Unravelling Media Impact on Public Discourse: An Examination of News Framing and Sentiment Analysis using NewsTalk

Robert Fleet – Queensland University of Technology

eResearch Australasia 2023

Introduction

CRICOS No.00213J

Data Scientist/Developer at the QUT Digital Observatory

Background in Criminology and Forensic Science

Work primarily with Human Data on the internet and elsewhere

Currently exploring the application of AI pipelines for research

Also exploring how to gather data in post API “Golden Age” and across multiple platforms

Research Space

A progressive shutdown of easy/free access to social media and media API for research

An unfolding response to AI/LLM web scraping for training data by limiting access to websites for scraping and enforcing terms of service

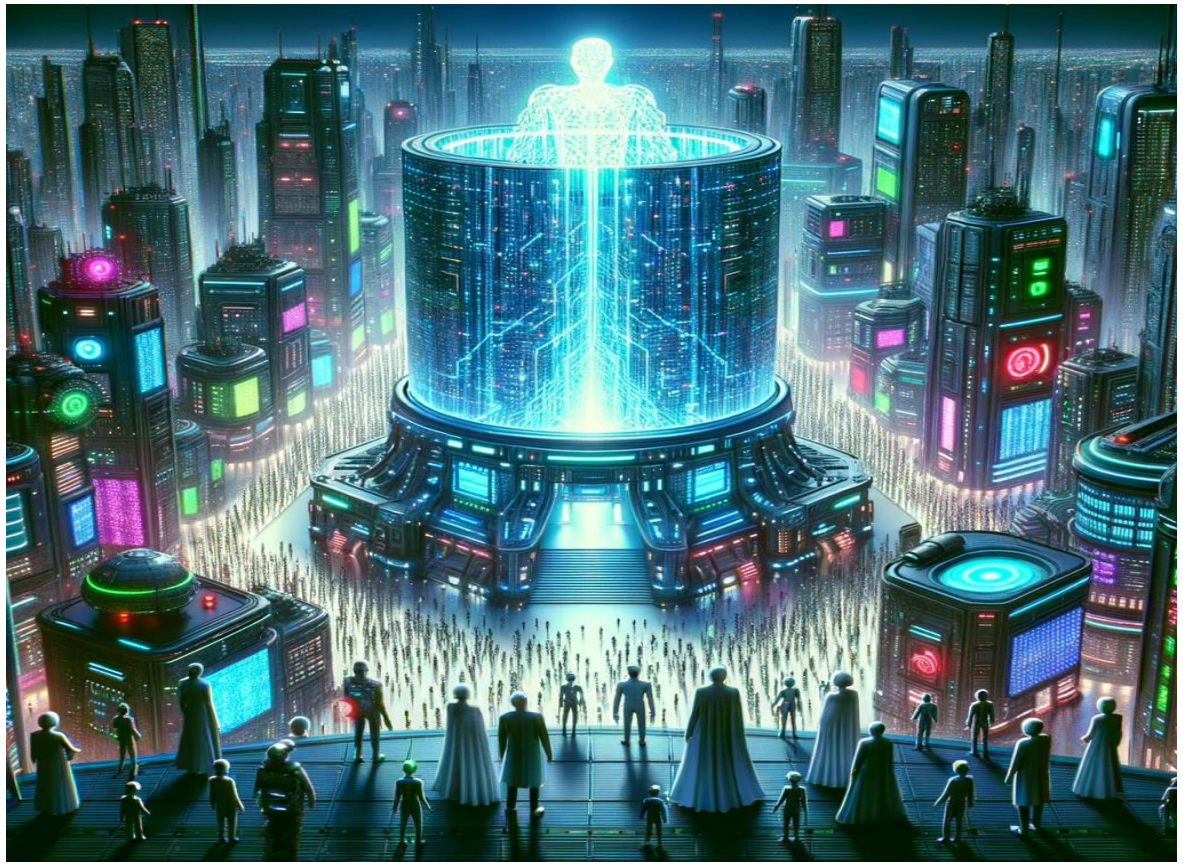
Big data methods no longer applicable in some cases

Research Space

There is research utility in bringing together a text and the commentary on or responses to that text – what discourse is being generated by the public

There is research utility in bringing together multiple locations of discourse on a topic even if those loci are small sets of data – where are the public talking about the topic

Rise of AI



Concurrently, we have seen the rise of AI/LLM based products that have generated a lot of excitement

LLMs do have great research potential

However, they aren’t science fiction either

There are some reality checks for using LLMs

Research Problem

- 2nd level agenda setting extraction from news stories and analysis of the associated commentary
- Scale – 1610 comments on one story
- Time consuming though not very sophisticated
- A person would need to read the article and all the comments to be able to make a suitable summary statement
- Build a system that leverages LLM and traditional NLP to solve this problem

Approach

Hybrid python pipeline

NLP based named entity extraction – spacy

Human intervention on choosing named entities (rank or interest)

LLM API driven vectorstore index queries

The data source – NewsTalk – brings together the news source and the news commentary in a single location

Strengths

- NER is deterministic and gives known entities with labels
- Queries are expressed in “natural language”
- Some control of the steps used in analysis
- LLMs give “natural language” answers
- Using an index approach allows for answers based on the provided documents – less scope for LLM “confabulation”

Limitations and Risks

- In many ways it remains a black box
- Answers are non-deterministic (never the same answer)
- LLMs are often wrong but never in doubt
- Context is always limited
- Prompt Engineering

Ethics

- Copyright content
- Implied consent from commentors
- Sensitive data
- Public facing data

Data sovereignty

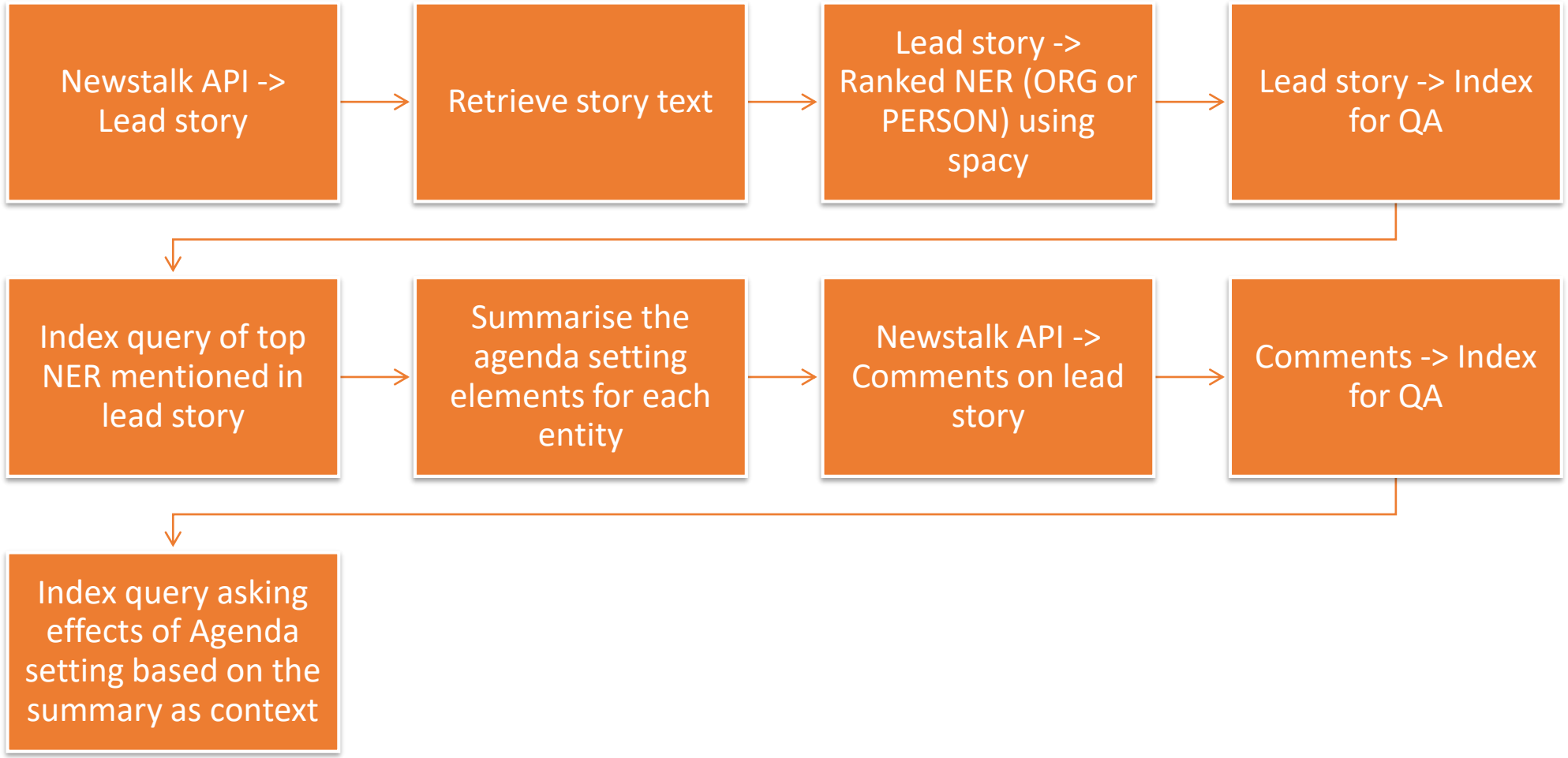
Where does the data reside

Is the data being used to train models by private API providers

Lack of infrastructure

- Lack of research services supporting local LLM
- Not always suitable for batch job based HPC
  - HPC resources may lack access to the outside world so cannot call API
- Higher performance open-source models are GPU/Memory intensive

Pipeline



Prompt – 2nd level Agenda Setting

- “Analyse the provided news article to identify second-level agenda-setting elements for each object, focusing on [labor]: Tone: Is the tone towards [labor]; positive, negative, or neutral? Frame: What aspects related to [labor] are emphasized? Themes: Are there recurring themes around [labor]? Language: Are specific words used to influence opinions on [labor]? Summarize your findings."

Results - Labor

- The tone towards Labor in the article is neutral. The article emphasizes Labor's primary vote, the performance of Prime Minister Anthony Albanese, and support for the Indigenous Voice. The recurring theme around Labor is that their primary vote has been decreasing since the May budget. The language used to influence opinions on Labor is not specified in the article.

Results - Coalition

- The tone towards the Coalition in the article is neutral. The article emphasizes the Coalition's primary vote increasing from 33 to 34 percent, and Peter Dutton's net performance rating improving from -17 to -8 percent. The recurring theme is that the Coalition's support is increasing. The language used to describe the Coalition is factual and does not attempt to influence opinion.

Results – the Voice

- The tone of the article towards the Voice is mostly negative. The article emphasizes the growing opposition to the Voice, the increasing frustration of respondents, and the lack of success of the Yes campaign. The recurring themes around the Voice are that it is divisive, a diversion, and unpopular. The language used to influence opinions on the Voice is mostly negative, with words like "exasperated," "frustrated," "diversion," and "divisive" used to describe it.

Prompt - Comments

- Given the context: *Overall, the tone of the article is neutral towards Labor and the Coalition, but negative towards the Indigenous Voice. In summary, the article presents a neutral tone towards Labor and the Coalition, with a focus on their primary vote and support for the Indigenous Voice. However, the tone towards the Voice is mostly negative, highlighting growing opposition and frustration among respondents. The language used to describe the Voice is also negative, portraying it as divisive and unpopular. Overall, the article presents a balanced view of the political landscape, but with a clear bias against the Indigenous Voice*. Analyse the comments responses to the article with regard to how the comments have been influenced by the agenda setting. Summerise your findings."

Results

- The article's second-level agenda setting has had a clear influence on public discourse, as evidenced by the comment responses. The negative portrayal of the Indigenous Voice has had a particularly strong impact on the sentiment and themes present in the comments.

- Many of the comments express frustration and disappointment with the lack of bi-partisan support for the Voice, and the negative campaigning by the Coalition

- There is also a sense of despair that the Voice may not be accepted, and that Australia will look silly on the international stage if it is rejected. The comments also reflect a sense of anger and frustration at the lack of progress on Indigenous matters, and the refusal of the LNP leadership to say sorry.

- Overall, the comments demonstrate how the article's framing of the Indigenous Voice has had a significant impact on the tone, sentiment, and thematic focus of the comment responses.

Conclusion

- Set out to build a pipeline to analyse articles and linked comments
- The results are on par with what an RA might be able to do
- This provides a useful and potential time saving tool to free up a researcher to focus on more ambitious research questions

Conclusion

CRICOS No.00213J

- However:
- There is a need to make sure that the system provides links back to the raw data
- And that the system can extract canonical examples from the raw data
- Not always accepting results at face value
  - Guiderails
  - Fact checking
  - Internal consistency

- There are also some commercial platform providers in the early stages of providing services that adhere to data management standards

Conclusion

- Early Days
- need for further research on the range of practical methods for hybrid LLM research data pipelines in:
  - HPC
  - local
  - secure cloud
  - tech giant APIs.
- Appropriate solutions will vary according to the specific data.

Thank You

- Ask us about access to Newstalk
- https://newstalk.digitalobservatory.net.au/

- Contact Us
- https://www.digitalobservatory.net.au/
- digtialobservatory@qut.edu.au