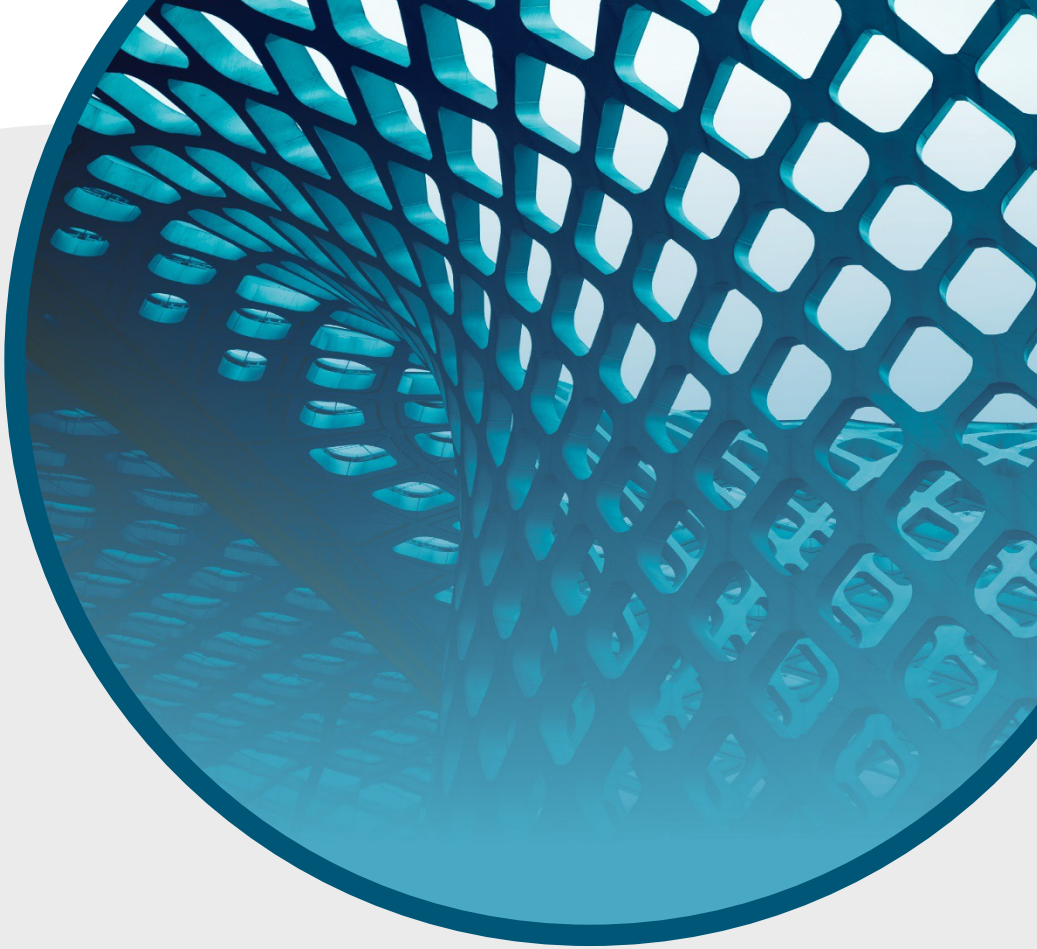


Comparing machine learning and human approaches for topic classification

Gerard Atkinson, Director, ARTD Consultants

Confession time

Contents

Rationale and Goal

Rationale and goal

Rationale and goal

Methods

Methods

Human Coding

- Traditional approach
- No automation

QAD Coding

- Semi-automated human approach
- Emergent framework
- Qualitative Automation by Direction (QAD)

Latent Dirichlet Analysis

- Statistical approach to topic classification
- Uses word frequency and proximity
- Results are presented as clusters of words

BERTopic

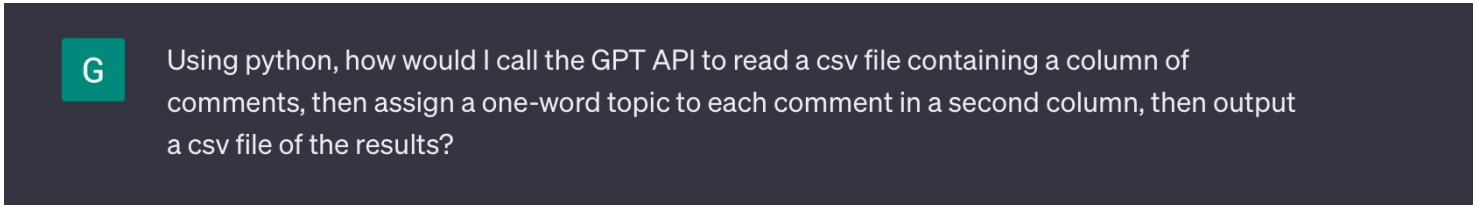
- Open source Machine Learning (ML) model
- Fine-tuned for topic classification tasks

GPT-3.5

- Online Large Language Model (LLM)
- Multi-purpose model behind Chat-GPT
- Application Programming Interface (API)

Where's ChatGPT?

- We used the same model ChatGPT uses!
- We also used ChatGPT to provide a little help in writing the computer code to call the API:



Where's <insert your favourite method here>?

Five other methods were trialled but not included:

- SAGE
- NMF
- GPT4All Falcon
- GPT4All LLaMa
- Bard LaMDA

Zero-shot and guided classification

Zero-shot

- No prior topic list
- Must generate best set of topics and assign based on content

Guided

- Existing list of topics
- Assigns probability based on content

Our test data

Li & Parikh (2020) dataset

N=1473 statements (diary entries)

Human tagged by emotion and topic (single classification)

Subset of 5 largest topics, with 200 entries selected at random as test set

Compared to other training sets (e.g. Twitter, IMDB, MNLI), the content more closely resembles evaluation qualitative data

Classification Rubric

Classification Rubric

- Rubrics have two elements:
  - Domains containing dimensions of merit (topics of interest)
  - Scale (levels of performance)

- The rubric fulfils three purposes:
  - Develop a consistent understanding of the effectiveness of different approaches
  - Enable a holistic assessment of approaches
  - Identify where there are gaps in methodologies that need to be addressed through further data collection

		Scale		
		Poor	Moderate	Good
Domain	Dimension A	☑		
	Dimension B			☑
	Dimension C		☑	

Classification Rubric

Dimension	Description
Accuracy	The approach was successful in matching the original coding
Speed	The approach delivered results in a timely fashion
Automation	The approach operated independently of human intervention
Ease of implementation	The approach was easy to set up and execute
Efficiency of implementation	The approach was cost-effective to implement
Efficiency of scale	The approach can be scaled to larger numbers of sources with minimal marginal cost

Scale point	Description
Low	The approach performs poorly on this dimension
Moderate	The approach provides reasonable performance but with some notable flaws
High	The approach performs well on this dimension with no or negligible flaws
N/A	It is not possible to make a confident judgement on this dimension

Findings

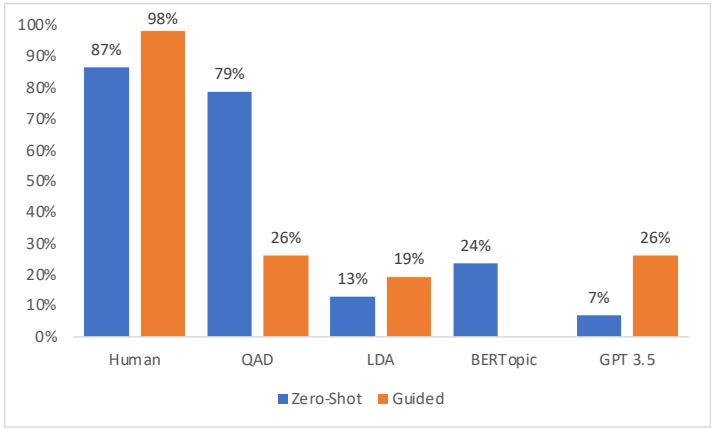
Findings

	Human Coding (Zero-shot)	QAD Coding (Zero-shot)	LDA (Zero-shot)	BERTopic (Zero-shot)	GPT-3.5 (Zero-shot)	Human Coding (Guided)	QAD Coding (Guided)	LDA (Guided)	BERTopic (Guided)	GPT-3.5 (Guided)
Accuracy	High	High	Low	Moderate	Low	High	Moderate	Moderate	N/A	Moderate
Speed	Low	Moderate	High	High	Moderate	Low	Moderate	High	N/A	Moderate
Automation	Low	Moderate	Moderate	High	High	Low	Moderate	Moderate	High	High
Ease of implementation	High	High	Low	Low	Moderate	Moderate	High	Low	Low	Moderate
Efficiency of implementation	Moderate	High	Moderate	Moderate	Moderate	High	High	Moderate	Moderate	Moderate
Efficiency of scale	Low	Moderate	High	High	Moderate	Low	Moderate	High	High	Moderate

Findings

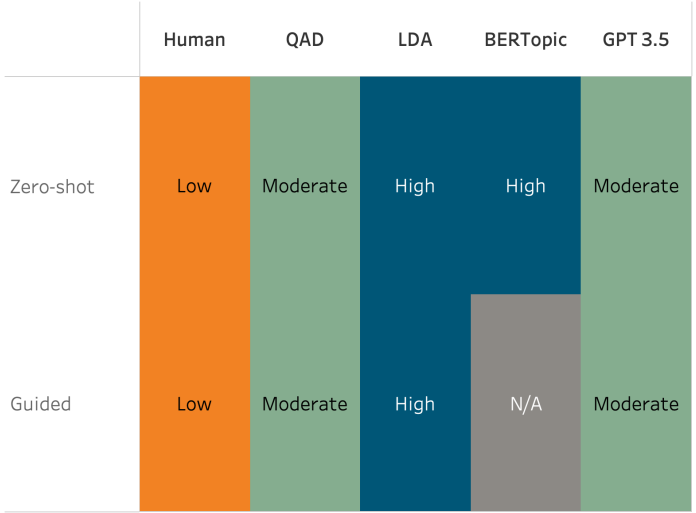
Accuracy

	Human	QAD	LDA	BERTopic	GPT 3.5
Zero-shot	High	High	Low	Moderate	Low
Guided	High	Moderate	Moderate	N/A	Moderate



Findings

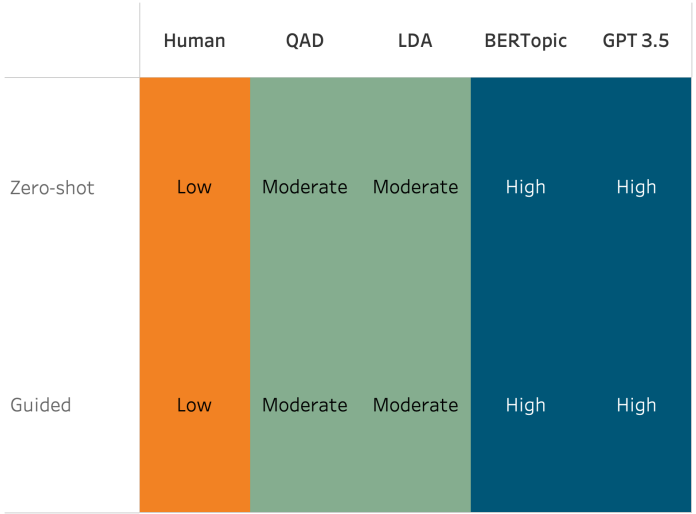
Speed



Method	Zero-shot	Guided
BERTopic	~12000/hr	N/A
LDA	~10000/hr	~10000/hr
GPT 3.5	~2500/hr	~3000/hr
QAD	~1400/hr*	~3000/hr*
Human	~550/hr	~900/hr

Findings

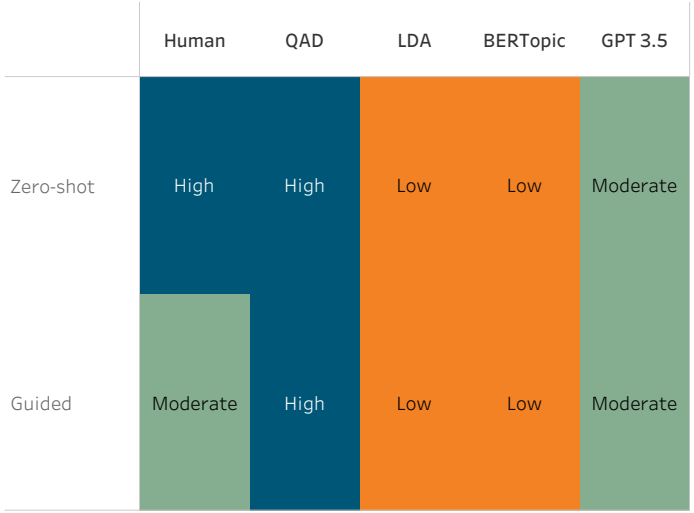
Automation



Approach	Automation
Human	None
QAD	Automated classification with human direction
LDA	Classification is automated, but human needs to select model and clustering
BERTopic	Near total automation
GPT 3.5	Near total automation

Findings

Ease



Approach	Knowledge required	Platform
Human	Minimal	Excel or NVivo
QAD	Minimal	Excel
LDA	Knowledge of Natural Language Processing (NLP) and programming	R, Python
BERTopic	Programming knowledge and understanding of BERT model	Python
GPT 3.5	Basic API knowledge	Python (for API)

Cost-effectiveness (implementation)

	Human	QAD	LDA	BERTopic	GPT 3.5
Zero-shot	Moderate	High	Moderate	Moderate	Moderate
Guided	High	High	Moderate	Moderate	Moderate

Approach	Setup Costs (estimated labour)
Human	$20-$100 depending on complexity
QAD	$20-$40
LDA	$80-$180
BERTopic	$80-$180
GPT 3.5	$80-$180

Cost-effectiveness (scaling)

	Human	QAD	LDA	BERTopic	GPT 3.5
Zero-shot	Low	Moderate	High	High	Moderate
Guided	Low	Moderate	High	High	Moderate

Approach	Commentary
Human	Almost no economies of scale
QAD	Marginal cost diminishes rapidly with scale
LDA	Near-zero fixed marginal cost
BERTopic	Near-zero fixed marginal cost
GPT 3.5	Cost is driven by token use; this is currently cheap but fixed cost

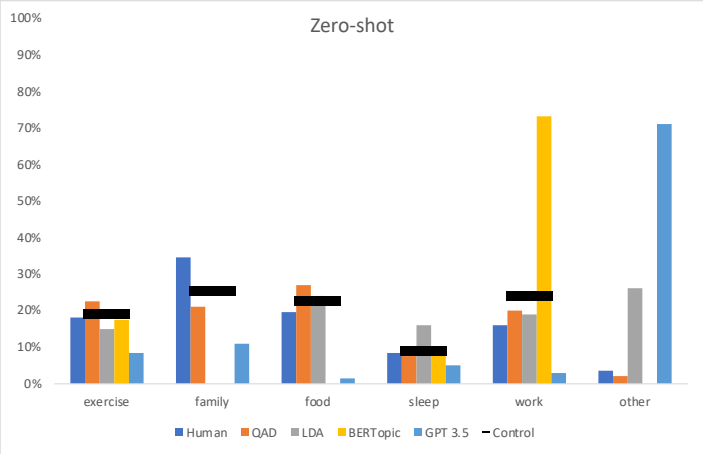
Findings

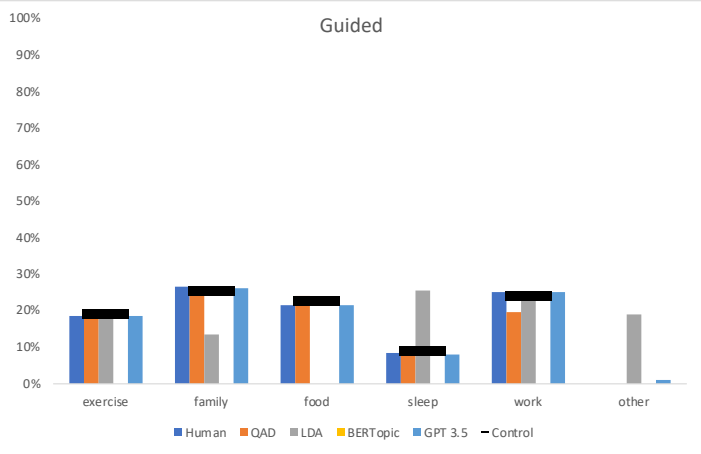
Implementation vs Scaling

	Human	QAD	LDA	BERTopic	GPT 3.5
Zero-shot	Moderate	High	Moderate	Moderate	Moderate
Guided	High	High	Moderate	Moderate	Moderate

	Human	QAD	LDA	BERTopic	GPT 3.5
Zero-shot	Low	Moderate	High	High	Moderate
Guided	Low	Moderate	High	High	Moderate

Topic prevalence tasks





Implications and next steps

Implications for practice

Machine learning approaches are cheap...

- ...but you get what you pay for

Choose the right tool for the job

- Most of these approaches have good use cases

When it comes to topic prevalence, guided QAD and GPT match human performance!

- Given that both these methods had only moderate accuracy, this is an interesting finding

Caveats and limitations

Security and privacy of data

Token limitations for APIs

Package dependency and deprecation

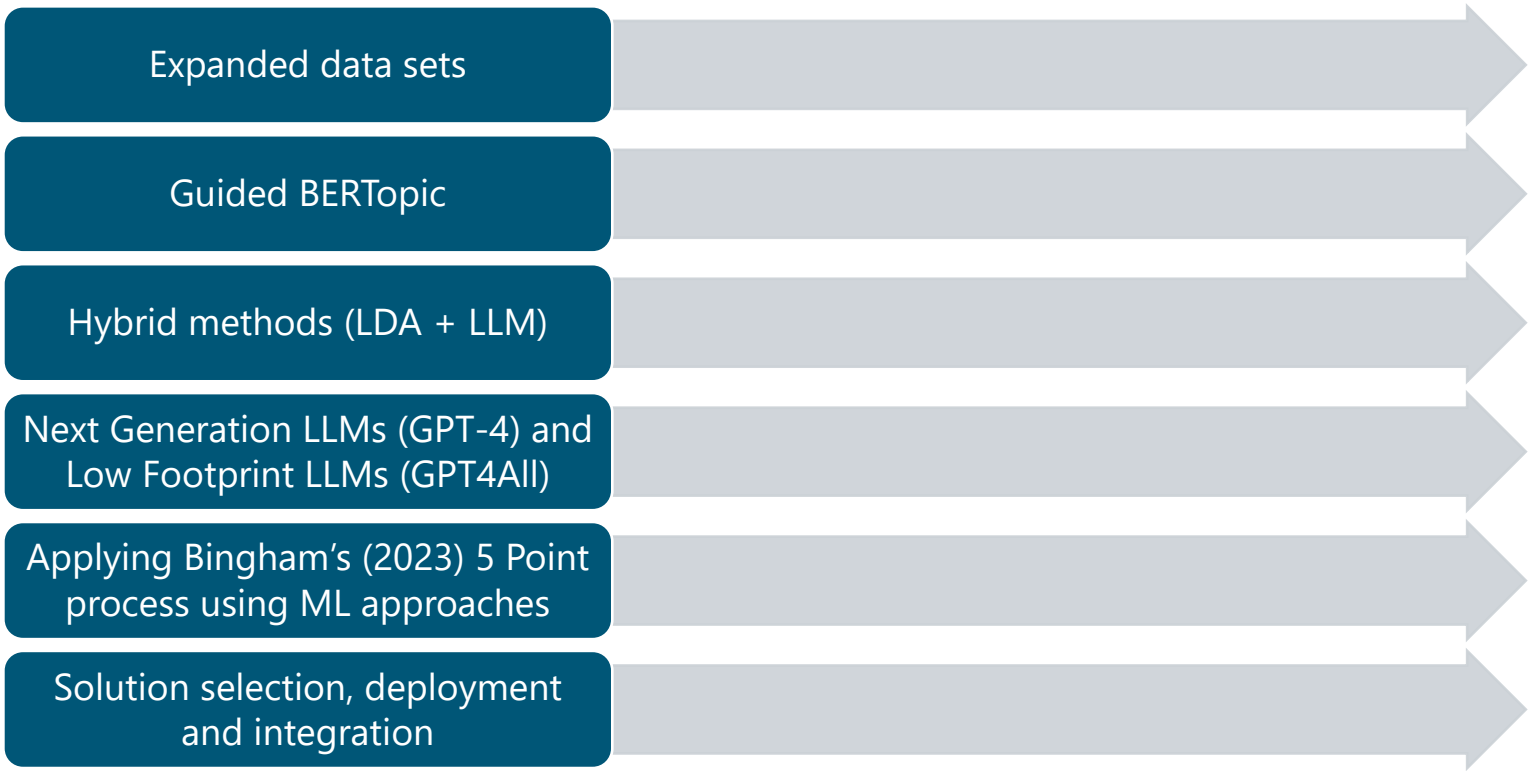
Algorithmic bias of training data

Hallucination is a problem

If you have a table of text comments in a spreadsheet, you can use the spreadsheet's built-in functions to perform the task that you want Bard to complete. For example, you could use the `VLOOKUP()` function to identify the main topic of each text comment and assign it to a specific topic category.

Once you have performed the task on the spreadsheet, you can copy and paste the results into Bard.

Next steps


Expanded data sets
Guided BERTopic
Hybrid methods (LDA + LLM)
Next Generation LLMs (GPT-4) and Low Footprint LLMs (GPT4All)
Applying Bingham's (2023) 5 Point process using ML approaches
Solution selection, deployment and integration

Confession time

Thanks!

CONSULTANTS

A quick guide to comparing selected methods for text classification in evaluation scenarios

Analysis and classification by Gerard Atkinson, ARTD Consultants (Gerard.Atkinson@artd.com.au)

	Human Coding (Zero-shot)	QAD Coding (Zero-shot)	LDA (Zero-shot)	BERTopic (Zero-shot)	GPT-3.5 (Zero-shot)	Human Coding (Guided)	QAD Coding (Guided)	LDA (Guided)	BERTopic (Guided)	GPT-3.5 (Guided)
Accuracy	High	High	Low	Moderate	Low	High	Moderate	Moderate	N/A	Moderate
Speed	Low	Moderate	High	High	Moderate	Low	Moderate	High	N/A	Moderate
Automation	Low	Moderate	Moderate	High	High	Low	Moderate	Moderate	High	High
Ease of implementation	High	High	Low	Low	Moderate	Moderate	High	Low	Low	Moderate
Efficiency of implementation	Moderate	High	Moderate	Moderate	Moderate	High	High	Moderate	Moderate	Moderate
Efficiency of scale	Low	Moderate	High	High	Moderate	Low	Moderate	High	High	Moderate