University of Sheffield

Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling

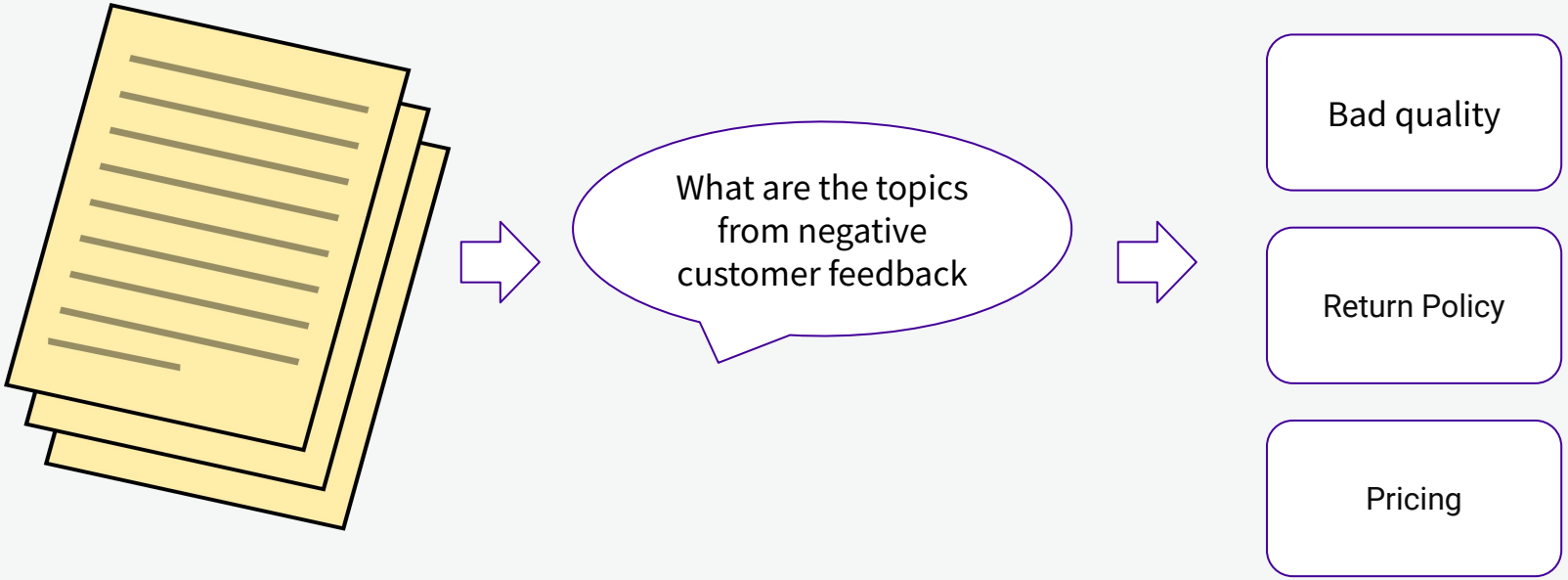
Yida Mu, Chun Dong, Kalina Bontcheva, Xingyi Song

LREC-COLING 2024

{y.mu, cdong8, k.bontcheva,x.song}@sheffield.ac.uk

Topic Analysis

- Gaining insights into topics
  - Drawing meaningful conclusions from vast amounts of text



Topic Analysis Approaches

- Probabilistic Topic Modelling (e.g. LDA)
  - Unsupervised
  - List of words to present topics
  - Difficult to control topic granularity
  - Difficult to interpretation the topic meaning

- Close-set topic classification
  - Supervised, Pre-defined a set of topics
  - Control topic granularity from the topic set
  - Difficult to discover new topics
  - Require human label

LLM Based Topic Modelling

- Open-set topic classification using pre-trained LLMs
  - Directly generate human interpretable topics
  - Open set topic classification
    - LLMs contains comprehensive global knowledge
    - Trained with billions of data

In our work: Conduct an exploratory study of open-set topic classification capability using off-the-shelf LLMs.

Experimental Settings

- Dataset: Two datasets with labelled topics
  - **20 News Group (**Lang, 1995**)**: documents labelled in 20 categories
  - **Vaccine Hesitancy Reason(**Poddar et al., 2022**)**: labelled under one of ten major vaccine hesitancy categories, such as ‘Side-effect’ and ‘Vaccine Ineffective
- Models:
  - ChatGPT3.5
  - Llama 2 Chat 7B

We also compare two traditional probabilistic topic models

- LDA and BERTTopic

Baseline Experiment

- In the baseline experiment we only instruct LLM to extract topic and the output format

<s>[INST] <<SYS>>
Read the given text and identify up to three topics, with each topic consisting of no more than three words.
Ensure that you return only the topics. The desired output format is:
Topic 1: xxx
Topic 2: xxx
Topic 3: xxx
<</SYS>>
The Given Text:
{list_of_text} [/INST]

Baseline Experiment

- In the baseline experiment we only instruct LLM to extract topic and the output format

<s>[INST] <<SYS>>
Read the given text and identify up to three topics, with each topic consisting of no more than three words.
Ensure that you return only the topics. The desired output format is:
Topic 1: xxx
Topic 2: xxx
Topic 3: xxx
<</SYS>>
The Given Text:
{list_of_text} [/INST]

System instruction

Baseline Experiment

- In the baseline experiment we only instruct LLM to extract topic and the output format

<s>[INST] <<SYS>>
Read the given text and identify up to three topics, with each topic consisting of no more than three words.
Ensure that you return only the topics. The desired output format is:
Topic 1: xxx
Topic 2: xxx
Topic 3: xxx
<</SYS>>
The Given Text:
{list_of_text} [/INST]

Format instruction

Baseline Experiment

- In the baseline experiment we only instruct LLM to extract topic and the output format

<s>[INST] <<SYS>>
Read the given text and identify up to three topics, with each topic consisting of no more than three words.
Ensure that you return only the topics. The desired output format is:
Topic 1: xxx
Topic 2: xxx
Topic 3: xxx
<</SYS>>
The Given Text:
{list_of_text} [/INST]

Input

Baseline Results

Finding 1: Granularity control is required. The model tends to return over general topics such as ‘Vaccine’, ‘COVID Vaccination’ and ‘Vaccine Hesitancy’ for Vaccine Hesitancy dataset

Finding 2: The topic naming is inconsistent. LLMs return around 2,500 extracted topics for Vaccine Hesitancy dataset (1000 samples) and most of them are near duplicates. For example:

‘side-effect’, ‘Side Effect’, ‘serious side effect’, ‘fear of side effects’ and ‘vaccine side effect’.

Adding Constraints and Postprocessing

- Based on the finding from baseline, we introduce three constraints instructions in the prompt and a set of hand-crafted rules to merge near duplicate topics.
  - Black list: A set of example topics to guide model not return over general topics
  - Seed topics: A set of example topics to guide to of the desired granularity and naming
  - Granularity_instruction: A description of topic granularity, such as ‘related to COVID-19 vaccination hesitation’
  - hand-crafted rules: Merge topics based on the lemmatisation, case and remove symbols.

Adding Constraints to the Prompt

<s>[INST] <<SYS>>
Consider the existing topics: {Seed_Topics}.
Read the given text and identify up to three topics {Granularity_instruction}, with each topic consisting of no more than three words.
Avoid general topics such as {Black_List}, which are already known.
Ensure that you return only the topics. The desired output format is:
Topic 1: xxx
Topic 2: xxx
Topic 3: xxx
<</SYS>>
The Given Text:
{list_of_text} [/INST]

Constraints

Constraints Results

- Top 10 Topics generated by GPT on the vaccination hesitation dataset

Baseline Prompt	**vaccine**, **covid vaccine**, vaccine effectiveness, **covid 19 vaccine**, vaccine safety, vaccine development, vaccine side effects, side effects, **covid, covid 19**
+Rules	vaccine hesitancy, vaccine safety, vaccine efficacy, vaccine effectiveness, vaccine side effect, hesitancy reason, side effect, hesitancy, trust issue, hesitancy factor

Constraints Results

- Top 10 Topics generated by GPT on the vaccination hesitation dataset dataset

Baseline Prompt	**vaccine**, **covid vaccine**, vaccine effectiveness, **covid 19 vaccine**, vaccine safety, vaccine development, vaccine side effects, side effects, **covid, covid 19**
+Rules	vaccine hesitancy, vaccine safety, vaccine efficacy, vaccine effectiveness, vaccine side effect, **hesitancy reason**, side effect, **hesitancy**, trust issue, **hesitancy factor**
+Seeds	vaccine safety, vaccine effectiveness, trust issue, vaccine hesitancy, vaccine efficacy, trust in vaccine, side effect, trust, safety, vaccine side effect

Topic Summarisation

- Merge topics into N classes

```
<s>[INST] <<SYS>>
Summarize and merge the following list of topics into {Fixed_Number} of final topics
<</SYS>>
{list_of_topics} [/INST]
```

Summarisation Results

- Top 10 Topics summarised by GPT on the VAXX dataset

Basic Prompt	**vaccine**, **covid vaccine**, vaccine effectiveness, **covid 19 vaccine**, vaccine safety, vaccine development, vaccine side effects, side effects, **covid, covid 19**
+Rules	vaccine hesitancy, vaccine safety, vaccine efficacy, vaccine effectiveness, vaccine side effect, **hesitancy reason**, side effect, **hesitancy**, trust issue, **hesitancy factor**
+Seeds	vaccine safety, vaccine effectiveness, trust issue, vaccine hesitancy, vaccine efficacy, trust in vaccine, side effect, trust, safety, vaccine side effect
LLM Summarise	Vaccine Safety and Efficacy, Vaccine Hesitancy and Trust, Access and Distribution of Vaccines, Government and Regulatory Influence, Vaccine Development and Testing, Adverse Reactions and Side Effects, Public Perception and Misinformation, Immune System Response and Effectiveness, Ethical Concerns and Transparency, Long-Term Impact and Future Immunity

Metrics

1. Granularity over Top N

We hypothesise that an increased number of final **N** topics results in decreased granularity (i.e., higher semantic similarity).

2. Recall

$$\text{Recall} = \frac{\text{No. Correct Extracted ST Samples}}{\text{No. Seeds Topic Samples}}$$

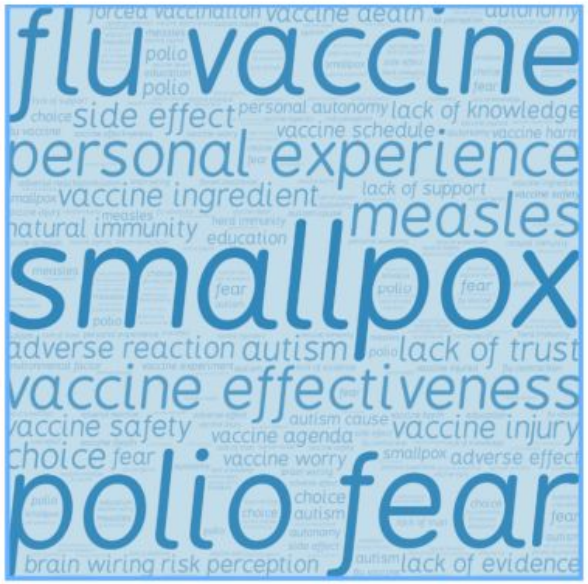
3. Precision

$$\text{Precision} = \frac{\text{No. Correct Extracted ST Samples}}{\text{No. Samples ST Extracted}}$$

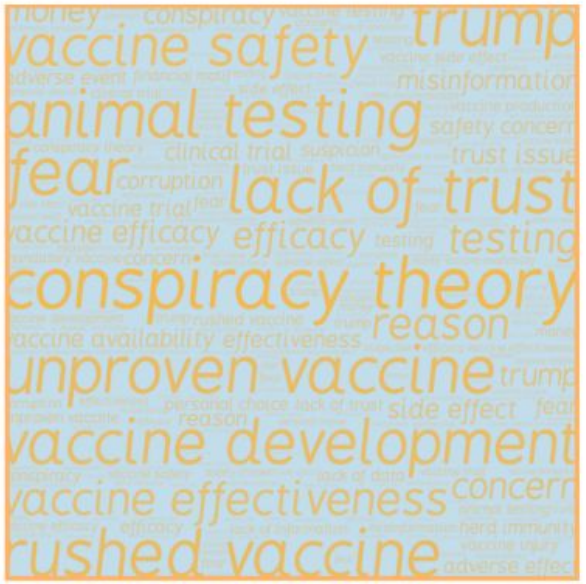
Results

	A	B	C	D	E	F	G	H
1	**LLM**	**Prompt Strategies**	**Vaxx**			**20NG**		
2			**Granularity Top 10 / 20 / 30**	**Recall**	**Precision**	**Granularity Top 10 / 20 / 30**	**Recall**	**Precision**
3	*GPT*	*Basic Prompt*	0.279 / 0.516 / 0.590	-	-	0.146 / 0.175 / 0.175	-	-
4	*GPT*	*Manual Instructions*	0.297 / 0.584 / 0.668	-	-	0.136 / 0.136 / 0.151	-	-
5	*GPT*	*Seed Topics*	0.282 / 0.259 / 0.564	0.716	0.488	0.132 / 0.145 / 0.159	0.171	0.351
6	*LLaMA*	*Basic Prompt*	0.320 / 0.662 / 0.679	-	-	0.392 / 0.318 / 0.317	-	-
7	*LLaMA*	*Manual Instructions*	0.341 / 0.541 / 0.545	-	-	0.182 / 0.188 / 0.230	-	-
8	*LLaMA*	*Seed Topics*	0.306 / 0.223 / 0.218	0.773	0.496	0.327 / 0.157 / 0.191	0.230	0.342

Case Study:Temporal Analysis of COVID-19 Vaccine Hesitancy



(a) Pre COVID-19



(b) COVID-19 Vaccine Development Period

(c) After the First Jab of COVID Vaccine

Future Works

- Fine-tuned model (such as LLaMA and Mistral)
- Larger Scale Experiment (on Wikipedia and Bills datasets)

Future Works

- Fine-tuned LLMs can generate significant less near-duplicate topics compared with off-the-shelf LLMs
- Also, less hallucinated topics…

Pre-print

Addressing Topic Granularity and Hallucination in Large Language Models for Topic Modelling

Code and Data:
https://github.com/GateNLP/TopicLLM_Granularity_Hallucination

arXiv: https://arxiv.org/abs/2405.00611

big data and cognitive computing

an Open Access Journal by MDPI

IMPACT FACTOR 3.7

CITESCORE 4.9

Natural Language Processing Applications in Big Data

Guest Editors

Dr. Xingyi Song, Dr. Ye Jiang, Dr. Yunfei Long

Deadline

31 December 2024

Special Issue

mdpi.com/si/199360

Invitation to submit