Uncovering hidden content in Australian newsprint articles on PNG

Author : Anthony Swan

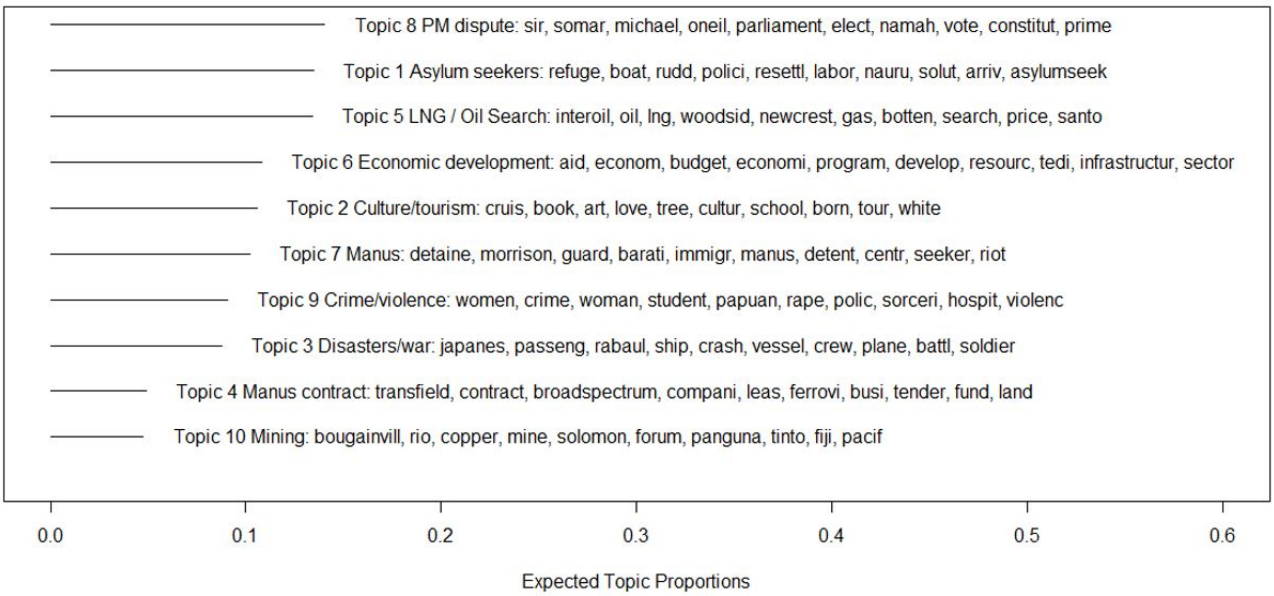
Date : August 23, 2016

Our lives are becoming increasingly saturated with text. Words convey information on events, opinions, ideology, preferences, ideas, and motivations – some of it useful and perhaps a lot of it not. The sheer volume of text poses problems on two fronts: important pieces of text can easily be missed or hard to find, and understanding the relationship of context across potentially hundreds or thousands of text-based documents can be mind-boggling. However, new advances in automated approaches to text analysis (also known as text mining) can help tackle these challenges. In this two-part blog post I highlight the strengths and limitations of text mining by drawing on my analysis of more than 2000 Australian newsprint articles (5 years' worth) focused on Papua New Guinea and other examples in the literature relevant to development policy.

One of the strengths of text mining is the ability to discover topics in texts. In qualitative research, for example, many questions are open-ended and it can be very time consuming to code the topic(s) that respondents focus on in their answers. For my corpus of newsprint articles on PNG, I use Structural Topic Modelling (STM) which is a relatively new approach for automating the discovery of topics in texts and estimating the probability that an article is related to a particular topic. It follows that STM does not assume that each article belongs to just one topic but potentially any number of topics based on these estimated probabilities. While there are alternatives to STM in the field of topic modelling (such as latent Dirichlet allocation), STM has the distinct advantage of being able to control for other variables of interest, such as time or treatment effects, and perform hypothesis testing based on these covariates.

In STM estimation (and most other forms of topic modelling), the total number of topics is user determined. The STM procedure outputs high probability terms associated with each topic and the user is required to interpret what each topic is about based on these terms. The figure below shows the results of STM estimation on my corpus of Australian newsprint articles on PNG based on a total of 10 topics: I have labelled each topic based on the associated terms for each topic (also shown) and the topics are ranked in terms of the estimated prevalence of each topic (i.e., the expected topic proportion) across the corpus.

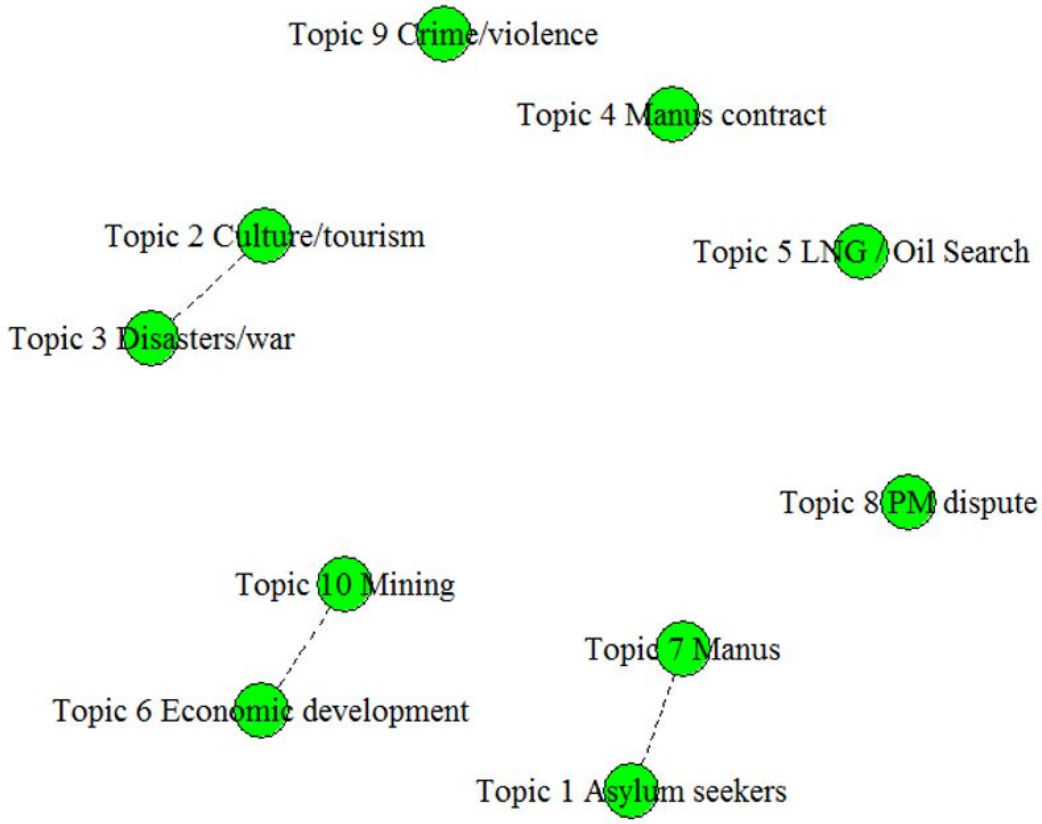
Australian newsprint topics on PNG: estimated topic proportions



Click to enlarge

The most prevalent topic is the dispute over the prime ministership between Sir Michael Somare and Peter O'Neill (Topic 8), which has an average topic proportion of nearly 15%, closely followed by the asylum seeker and LNG/Oil Search topics (Topics 1 and 5, respectively). Economic development (Topic 6), culture/tourism (Topic 2), and Manus Island (Topic 7) are also relatively prevalent topics with an average topic proportion of slightly more than 10% each.

The relationships between the estimated topics can also be easily displayed. The figure below connects topics that include terms that are positively correlated with each other; connected topics imply that those topics tend to be discussed within the same article.



At this stage it is useful to verify the classification of topics based on STM. A good first step in this process is to (manually) read articles that are predicted to be most relevant to each topic. Below, I show the headings of these articles for four topics of interest.

Topic 8: PM dispute

ASIA-PACIFIC Respite in PNG's reign of confusion -- 31/12/2011

PNG court restores Somare -- 13/12/2011

PNG high court dismisses PM -- 13/12/2011

Moresby mayhem Rival governments digging in -- 15/12/2011

Somare defiant but rival O'Neill has the support -- 17/12/2011

Topic 6: Economic development

Solar power could be answer to PNG electricity crunch -- 12/08/2015

PNG Vietnam face services test -- 12/08/2013

Bank chief and PM see resources slump as great opportunity for PNG -- 14/08/2015

PNG hailed for looking past mining -- 4/07/2013

PNG to spend the spoils of gas boom -- 22/11/2013

Topic 1: Asylum seekers

Labor's head in the sand on Pacific detention -- 22/11/2012

UN slams plan for asylum seekers -- 27/07/2013

Whatever capacity required will be built for detention PNG SOLUTION - ELECTION 2013 - -- 8/08/2013

Work rules for asylum seekers to be relaxed -- 27/11/2012

Unapologetic Rudd anticipates legal challenge to new policy -- 23/07/2013

Topic 9: Crime/violence

Accused priest preaches in PNG -- 1/12/2012

Crowd tries to burn women alive -- 14/02/2013

PNG women in Easter torture -- 6/04/2013

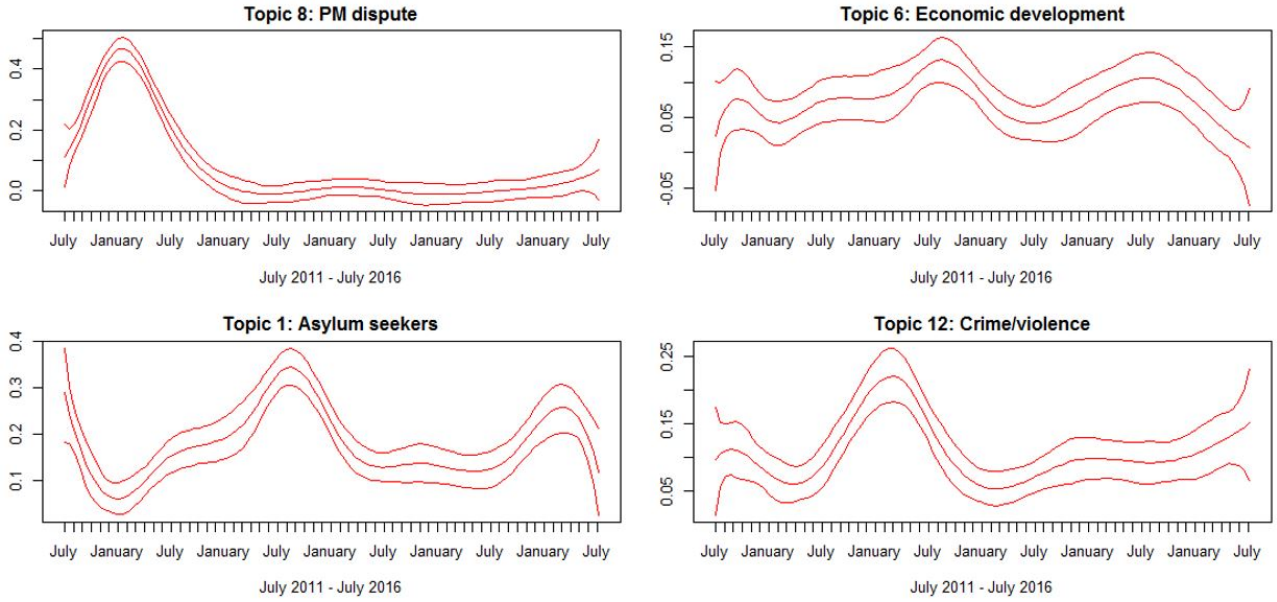
Action urged on sorcery killing spree -- 10/02/2013

PNG women tortured beheaded -- 9/04/2013

Click to enlarge

How does the estimated prevalence of these topics vary over time?

Below I plot expected topic proportions smoothed over time (the middle red line) as well as the 95% confidence interval (bounded by the outer red lines). The PM dispute topic, for example, peaks at an expected topic proportion of near 50% at the beginning of 2012 (i.e., around half the content of articles were devoted to this topic at this time) but drops to around zero by 2013. The asylum seeker topic has a peak around the time of the Australian Federal election in 2013 and remains an ongoing issue of importance. Similarly, economic development and crime/violence topics hold their importance over time, although there is a noticeable peak for the crime/violence topic reflecting a spate of sorcery related violence in PNG in 2013.



In a follow-up blog post I will show how the STM approach can be used to estimate treatment effects on topic proportions, which is an incredibly useful tool when using qualitative research methods to conduct experiments. In the meantime, it is useful to highlight underlying assumptions of these methods and the limitations they imply.

A major limitation is that that text mining usually involves focusing on just keywords – punctuation and uninformative words are removed and remaining words are stemmed (so that “govern” and “governing”, for example, are treated identically). Importantly, this approach means that the order of words is ignored.

This is a problem when the focus of text mining is to try to understand the meaning of text or sentiments being conveyed. Sentiment analysis, for example, attempts to identify and categorise opinions contained in text by measuring the relative prevalence of words that reflect a positive or negative attitude towards a subject. Despite the growing popularity of sentiment analysis, ignoring the order of words, as well as the presence of sarcasm or double entendre, means that this aspect of text mining still faces major challenges. That being said, topic modelling is not particularly susceptible to these particular problems, although there are other challenges (as summarised here [pdf]), including validation of results to demonstrate that topics are assigned in a meaningful way.

Anthony Swan is a Research Fellow at the Development Policy Centre.