News in the time of COVID
Creating a high-frequency news sentiment measure for New Zealand using GDELT

Christopher Ball    Reserve Bank & Stats NZ
25 August, 2020
Outline

1. News sentiment
2. Source data
3. Data processing
4. Series construction
5. Comparisons to other sentiment measures
6. Summary & future applications
News sentiment
What have others done with GDELT?

Mentions Of “Stress” Increase To Highest Levels In Decade

RealClear Public Affairs
What have others done with GDELT? (II)

Climate Change – On The Rise Since Mid-2018 – Hits 3.5-Year Low

RealClear Public Affairs

News in the time of COVID | Christopher Ball
Page 3/29
What have others done with GDELT? (III)

BBVA: Tracking Chinese Vulnerability in Real Time Using Big Data

Chinese Vulnerability Sentiment Index (CVSI) combining key indicators with big data

Evolution of the “tone” of the main themes on vulnerability in China. Higher values indicate an improvement in sentiment and lower vulnerability.
What have others done with GDELT? (IV)

Gauging the State of the Economy with News Narrative and Sentiment

![Graph showing time series data from January 2020 to June 2020 with net sentiment, confidence, and financial uncertainty trends]

News in the time of COVID  |  Christopher Ball
Page 5/29
SOURCE DATA
What is GDELT?

Supported by Google Jigsaw, the GDELT Project monitors the world’s news and identifies the people, locations, organizations, themes, sources, emotions and events driving our global society every day, delivered on a free open platform.

Shortened from https://www.gdeltproject.org/
But I want details!

• GDELT updates every 15 minutes.

• Realtime Translation of 65 Languages, representing 98.4% of its daily non-English monitoring volume.

• Realtime Measurement of 2,300 Emotions and Themes, including natively assessing the emotions of 15 languages.

• Tracking Event Discussion Progression, which allows the emotional reaction and resonance of an event to be assessed as it sweeps through the world’s media.

Shortened from https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/
What does a typical record look like?

Christopher Ball
What does a typical record look like?

Part 2
News in the time of COVID

What does a typical record look like?

Part 3
What does a typical record look like?
Part 4

https://resources.stuff.co.nz/content/dam/images/y/m/y/z/w/image.related.StuffLandscapeSixteenByNine.1420x800.4ynjvw.png/1596155347262.jpg

I have been working on the details of a solution that pretty much there x2013; more to say very soon ..#2008|x31|one of those outstanding issues #2064|35||resolved by the previous Government#2672|47||park-and-ride system on the old salp yard site#3162|26||just front up and sort out Christchurch Hospital, 53; Christchurch Regeneration Minister Megan Woods, 196; Christchurch Hospital, 396; Hagley Ave, 906; Canterbury District Health Board, 1126; Deans Ave, 1280; Gerry Brownlee, 1833; Judith Collins, 2233; Christchurch Hospital, 2299; Deans Ave, 2657; Hagley Ave, 2961; Hagley Park, 3223 350, car parks on Deans, 1034; <PAGE_LINKS>https://www.stuff.co.nz/business/105190635/multiple-hospital-parkandrides-can-only-be-positive--cdhb; https://www.stuff.co.nz/national/health/121977340/christchurch-hospital-car-parking-likely-heading-back-to-old-deans-ave-site; https://www.stuff.co.nz/national/health/122333626/national-leader-judith-collins-makes-confusing-christchurch-hospital-car-parking-promise; https://www.stuff.co.nz/national/health/89425133/parking-crisis-at-christchurch-hospital</PAGE_LINKS><PAGE_PRECISEPUBTIMESTAMP>20200730093200</PAGE_PRECISEPUBTIMESTAMP> <PAGE_TITLE>Christchurch Hospital car park plan to be released within weeks - Megan Woods</PAGE_TITLE>
A more useful way of looking at these data

<table>
<thead>
<tr>
<th>Difficulty</th>
<th>Variables</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>Source, tone, word count</td>
<td>Comma delimited</td>
</tr>
<tr>
<td>Medium</td>
<td>Locations, People, Organisations, Themes</td>
<td>Nested semi-colon delimited list</td>
</tr>
<tr>
<td>Hard</td>
<td>Alternative sentiment measures, alternative thematic coding</td>
<td>Key-value pairs, semi-colon delimited</td>
</tr>
</tbody>
</table>
Themes

11.3 million coded themes in 2020 with 12,000 distinct themes

<table>
<thead>
<tr>
<th>Level 1</th>
<th>Count</th>
<th>Level 2</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAX</td>
<td>3,598,756</td>
<td>TAX FNCACT</td>
<td>1,852,791</td>
</tr>
<tr>
<td>WB</td>
<td>3,040,928</td>
<td>TAX ETHNICITY</td>
<td>501,649</td>
</tr>
<tr>
<td>EPU</td>
<td>810,019</td>
<td>EPU POLICY</td>
<td>456,836</td>
</tr>
<tr>
<td>CRISISLEX</td>
<td>507,887</td>
<td>TAX DISEASE</td>
<td>452,816</td>
</tr>
<tr>
<td>SOC</td>
<td>334,491</td>
<td>TAX WORLDLANGUAGES</td>
<td>285,111</td>
</tr>
<tr>
<td>UNGP</td>
<td>284,897</td>
<td>SOC POINTSOFFINTEREST</td>
<td>257,295</td>
</tr>
<tr>
<td>ECON</td>
<td>267,529</td>
<td>EPU ECONOMY</td>
<td>191,049</td>
</tr>
<tr>
<td>GENERAL</td>
<td>255,816</td>
<td>EPU CATS</td>
<td>147,321</td>
</tr>
<tr>
<td>USPEC</td>
<td>196,675</td>
<td>GENERAL HEALTH</td>
<td>140,212</td>
</tr>
<tr>
<td>MEDICAL</td>
<td>109,210</td>
<td>WB 621</td>
<td>131,615</td>
</tr>
</tbody>
</table>
Social Network Analysis - 2020 Themes
What is “tone”?

• The metadata states the tone measures the difference between the % of positive words and the % of negative words (where words are classified as positive, negative or neutral).

• The metadata doesn’t state how words are classified as positive, negative or neutral. Earlier GDELT blog posts suggest that the measure is the same one used in the lead developer’s early work.

• Alternative tone measures with better documentation are provided in GCAM, the hard part of the source data to process.
DATA PROCESSING
How much data?

- Each 15 minute block is supplied as a 5MB to 50MB compressed zip file, which extracts to a 20MB to 200MB csv file.

- A weekly update requires downloading about 10GB of data, and processing about 30GB. A full year requires processing more than 1TB.

- Only a small fraction of the articles reference New Zealand.
Processing

• Historical files are not revised, so we can download the continuously updated master file list and check if the file has already been processed.

• Use Python to stream each 15 minute files line-by-line, check if article is relevant to New Zealand and if so extract the relevant fields to a new csv file (also written line-by-line).

• Can be run asynchronously to improve performance. Typically limited by internet download speed, but could also be CPU or read/write limited with a good enough connection.
Other notes

• News is frequently syndicated, with publication of essentially the same article on a number of affiliated websites on the same day. These articles are compressed to the first publication of a given article.

• Ninja edits are also an issue with online news publication. Edits which don’t affect the tone will typically be picked up when dealing with syndication.

• As far as I can tell from the metadata, continuously updating stories will be picked up in each 15 minute batch and if the tone changes they will show up as multiple articles. It may be possible to use the source url to deal with this type of publication, although taking the first occurrence may not be the best approach.
Series construction
Weighted mean

Article Weight = (Relevant themes as % of total themes) \times (New Zealand as % of total locations)

\[
\text{Theme Tone} = \frac{\sum (\text{Weight} \times \text{Tone})}{\sum \text{Weight}}
\]

The economics series also filters to articles with a relevant themes more than 15% of the total themes present in the article.
How are “relevant themes" determined?

• Start by building a list of (root) keywords for the topic of interest, e.g. Econ, Finance, Market.

• Look through the list of themes which text match the keywords to ensure the matches are accurate.

• Find the most frequent themes which occur in articles containing the keyword list. Decide whether to include these in the keyword list.

• Repeat this process until you are satisfied with the keyword list.
Final economic news sentiment series

Themes: ECON;FINANC;UNEMPLOY;GOVERNOR;MARKET;AUSTER;DISEASE;CORONA;COVID

28 day rolling average, no seasonal adjustment
COMPARISONS TO OTHER SENTIMENT MEASURES
Other measures of uncertainty and confidence

There are a number of other measures of uncertainty and confidence in New Zealand, although they are provided on a monthly or quarterly basis and with a significant lag compared to GDELT.

- Monthly ANZ Business Outlook - Questions
- Monthly ANZ Roy-Morgan Consumer Confidence
- Quarterly Westpac McDermott Miller Consumer Confidence
- Quarterly Survey of Business Opinion - Questions
Comparisons to ANZ Business Outlook

ANZ Business Outlook

Series: business confidence, activity outlook, profits, capacity utilisation

Date
Normalised value
ANZ Business Outlook

News in the time of COVID | Christopher Ball
Page 23/29
Comparisons to ANZ Roy-Morgan

ANZ Roy-Morgan Consumer Confidence

Series: current conditions, financial conditions, short-term economic outlook

Date
Normalised value

2016 2018 2020

News in the time of COVID
Christopher Ball
Page 24/29
Comparisons to Westpac MM

Westpac McDermott-Miller Consumer Confidence

Series: overall, profitability and domestic trading activity

Date
Normalised value
Comparisons to QSBO

Quarterly Survey of Business Opinion

Normalised value

Date

2016 2018 2020

Series: overall, profitability and domestic trading activity

News in the time of COVID | Christopher Ball
Page 26/29
SUMMARY & FUTURE APPLICATIONS
Summary

• Present a high-frequency low-lag measure of sentiment from GDELT’s news database

• It compares well to existing lower frequency measures of economic confidence (particularly consumer confidence), although it is an independent measure of sentiment.

• Only skimming the surface so far. There is a lot of extra information on source, people, organisations and themes within the GDELT. There are also other collections like the Global Front Page Graph which could provide a better measure of “importance".
Future applications

• Sentiment series for other concepts like employment, racism, xenophobia or other topics and themes commonly reported in the news.

• Social network analysis of the organisations or people mentioned in the news.

• GDELT has captured the news on the “front page" of most major news publications since 2017. This could be used to create a prominence or importance measure to supplement the current approach which essentially looks at volume.
People driving NZ news

News in the time of COVID  |  Christopher Ball
Page 29/29
Fin
APPENDIX SLIDES
Employment series

A higher value indicates more negative sentiment

News in the time of COVID | Christopher Ball
Page 1/3
ANZ Roy-Morgan Questions

1. Would you say you and your family are better off financially or worse off than you were at this time last year?
2. This time next year do you and your family expect to be better off financially or worse off than you are now?
3. Thinking of economic conditions in New Zealand as a whole, in the next 12 months, do you expect we’ll have good times financially, bad times or some good and some bad?
4. Looking ahead, what would you say is more likely: that in New Zealand as a whole we’ll have continuous good times during the next five years or so, we’ll have bad times, or some good and some bad?
5. Generally, do you think now is a good time, or a bad time, for people to buy major household items?
6. During the next 2 years do you think that prices in general will go up, go down, or stay where they are now? And if up, what is the expected percentage per year?
7. Specifically thinking about the price of houses during the next 2 years, do you think that the price of houses in general will go up, go down, or stay where they are now? And if up, what is the expected percentage per year?
The Westpac McDermott Miller Consumer Confidence Index summarises the net balance of optimistic/pessimistic responses to five questions: how households’ financial situation has changed over the past year; whether now is a good time to buy a major household item; how households expect their financial situation to change over the coming year; and near term and longer-term prospects for the New Zealand economy as a whole.

The first two of these questions are summarised in the Present Conditions Index, and the last three are summarised in the Expected Conditions Index. An index number over 100 indicates that optimists outnumber pessimists, though the series may be above or below 100 on average.