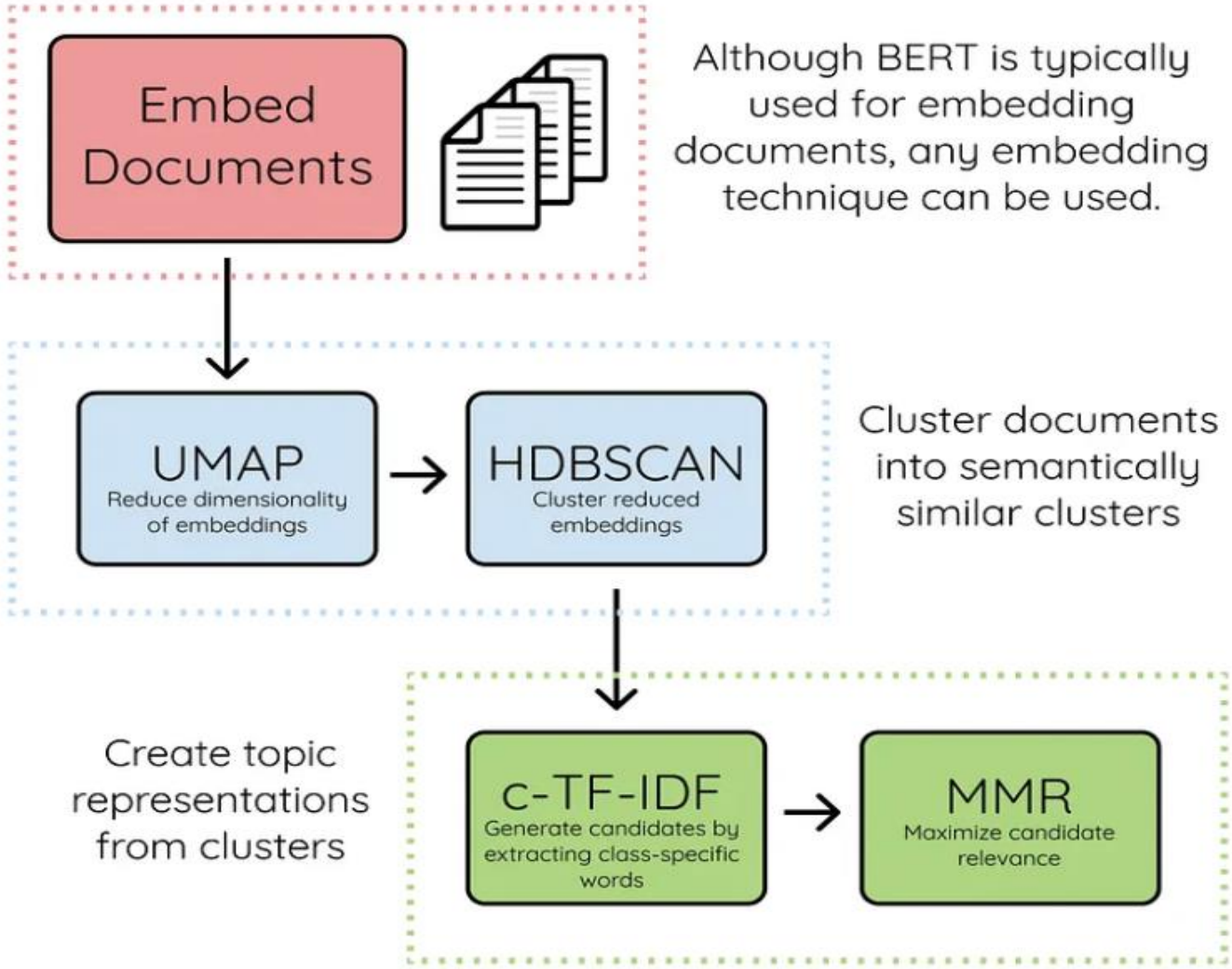


Tokenization, Pos Tagging etc.

LDA, NMF, LSA etc.

Topic Labeling, Evaluation, Visualization

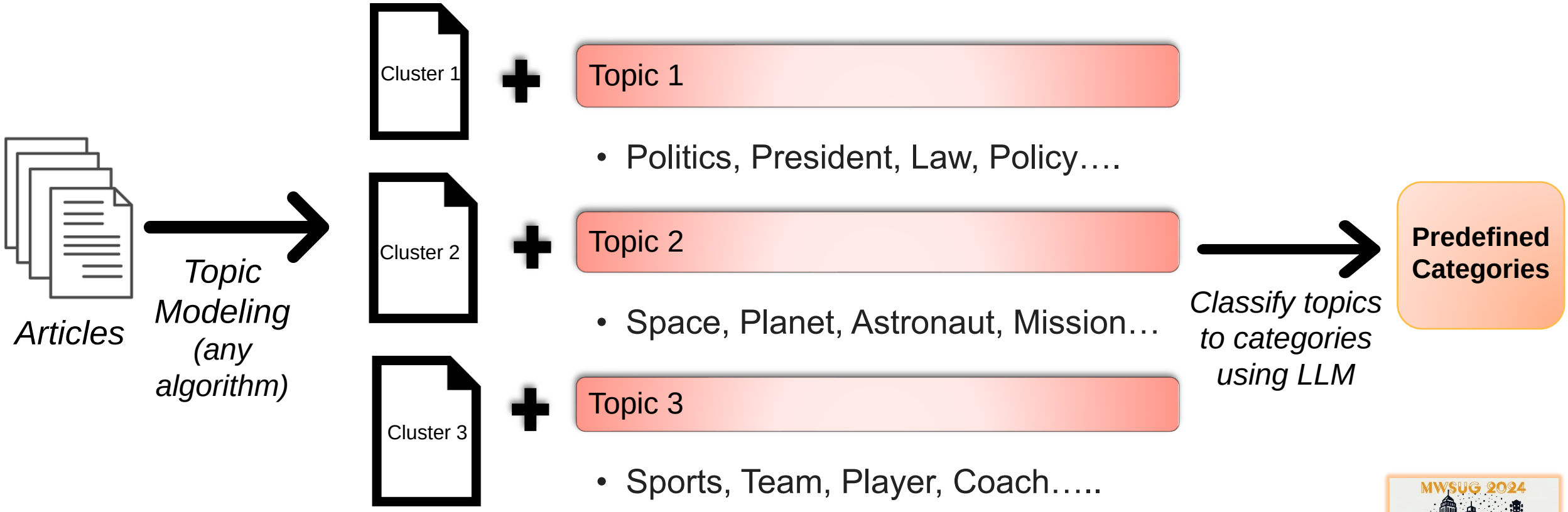
Transformer Based Architecture



Embedding Example:
"The cat sat on the mat"

	cat	mat	on	sat	the
the =>	0	0	0	0	1
cat =>	1	0	0	0	0
sat =>	0	0	0	1	0

Proposed Architecture (Topic Modelling + Classification)



Evaluations

Topic Modeling

Technique	# Topics	Coherence score
Lda	14	0.34
Top2vec	10	0.47
SAS	11	0.49
BERTopic	53	0.62

Classification

Model Type	Accuracy
BERT-based Model (BERTopic + LLM)	88.3%
LDA-based Model (LDA + LLM)	72.6%

Difference between two approaches

Traditional

- Based on probabilistic models or matrix factorization.
- Uses bag-of-words assumption (ignores word order).
- Focuses on identifying word co-occurrence patterns in documents.

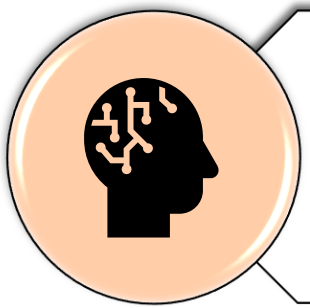
Modern

- Leverages word embeddings (BERT, GPT) to capture context.
- Uses deep learning and transformers for topic inference.
- Focuses on generating context-aware topics using semantic relationships.

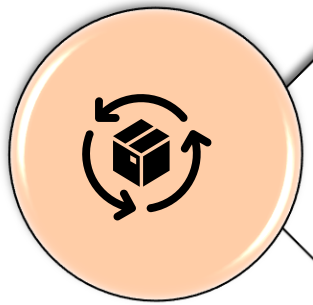
Use Cases

Calculator is to Math, LLMs are to ?

Conclusions



Combining topic modeling with language models enhances categorization.



Traditional algorithms often miss text semantics, but LLMs provide significant improvements.

Future work will focus on optimizing parameters, conducting detailed studies on specific user groups

Thank You!

Vikranth Reddy Chapaala

Graduate Student, Business Analytics and Data Science

Oklahoma State University

vikranth.reddy@okstate.edu

https://www.linkedin.com/in/vikranth-reddyc/