Critical Reflections on Methodology and Technology: Gamification, Text Analysis and Data Visualisation
Edited by D Birks *et al*
Compilation ©2013 Association for Survey Computing

Exposing and Quantifying Narrative and Thematic Structures in Well-formed and Ill-formed Text

Dale Chant, Red Centre Software Pty Ltd

Abstract

This paper applies some of the techniques of time series analysis, as developed over the past few decades for longitudinal (tracking) studies of consumer behaviour, to various bodies of text. The intention is to expose and quantify narrative and thematic structures – do they exist, and if so, to what extent do they dominate or share the text space? The methodology is first verified by application to two well-formed texts: *Romeo and Juliet* and *Wuthering Heights*. Once verified in these ideal scenarios, the methodology is then extended to ill-formed texts, specifically Twitter interactions.

For both well-formed and ill-formed texts, a coding regime is necessary. For ill-formed – characterised by inconsistent morphology, arbitrary or counter-intuitive punctuation and random white space – the Damerau-Levenshtein Approximate String Matching (ASM) algorithm is used to multi-code responses against target synonym/variant lists. The algorithm has been extended so that the tolerable *Levenshtein distance* can be scaled against the length of each whitespace/punctuation-delimited sub-string. In effect, longer strings will have potentially more deviations from the official lexicon, and so can tolerate a higher distance (without loss of accuracy when matching a synonym/variant list) than shorter strings.

The results will show the extent and rate of narrative and thematic reaction in Twitter comments to the events and issues of the political campaign leading up to the Australian 2013 federal election held on 7 September 2013.

The Problem

The primary bottle-neck in complex tracking operations is the human coding of open-ended questions. This can add days to turn-around time, and inconsistencies in coding judgements can wreak havoc on small weekly samples.

Various recent attempts to take a novel approach to bodies of free text, such as Wordle and related mapping techniques, may create appealing graphics, but essentially contain no more information than a sorted frequency count.

Ideally, we all eagerly await the promise of AI to break the semantic barrier, to see software which can actually and really make sense of text the way a literate human would, but there is no reason to think that such a thing will ever emerge from the confines of science fiction into the real world, despite the many advances in machine translation. Machine translators, like all software, just follow the programmed rules, and are only now getting to a point of usefulness because recent CPUs can work fast enough to statistically canvass the options. But this is not AI, and will never deliver, say, an accurate unaided one-page summary of the major themes in *War and Peace*.

This does not mean that the cause is lost, however. A common sense approach to coding ill-formed text is seen with search engine prompts like 'Did you mean to search for...' and the lists of close-enough words presented by spell-checkers. These all employ various forms of approximate string matching, which is exactly what we need to be able to routinely and consistently auto-code things like ill-formed brand list verbatims. We can go further than this, however, and adapt these techniques to get as close as we can to having a machine extract, if not meaning itself, then at least results which are sufficiently meaningful to a human who understands the context. This paper discusses and demonstrates how to achieve this goal.

Coding Text – Exact and Approximate String Matching

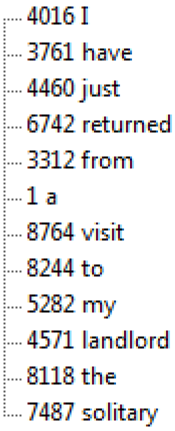
The naive approach to auto-coding text is

- Read all source words (or complete strings) into an array
- Sort alphabetically
- Assign codes from 1 to N, where N is the number of unique words (or unique strings)
- Write the assigned codes in the original word (or string) order

Reading and sorting all individual words shows which ones are dominating, and gives a minimal code frame, but all context is lost. This is the first step in a word map, where frequency directly determines font size. Reading and sorting complete strings instead, such as the full response to an open-ended question like ‘What do you think of the state of the world today?’ will usually result in nearly as many codes as there are respondents, but the context is retained for human review, typically in a table of raw responses by the human-categorized (coded) version.

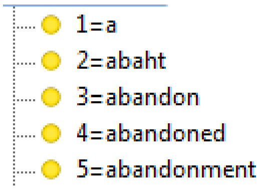
For the analysis examples in this paper, individual words or discrete phrases are read and coded. The presumption is that auto-coding is appropriate for huge volumes where attempting to review the responses individually and manually is not feasible. It is the overall patterns and associations which are sought, and therefore human review, other than to spot-check a few cases, is in any regard not required or useful. One can quantify that love unfulfilled is the dominant theme of both *Wuthering Heights* and *Romeo and Juliet* by simply reading the text, but that is not possible when faced with total word instance counts in the tens of millions or more, short of employing a team of hundreds of human coders (and then issues of consistency across coding judgements inevitably arise).

Static well-formed text, such as a published novel, is code complete. The number and variety of words is bounded at publication time, so each can be assigned a code which, when decoded and read sequentially, exactly reconstructs the original word order. The first line of *Wuthering Heights* is



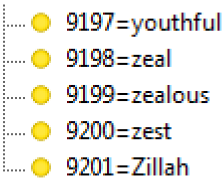
```
4016 I
3761 have
4460 just
6742 returned
3312 from
1 a
8764 visit
8244 to
5282 my
4571 landlord
8118 the
7487 solitary
```

Note code 1 = ‘a’, because ‘a’ is first in the alphabetical sort. The total number of unique words is a known constant, at 9,201.



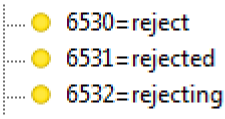
```
● 1=a
● 2=abaht
● 3=abandon
● 4=abandoned
● 5=abandonment
```

through to



```
● 9197=youthful
● 9198=zeal
● 9199=zealous
● 9200=zest
● 9201=Zillah
```

The traditional approach to categorising such data is a Boolean net. For example, to capture the theme of abandonment from the *Text* variable, define a code 1 for a constructed variable *Themes* as the net of codes 3, 4 and 5 of *Text* (noted hereafter as Text(3/5)). Abandonment is a similar concept to rejection, so observing that



```
● 6530=reject
● 6531=rejected
● 6532=rejecting
```

the net can be expanded as Theme(1) = Text(3/5,6530/6532). The netted codes are effectively a synonym list, so decoding we have

abandonment = abandon/abandoned/abandonment/reject/rejected/rejecting

In contrast, dynamic ill-formed text in a tracking context, such as verbatim responses to an open-ended Brand Awareness question, or time-dependent blog or social media exchanges, can never be code-complete, because forthcoming data may throw up unanticipated variations. Variations can be as elementary as an informal deviation from the official lexicon (in English-speaking countries, the Oxford English Dictionary), an official variation across subcultures, such as USA vs British English, or a completely different word morphologically but closely related semantically.

For example, in a study of pet ownership, it is observed from the raw data that the text items *dog* and *dogs* often occur, so the synonym and variants list commences as ‘dog/dogs’. But later cases may have *mongrel*, or *mongrels*, *mutt* or *mutts*, or a related species as *wolf* or *dingo*. Reruns of naive auto-coding will of course capture all new instances, and the synonym list of netting candidates can be manually expanded to accommodate, but what we really want is a single code for all mentions of *dog*, no matter how the concept of dogness is actually expressed by the author(s), and which will cover at least unanticipated minor variations in morphology and grammatical form. Since the number of such instances is indefinite until such time as data collection concludes, exact string matching is untenable in practical terms.

Damerau-Levenshtein

The classic approach to handling the issue of unanticipated variations is the Damerau-Levenshtein algorithm for approximate string matching, acronymed in the literature as ASM [1].

The original Levenshtein algorithm calculated a *distance* for each source word against a target, where distance means how many times a character must be inserted, removed, or replaced to transform the source into the target. Additionally, transposition of two otherwise correct characters can be covered by a simple extension known as Damerau-Levenshtein. The significance of these four edit operations is that they collectively address the most common errors introduced by manual data entry [2].

To be useful, it is further necessary to scale the allowable distance for a target match according to the source string's length. Short strings should match exactly because the risk of ambiguity is high (eg *fox* and *ox* have a distance of 1), intermediate lengths can allow a distance of 1 to 2 so that typical misspellings and typographical errors do not prevent a correct match, and longer strings such as multi-word phrases could support a distance of 3 or more, depending on context. In this paper, the allowable distances at short, intermediate and long thresholds are collectively called the *fuzz parameters*.

Setting the thresholds for the scaled distances is determined empirically – if the current choices result in too many false positives then increase the thresholds, reduce the distances, or both.

Context is an important consideration – in a study of adolescent clumsiness, *spill* is likely to have its common meaning, to be synonymed with *splash*, *slop*, *drip* etc, and categorised as something like *problems with fluids*, but if a study of leadership struggles in a Westminster system, then its metaphorical meaning of declaring all positions vacant would be preferred, to be synonymed with *plot*, *treachery*, *coup*, *betrayal*, etc, and characterised as something like *leadership* or *power struggles*.

Another extension in the implementation here is to handle multi-word phrases. Kevin Rudd, Australia's prime minister from 2007-2010 and from 26Jun13 to 7Sep13, is also known as *The Milky Bar Kid*, and more recently, as *The Messiah*. Tony Abbott, the leader of the opposition, is widely known as *The Mad Monk*. So to accurately quantify presence (or Share of Voice in advertising tracking terms), these must be included among the more obvious synonyms such as *KRudd*, *KevinRuddMP*, *KR*, and *Tony Abbott*, *Abbott*, *TA*, etc.

To clarify what precisely is gained by using Damerau-Levenshtein for ill-formed text, a target word like *megalomaniac* at distance=1 will cover 12 letters * 26 possible in situ typos (negalomaniac, megalonaniac) + 12 missing (megaomaniac) + 12*26 extraneous (megaloomaniac) + 11 transpositions (meglaomaniac) + 2*26 for an extra pre/post character (mmegalomaniac), giving 699 possible variations, any of which, as source input, will be automatically matched and coded against the target *megalomaniac*. Not all of these will be aberrations from the lexicon – distance=1 also matches on *megalomaniacs*, *megalomania*, and *megalomanias*. When defining the synonym lists, count how many moves are required to transform the source to the common lexically correct variations, and set the distance and scaling thresholds accordingly.

It is important to note that the methodology professes to no intelligence. A human seeing the sentence 'the vat sat on the nat' in an ill-formed context would most likely assume typographical errors for cat and mat, and impute the meaning to be 'the cat sat on the mat', especially since c/v and m/n are adjacent on the QWERTY keyboard. The algorithm cannot deal with this as single words because distance must be at least 1 to get from vat to cat, and from nat to mat, so vat will match on both cat and mat (whichever is encountered first) and nat will match on both mat and cat (whichever is encountered first). To find the correct match, the algorithm would require a target of 'the cat sat on the mat' and distance=2, one step for *vat* to *cat*, and one step for *nat* to *mat*. Similarly, *just* and *rust* are 75% equivalent morphologically, at a distance of 1 over 4 letters, yet unrelated in meaning. A human of course knows that the meanings are unrelated, but to avoid a false positive from the algorithm would require enforcing distance=0 for source strings of four or less characters. Damerau-Levenshtein cannot impute meaning. That remains the province of the human.

Code Matching vs String Matching

An exact string match on a set of strings is logically equivalent to an exact code match on a set of codes, so for well-formed texts, one can either net coded synonyms, or net uncoded strings. That is, if *cat* is coded as 1, then there is a code match on {1,2,3}, and there is a string match on {*cat*, *dog*, *rabbit*}. Computationally, matching codes should be faster than matching strings, but to verify the methodology, the Damerau-Levenshtein algorithm is used for both, the difference being that for well-formed, the fuzz parameters are set to zero and the synonym lists are exhaustive. The output is therefore functionally equivalent to a precise Boolean code net. Since most well-formed texts are short by comparison to something like the results of a Twitter search on a trending topic, in practice code matching by Boolean algebra evaluations would be preferred, because that way 100% accuracy can be guaranteed at a lower runtime cost. For Twitter, on the other hand, 100% accuracy is impossible. All analysis, even rudimentary searches through the public Twitter API, will lose information and suffer noise [3]. However, false positives, if the fuzz parameters are set appropriately, are usually insignificantly small.

If matching on multi-word strings such as a key phrase which could occur in a sentence, then Boolean netting will have difficulties – how to assemble unique word codes into an ordered-by-syntax sub-string? On a code frame of 1=bar, 2=kid, 3=milky and 4=the, the expression Text(1/4) will return true on source like 'I gave my kid the last milky bar'. The extended Damerau-Levenshtein, by contrast, will return a distance of zero only on exactly "The Milky Bar Kid". Damerau-Levenshtein is therefore the preferred approach.

The Methodology

The procedure is

- Code the source text, one code per unique word
- Run a sorted frequency count to identify recurrent themes and concepts
- Review actual instances of these words *in situ* to determine appropriate fuzz parameters and the thematic and conceptual contexts, and if they are part of a recurrent phrase
- With the fuzz parameters and contexts in mind, devise a compact target code frame which maps the themes and concepts words of interest to synonym and variant lists
- Process the source text against the targets, to create a categorical variable which can be tabulated in the normal manner against any other variable

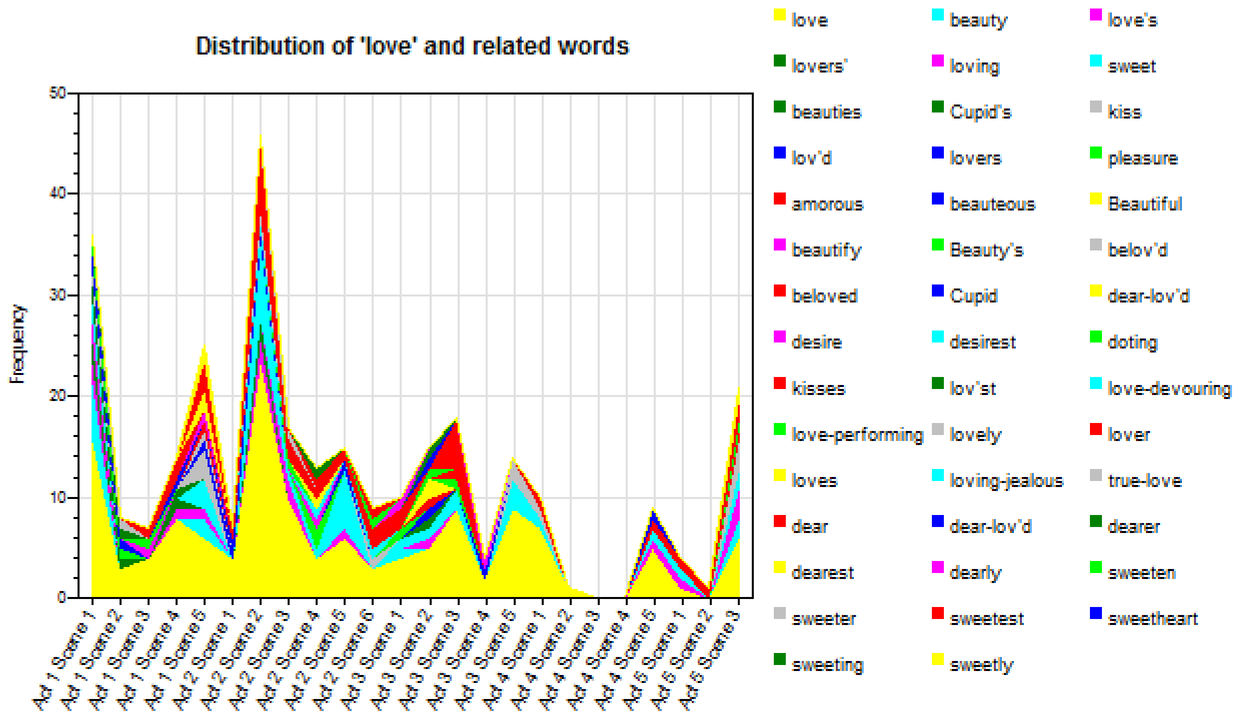
Verifying the Methodology against Well-formed Texts

Romeo and Juliet and *Wuthering Heights* are used as the in-sample validation tests. We know, after some centuries of critical evaluation, that they contain structured themes and narratives, and in the case of the dominant themes and narratives, the critics and general readership are mostly agreed as to what they are. Since our primary intention here is to look at ill-formed text, a few validation examples of well-formed must suffice.

Since it is the number of times an audience or reader hears or encounters a word which builds on the perception of a theme or narrative, rather than the share of them among all words in a chapter or scene, the frequency counts are useful for quantifying exposure. That is, the length of a chapter or scene is not necessarily taken into account. This is viable because the text units (a scene, act or chapter) are usually sufficiently short to be retained in immediate memory. Share within a chapter or scene can however illuminate intensity. Both frequencies and percentages are used below.

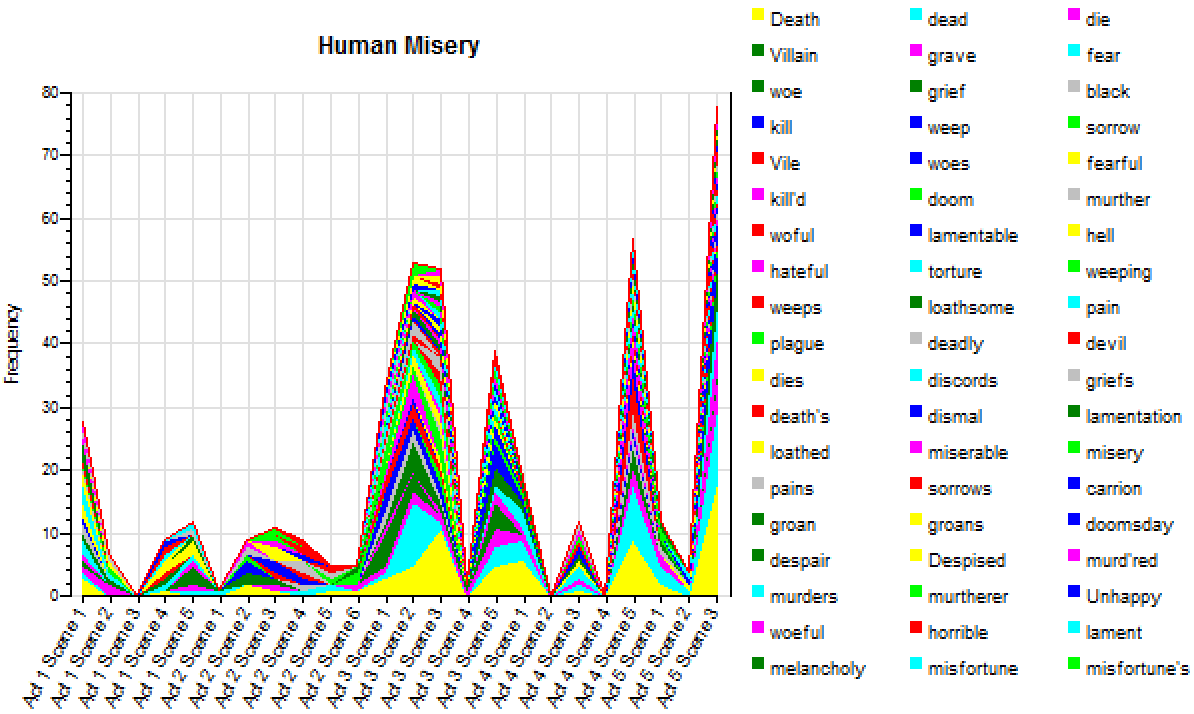
Romeo and Juliet

As a teenagers-in-love narrative, the obvious move is to net all words related to love, and plot their frequency over the scenes.



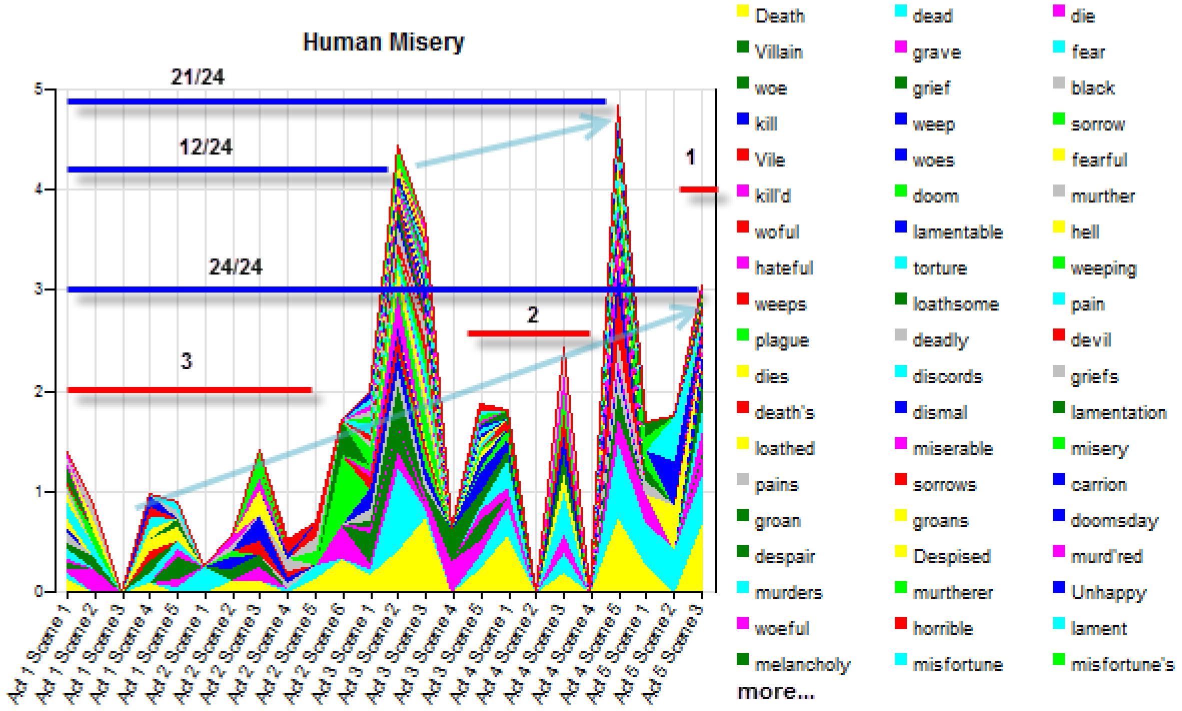
The legend shows the breakdown of the synonym list. The area plot shows the relative contribution of each target. The peak at Act 2 Scene 2 is the orchard/balcony scene: ‘Romeo, Romeo! wherefore art thou Romeo?’

The love theme recedes in diminishing waves from Act 2, to be replaced increasingly by misery and pain.



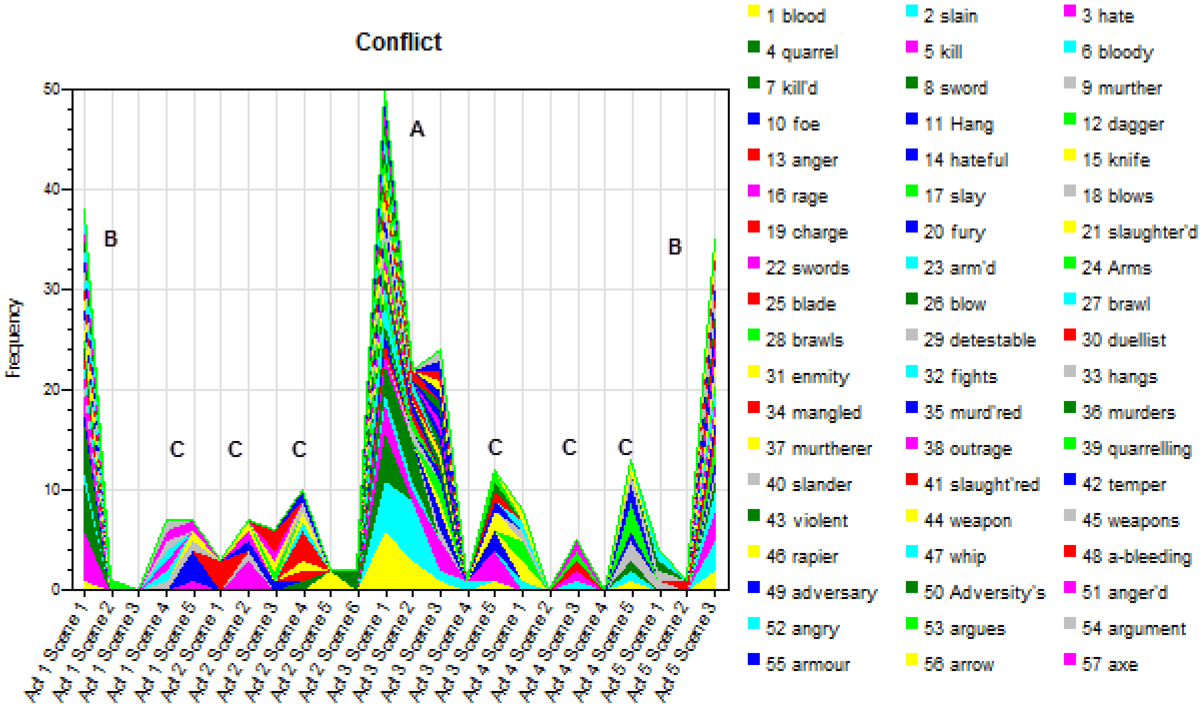
The peak at Act 3 Scene 2 is the aftermath of the deaths of Tybalt and Mercutio. The peak at Act 4 Scene 5 is the report of Juliet's apparent death. The climax peak at Act 5 Scene 3 is where Romeo kills Paris, and then suicides, followed by Juliet.

As a Share of Scene measure (percentage of total), the structure is remarkably regular.



The two dominant peaks are spaced by harmonics at an average period of three scenes. There are eight peaks, giving 3*8 = 24 scenes. The middle peak, at scene 13 (Act 3 Scene 2) bulges somewhat, but the average still stands. Rounding the sequential scene 21, at Act 4 Scene 5, to 20, gives the three highest peaks at 12/24=1/2=halfway, 20/24 = (3+2)/6 = 5/6, and 24/24 = (3+2+1)/6=1=the exodus.   The upward trend on the sub-peaks is nearly parallel with the upward trend on the two dominant peaks, in a pattern of escalating advance and retreat, of flow and ebb.

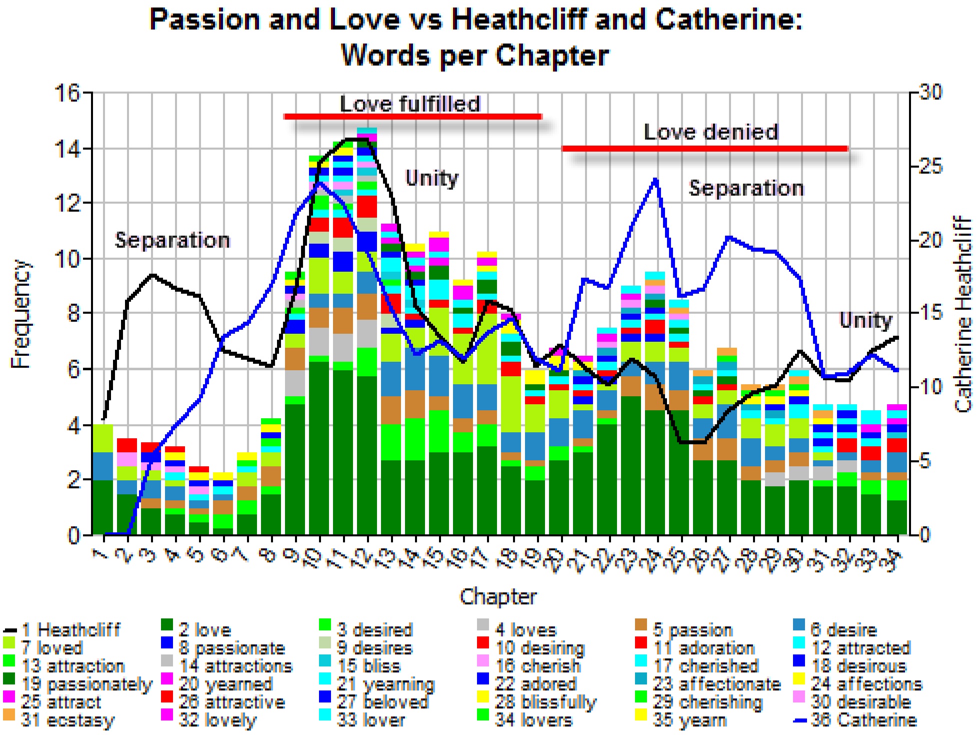
The distribution of the theme of conflict is remarkably symmetrical.



The intermediate peaks B are at scenes 1 and 24. The dominant peak A is at scenes 12 to 13, with three minor peaks C either side.

Wuthering Weights

Not all themes can be inferred by word counts – sometimes a proxy is required. Unity and separation can be inferred by counts on Heathcliff and Catherine/Cathy, and as would be expected, the theme of love and passion tracks where Heathcliff and Catherine are mentioned most often in conjunction, across chapters 9 to 17 or so. Smoothing at moving average = 4 shows



The separation/unity proxy also allows us to distinguish the different aspects of love, as fulfilled/denied.

The validation on well-formed texts is extremely good, with (I would hope) some additional insights to the established critical literature.

Applying the Methodology to Ill-formed Texts:

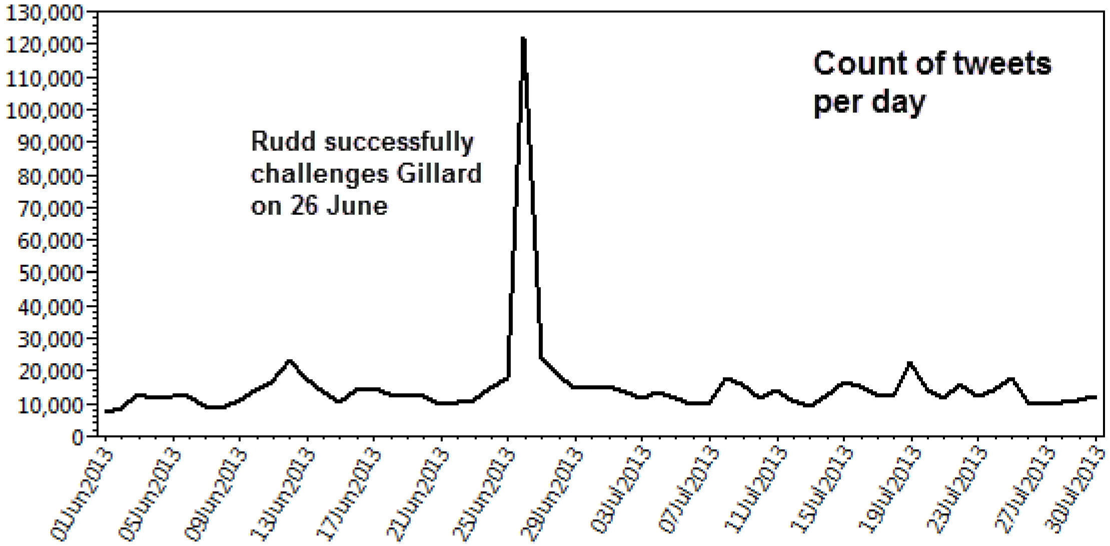
The body of ill-formed text examined here comprises all tweets between 1 June 2013 and 30 July 2013 which are returned on a search for the most common unambiguous hashtags relating to Australian federal politics: #auspol OR #auspoll OR #ausvotes OR #ozcot. These search items were chosen to minimise noise from other jurisdictions - #election2013 and #Spill are also often used, but these give no indication of country, and #Spill is too generic.

Two commercial data source providers were used, Gnip and ScraperWiki [4]. The Gnip data was collected in a single 28 hour run conducted on 15 Aug 2013. ScraperWiki provides user-initiated searches for up to the prior seven days. There is a major difference in the two data sets: because ScraperWiki is near real time, accounts later banned or suspended by 15th August, and hence not in the Gnip data, remain present. The ScraperWiki data is used below only to demonstrate this point.

The search term as executed by Gnip returns 927,190 cases, where each case is a tweet and associated metadata. The major narrative over this period is the defeat of Julia Gillard by Kevin

Rudd on 26Jun13 on the basis of incompetence (himself removed by Gillard in 2010 on that same basis), to return as prime minister to lead the Labor party at the election on 7Sep13 [5].

The activity counts show around 10,000 to 20,000 tweets per day, with a huge spike on 26 June.

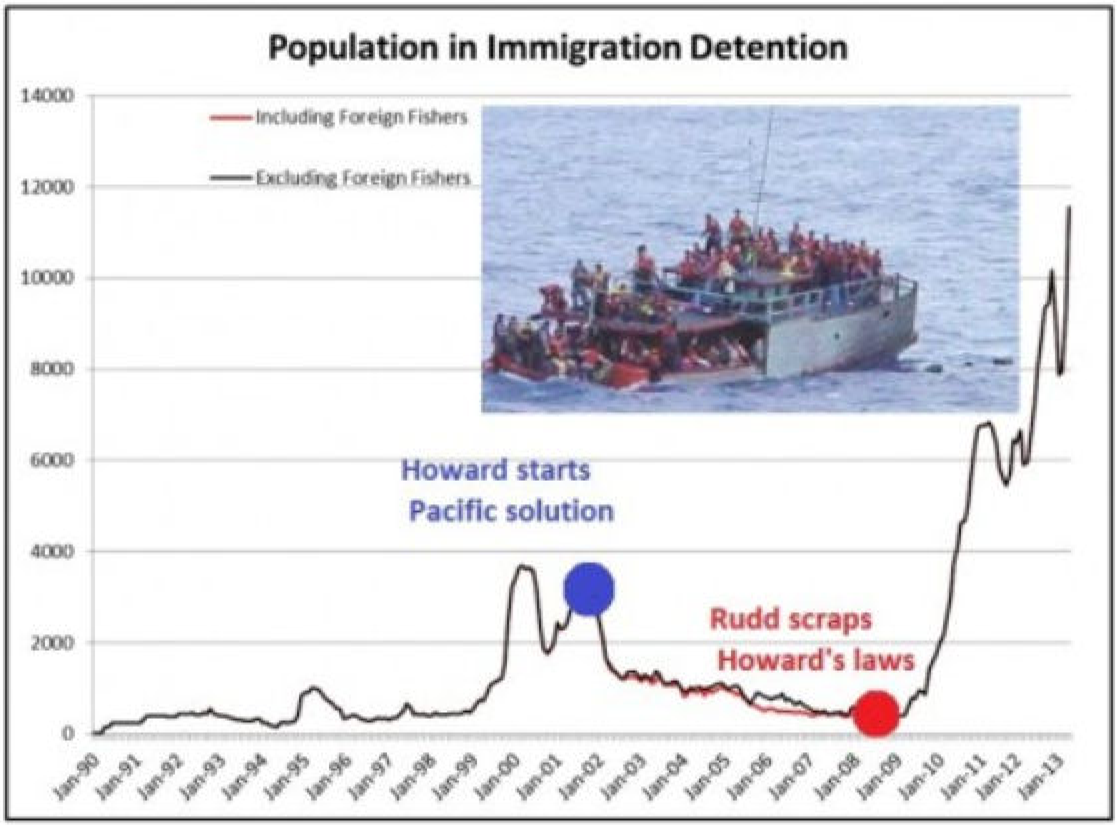


The major issue in the popular mind is border control. This cartoon neatly summarises the situation:

YOU COULD CALL AN ELECTION FOR NOVEMBER SO VOTERS CAN SEE THE SORT OF PRIME MINISTER YOU ARE
OR YOU COULD GO IN AUGUST BEFORE THEY REMEMBER THE SORT OF PRIME MINISTER YOU WERE...
BORDER PROTECTION

Zanetti [6]

Rudd opted for the second alternative, with the election called on 4 August for 7 September. Nearly all illegal entry is by boat, mostly from Indonesia. This chart, hammered by conservative commentators at every opportunity, shows Rudd's essential problem.



To placate the progressive left, Rudd dismantled Howard's border controls in 2008, resulting in an immediate surge in undocumented arrivals.

Thus the scene is set. With the pretender to the throne (Gillard) summarily dispatched, the true and rightful king (Rudd), triumphantly returned from (backbench) exile, now faces the great conservative adversary (Abbott) in a battle to the (political) death for control of the realm – it is *Game of Thrones* where words and image, impressions and appearance, half-truths, spin and lies are the weapons instead of swords and spears. Can this grand narrative be told, and the underlying themes exposed, from the raw mass of Twitter case data?

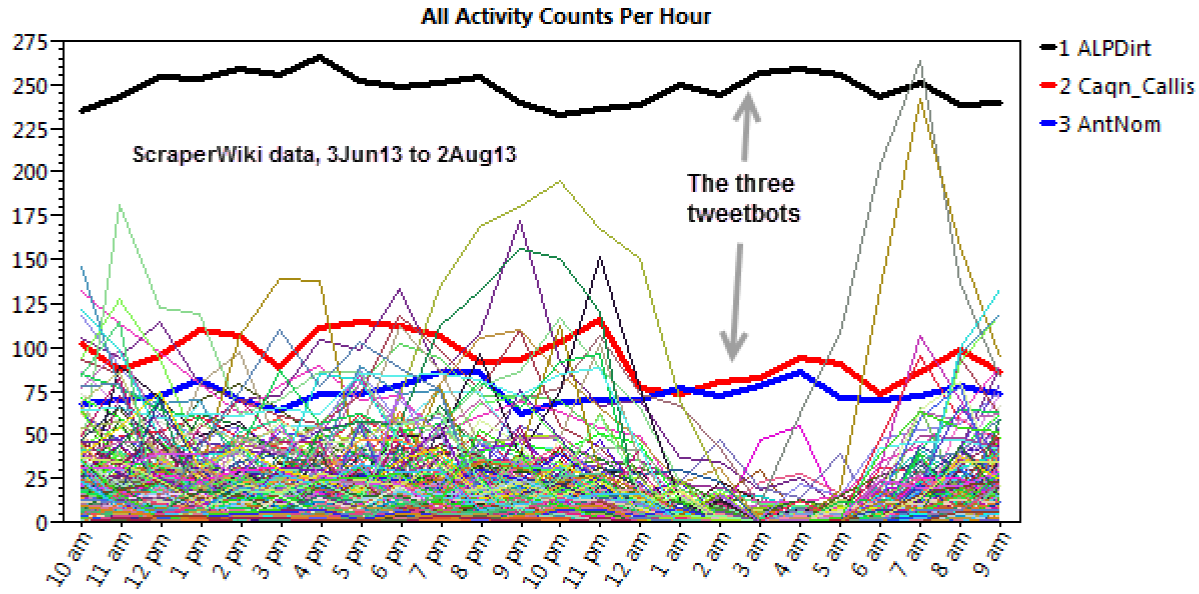
The Twitter narrative form is much more like a play (a collection of 1st person speech instances) than a novel (predominantly 3rd person omniscient). Twitter has no third person commentary – not even stage directions. The closest action to a single point of authorial control over the text is to ban users (and disappear their tweets) who abuse the terms of service. A tweet is a self-authored transcript of a person-to-person communication (whether one-to-one or one-to-many or a retweet) . The Twitter corpus therefore entirely comprises 1st person direct speech, even if framed as commentary. If no explicit addressee, then the stage analog is a soliloquy – an announcement to the world at large of an opinion or insight into a matter for all posterity, but at heart, as with an actor addressing an audience, it remains direct speech. This actually seems implicitly understood by tweeters, as indicated by the extensive use of double quotes.

From the direct speech interactions (the conversations), or declarations (the soliloquies), we can first infer the narratives – the stories being told or recounted or commented – and from the interplay of narratives we can infer the underlying themes. A narrative will be specific – Rudd defeats Gillard – whereas a theme will be generic, and informed by many overlapping or intersecting narratives.

Exposing and Quantifying

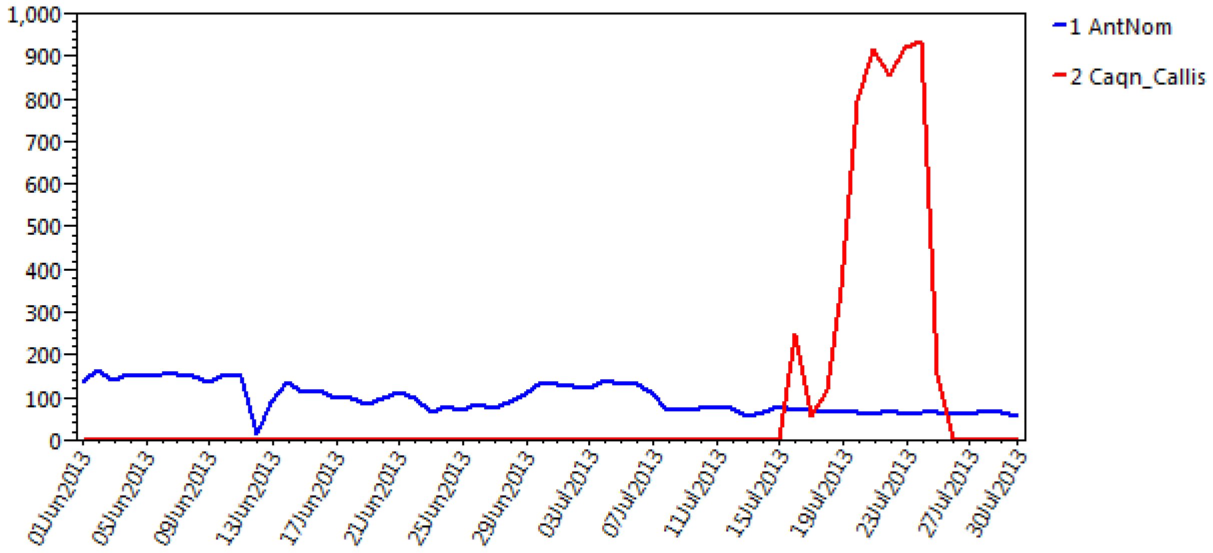
Distorting the Message (1) – Attack of the TweetBots

Since the subject matter is overtly political with tangible outcomes – winners and losers – there are some users identified by screen name whose agenda is clearly to bias the commercial and public search engines. The worst offenders are the tweetbots. The bots can be identified by their inhuman post rate.

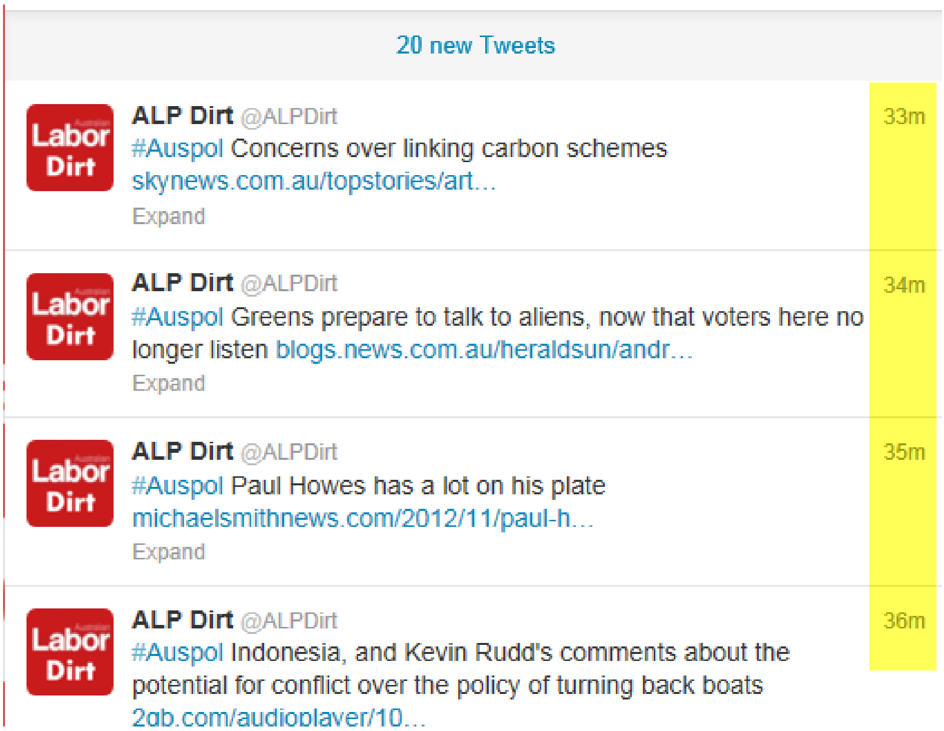


The above chart show that, among all 26,662 screen names for this period, ALPDirt, Caqn_Callis and AntNom never sleep, tirelessly tweeting at a uniform rate across the 24 hour daily cycle. ALPDirt and Caqn_Callis post similar material, and Caqn_Callis is most likely ALPDirt (banned at 13 July) under a new screen name. AntNom is neutral, and behaves more as a news aggregator.

In the Gnip data ALPDirt has disappeared, and Caqn_Callis was suspended at 26 July. AntNom remains.



The tweet rate for ALPDirt was about once a minute.



20 new Tweets

ALP Dirt @ALPDirt 33m
#Auspol Concerns over linking carbon schemes skynews.com.au/topstories/art…
Expand

ALP Dirt @ALPDirt 34m
#Auspol Greens prepare to talk to aliens, now that voters here no longer listen blogs.news.com.au/heraldsun/andr…
Expand

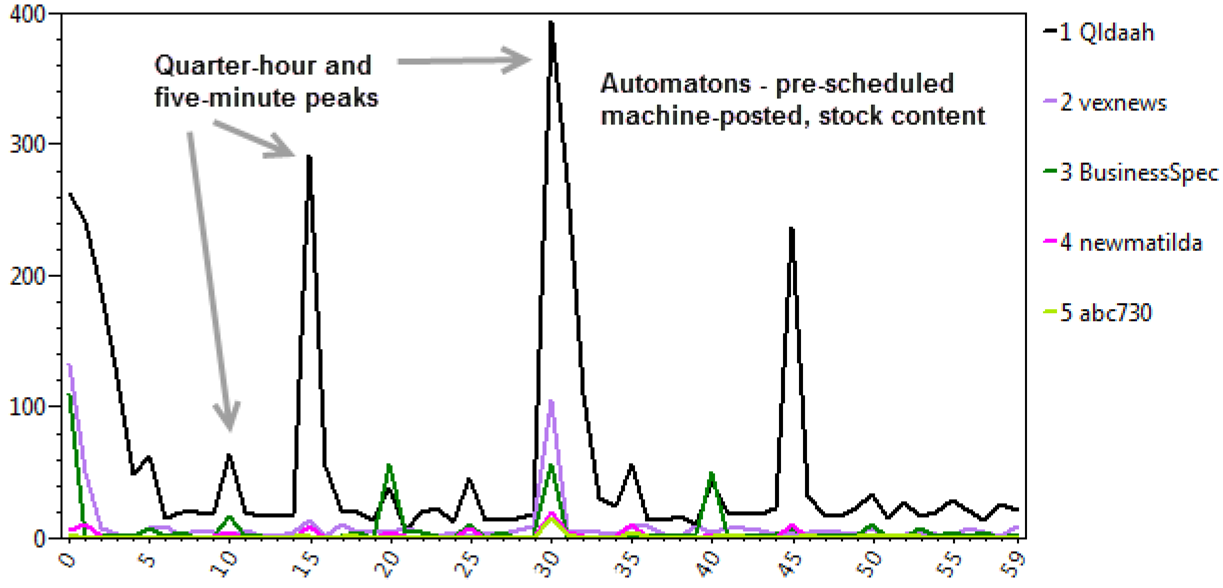
ALP Dirt @ALPDirt 35m
#Auspol Paul Howes has a lot on his plate michaelsmithnews.com/2012/11/paul-h…
Expand

ALP Dirt @ALPDirt 36m
#Auspol Indonesia, and Kevin Rudd's comments about the potential for conflict over the policy of turning back boats 2gb.com/audioplayer/10…

This of course raises a philosophical issue – the impact of the tweetbots is not removed by purging the record after the fact, so we are left with causeless events.

Distorting the Message (2) – Scheduled Automatons

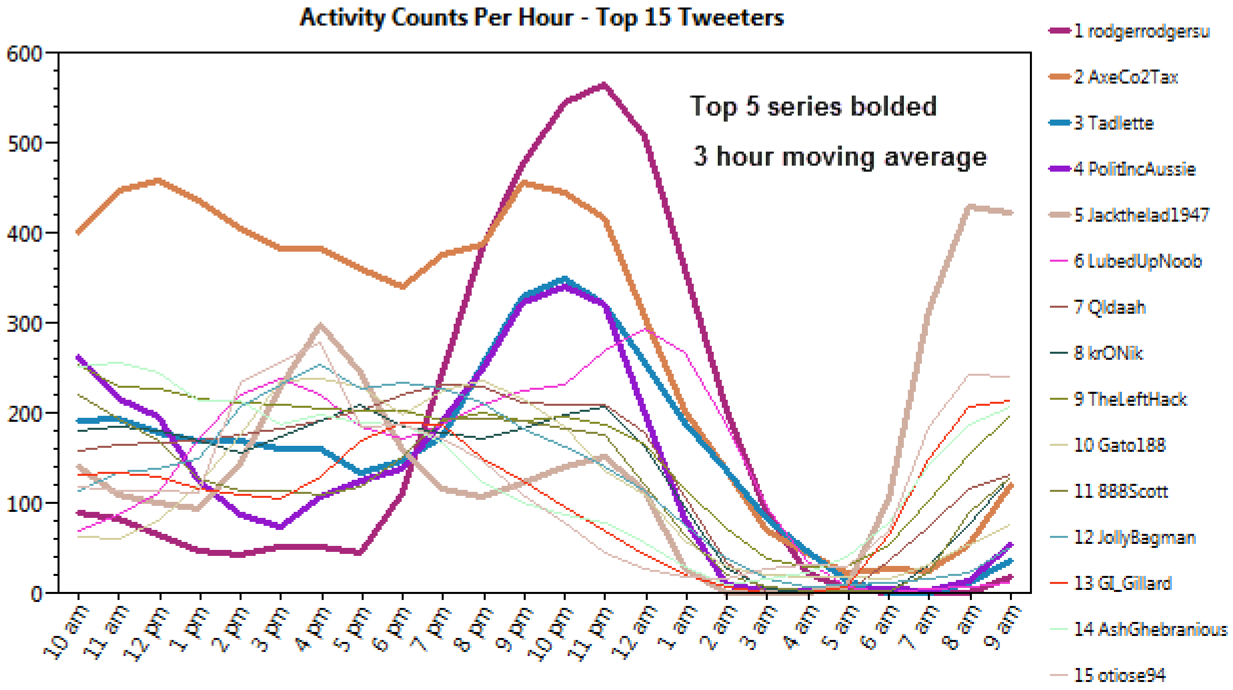
First cousin of the tweetbot is the scheduled automaton. These can be identified by the inhumanly regular posting times, at multiples of 5 and 10, with peaks at the quarter hour.



The automatons are all media related rather than agenda-driven, and also often stimulate conversations in their own right, and so are retained in subsequent analysis. At worst, they may amplify slightly by reposting duplicate material over the news cycle.

Distorting the Message (3) - Obsessive Compulsives

Looking at the top 5 of the top 15 tweeters, we see clear evidence of addictive/obsessive/compulsive behaviour, but no reason to eliminate their contribution, since they do appear to sleep at least a few hours a night.



At our level of granularity, we can detect and eliminate by applying a global filter to remove the tweetbots. Retweets of tweetbot-originated material stands, however, since that represents a considered endorsement by another user. In all subsequent charts, the tweetbots Caqn_Callis and AntNom have been filtered out. Since only Gnip data is used hereafter, ALPDirt is not present, so no filter required.

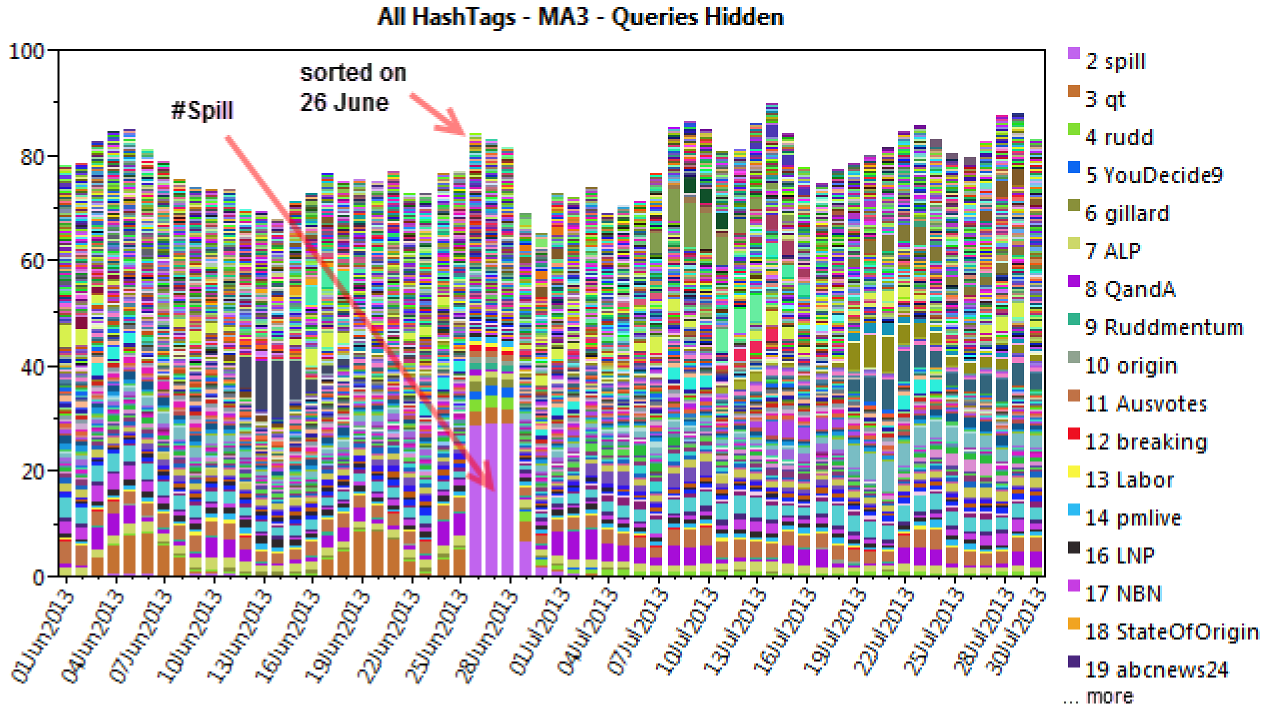
What the HashTags Tell us

One of the problems of conventional machine analysis of free text is the need for metatags to clarify context and intent. In the case of Twitter, we have the advantage that the hashtag convention is followed by many, and so the metatagging to some extent comes ready-made. The hashtag tells us that, from the author's point of view, the tweet itself is intended as part of this or that narrative, which will in turn exemplify or inform a theme or themes.

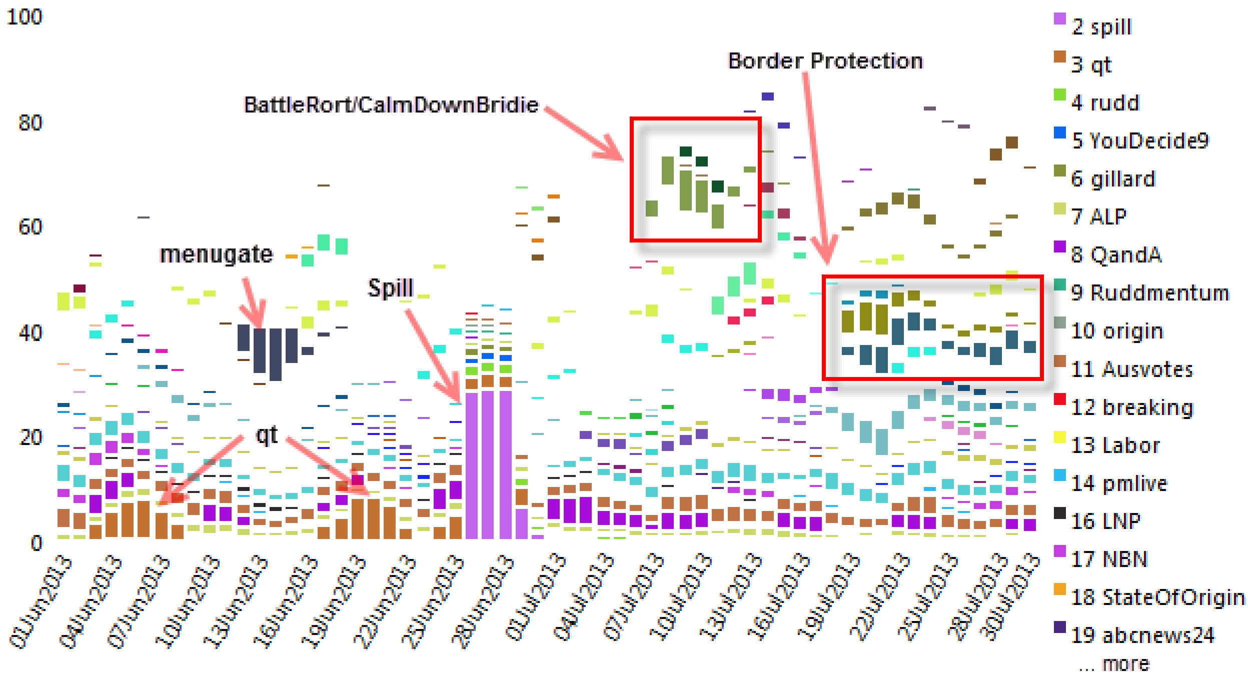
In our case, the search term ensures that every returned tweet has at least one hashtag. Collectively, the search term ensures that every returned tweet is regarded in the mind of the author as relating at least tangentially, and at most directly, to the federal election. But the hashtag is much richer than just a *de facto* search or indexing term. They can be used also as tokens to carry a further message: commentary on current affairs (#1000BoatDeaths, #20000JobCuts), calls to action (#2013electiondateplease, #AbolishParliament), political attack (#AbbotLies), to take a position (#AgeOfEntitlement), and so on.

There is no official process for registering a hashtag in such a way as to ensure that all users employ the same string, so there is a lot of variety, although some formalism is evident since users would generally want their tweets to be found. The initial character must be a # followed by an alphanumeric, underscore as only punctuation mark, no white space, and if multiple tags then space- or punctuation-delimited – these are the same rules which apply to variable names in most computer languages, and are clearly intended to keep the hashtags parsable by search and indexing routines.

There are currently 33,206 unique hashtags returned by the search term over the period. A column percent stacked bar chart of Day by them all except the four search term items on a 3 day moving average is

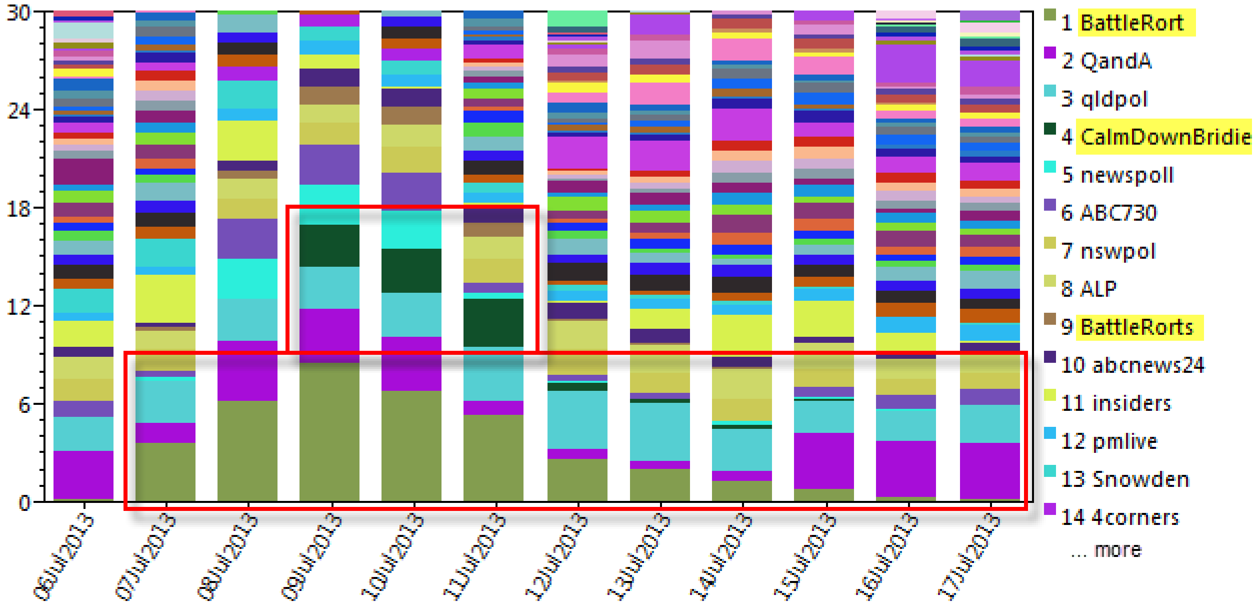


There are clear trends here. Applying a white colour mask to hide all small cells not already visually lost due to screen resolution exposes the structures:

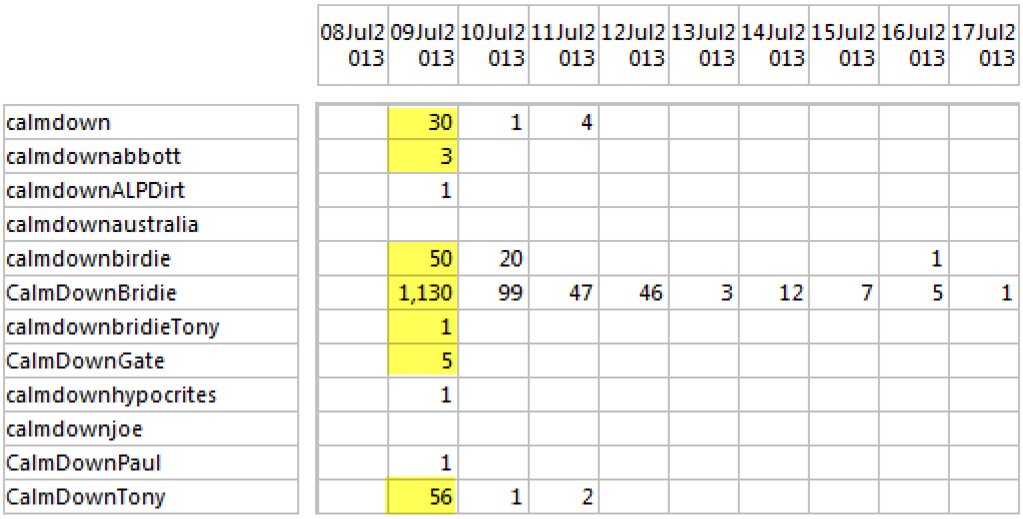


This shows which hashtags are *trending*, in Twitter parlance. To identify the actual tags, sort on a day of interest. There are too many narratives happening here to discuss them all. The main ones are #menugate (a sexist menu used at a Liberal Party function), #qt (parliamentary question time), #spill (leadership spills), #BattleRort (a dodgy expenses claim by Abbott, and put-down of journalist Bridie Jabour #CalmDownBridie) and multiple interrelated tags for border protection. As representative, two only will be followed through here – the BattleRort/CalmDownBridie nexus, commencing July 7 and 9, and the dominant narrative from 19 July, border protection, arising from Rudd's announcement of the PNG-Solution [see note 5]. Theme-wise, BattleRort informs sexism and corruption, and border protection informs racism and xenophobia.

Sorting on 9 July gives (in zoom view)



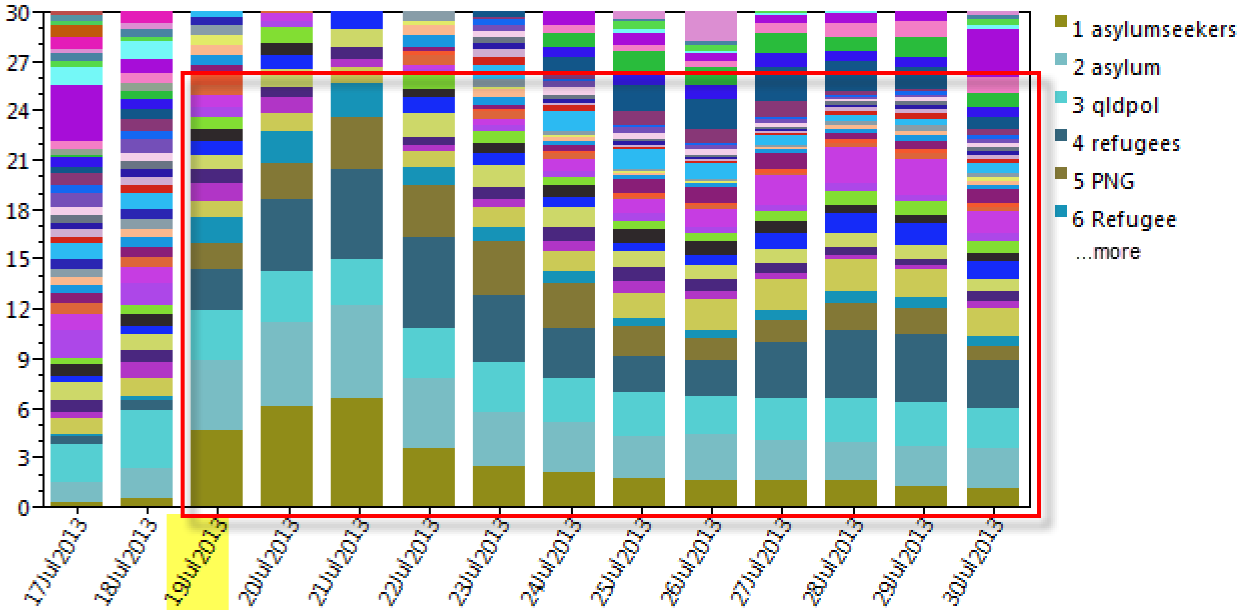
Ignoring QandA and qldpol, the two spontaneous narratives which emerge from the noise are #BattleRort and #CalmDownBridie. On 7 July, Guardian journalist Bridie Jabour persistently questioned Abbott about an improper expenses claim when promoting his book *Battlelines*. Abbott eventually suggested she 'calm down', which was promptly deconstructed by the progressive left as sexist because it implied hysteria. #CalmDownBridie, coined by Bridie Jabour herself [7], immediately spawned many variants:



	08Jul2013	09Jul2013	10Jul2013	11Jul2013	12Jul2013	13Jul2013	14Jul2013	15Jul2013	16Jul2013	17Jul2013
calmdown		30	1	4						
calmdownabbott		3								
calmdownALPDirt		1								
calmdownaustralia										
calmdownbirdie		50	20						1	
CalmDownBridie		1,130	99	47	46	3	12	7	5	1
calmdownbridieTony		1								
CalmDownGate		5								
calmdownhypocrites		1								
calmdownjoe										
CalmDownPaul		1								
CalmDownTony		56	1	2						

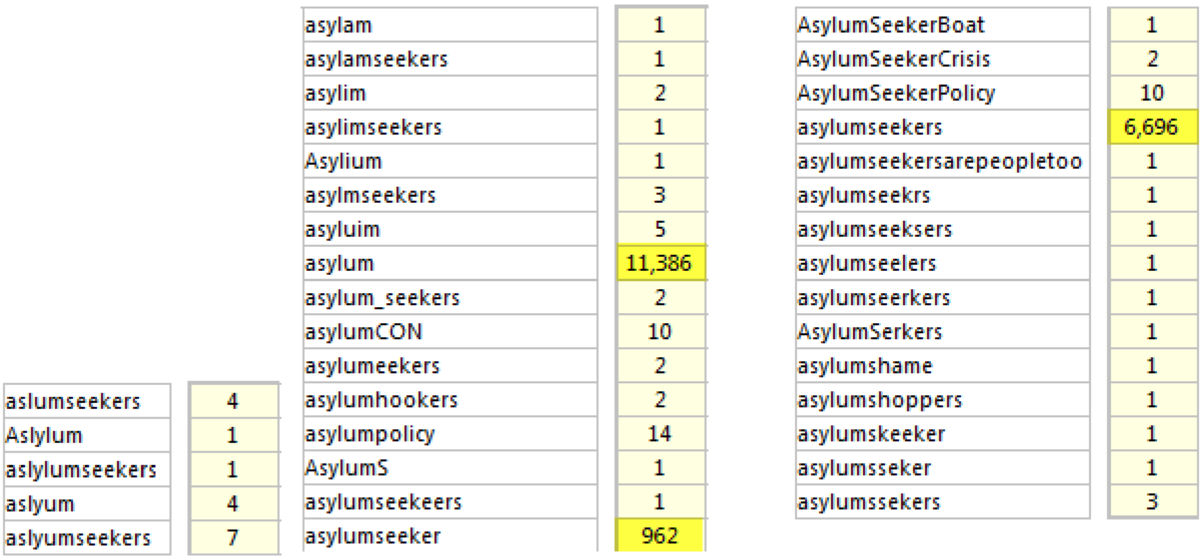
The dominant hashtag is #CalmDownBridie, at 1,130 mentions on 9 July, but #calmdownb**ir**die, either as a typo or as a pun (*bird* being slang for *female*) has 50 mentions. Making a human judgement as to which are intended by the author to be a part of this narrative gives 30+3+50+1+5+56 = 145/1130 = 12.8% of mentions which would be missed by a search on just #CalmDownBridie – lost in the noise forever. Thus, assuming accurate and coherent hashtags will result in the loss of quite a bit of otherwise unambiguous information.

Similarly, sorting on 19 July shows this:



Of the first six dominant hashtags, five deal with border protection, PNG (Papua New Guinea) being used for refugee detention sites.

Looking at just #asylumseekers case totals in the full table context (sorted by label), there are many variants:



aslumseekers	4
Aslylum	1
aslylumseekers	1
aslyum	4
aslyumseekers	7

asylam	1
asylamseekers	1
asylim	2
asylimseekers	1
Asylium	1
asylmseekers	3
asyluim	5
asylum	11,386
asylum_seekers	2
asylumCON	10
asylumeekers	2
asylumhookers	2
asylumpolicy	14
AsylumS	1
asylumseekeers	1
asylumseeker	962

AsylumSeekerBoat	1
AsylumSeekerCrisis	2
AsylumSeekerPolicy	10
asylumseekers	6,696
asylumseekersarepeopletoo	1
asylumseekrs	1
asylumseeksers	1
asylumseelers	1
asylumseerkers	1
AsylumSerkers	1
asylumshame	1
asylumshoppers	1
asylumskeeker	1
asylumsseker	1
asylumssekers	3

The three dominant tags are clear, but the variants will be lost under a search on just *asylum* OR *asylumseeker/s*.

Quantification

The above shows that a smoothed percentage chart of all instances will expose the narratives, but to quantify them accurately, we cannot forego counting the variants. To get a more precise read, we apply Levenshtein. Recalling the four transformation rules (insert, delete, replace, transpose), the following matches (among many others) will be made to the dominant forms at run time:

batt**el**rort->batt**ler**ort (transpose once)
calmdownb**ir**die->calmdownb**ri**die (transpose once)
asylumseeke ->asylumseeke**rs** (insert twice)
asylumseeke**e**rs->asylumseekers (delete once)
asyl**y**mseekers->asyl**u**mseekers (replace once)

To prepare the synonym/variants lists for the dominant tags, the procedure is

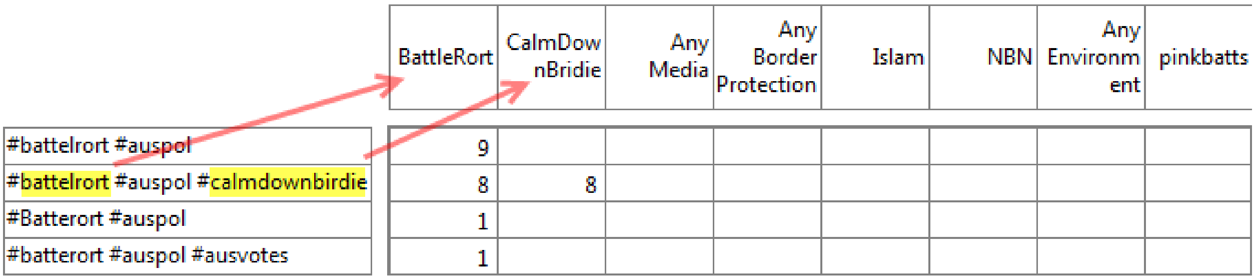
- Code the hashtags, one code per unique tag
- Generate a sorted frequency count table
- Choose a cut-off point - I have used 30
- Review all items > 30, define and initialise a coded synonym/variants list with the dominant tags
- Sort the table alphabetically by label
- Review label blocks for any variants which are too coarse for Damerau-Levenshtein, and add to the relevant synonym/ variants list

This results in the following code definitions:

Code	**Category**	**Synonym/ Variant Targets**
1	BattleRort	#BattleRort/#BattleRortAbbott/#BattleRortGate/#BattleRortMovies/#BattleRortSongs
2	CalmDownBridie	#CalmDownBridie/#CalmDownBr/#CalmDownTony/#CalmDownAbbott/#CalmDown
3	Any Media	#qanda/#abcnews24/#abcnews/#abc24/#Insiders/#730report/#abc730/#730/#lateline/#thedrum/#pmlive/#pmagenda/#amagenda/#4corners/#contrarians/#abc/#MSMfail/#MSM/#contrarians/#theboltreport/#viewpoint/#media/#Murdoch/#ABC1/#mediawatch/#datelineSBS
4	Any Border Protection	#asylumseekers/#refugees/#boatpeople/#asylum/#PNGSolution/#PNG/#Nauru/#ManusIsland/#Manus/#humanrights/#stoptheboats/#Indonesia/#immigration/#boats/#operationsovereignborders
5	Islam	#islamophobe/#islamist/#islamlaw/#islamic/#Islam/#muslim
6	NBN	#NBNCo/#NBN/#fraudband
7	Any Environment	#climate/#coal/#fracking/#carbon/#energy/#CSG/#climatetax/#climatechange/#environment/#ETS/#carbonscam/#carbontaxscam/#climatescam/#climatecon/#green/#AGWHoax/#globalwarming/#greenarmy/#renewables/#votegreen/#climategate/#naturalcsg/#naturalgas
8	pinkbatts	#pinkbatts

- Run these targets against the source hashtags, with fuzz parameters as distance=0 for strings of 4 characters or less, distance=1 for 9 characters or less, and distance=2 for 10 characters or more, to create a new variable comprising eight codes

To confirm, run a table of the eight coded categories against the original raw hashtag text:



	BattleRort	CalmDownBridie	Any Media	Any Border Protection	Islam	NBN	Any Environment	pinkbatts
#battelrort #auspol	9							
#battelrort #auspol #calmdownbirdie	8	8						
#Batterort #auspol	1							
#batterort #auspol #ausvotes	1							

The source strings batt**el**rort and calmdownb**ir**die are both correctly captured.

What the Text tells us

Dealing with the hashtags is the easy part, because they are simple multi-coded lists. The tweet text is considerably more complex, comprising sentences, punctuation, phrases and all manner of randomness, but the same procedure can be applied. The full set of targets and the category codes they map to is

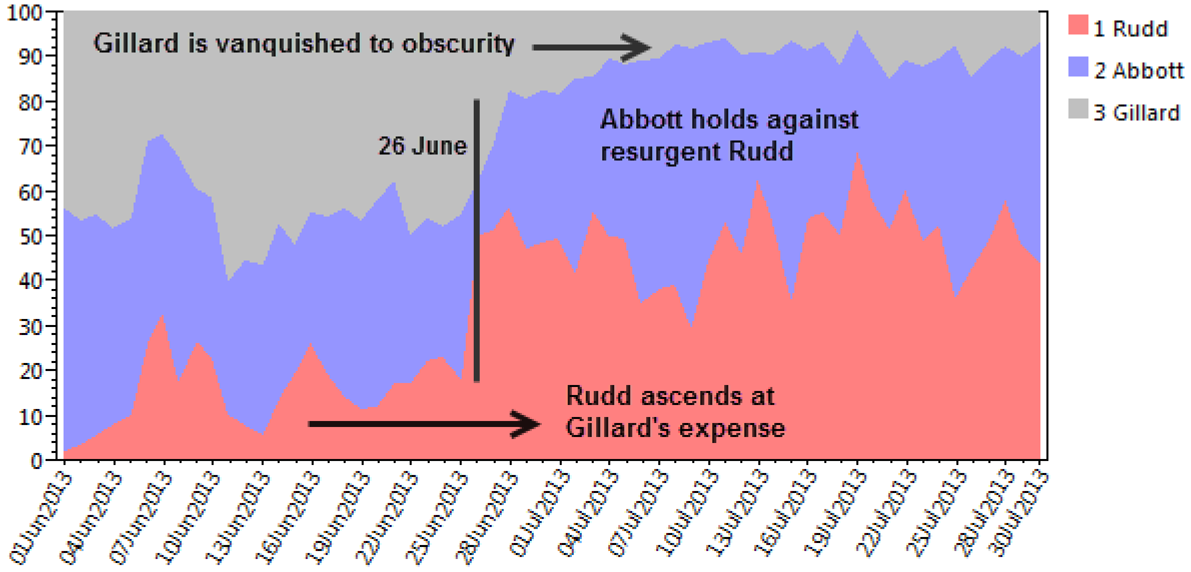
Code	**Category**	**Synonym/variant Targets**
the government		
1	Rudd	Kevin Rudd/KevinRudd/Kevin13/Kevin747/Kevin/@KRuddMP/KRudd/Rudd/CrudDudd/KR/milky bar/messiah
2	Albanese	Albanese/Albosleaze/Albo
3	Gillard	Julia Gillard/JuliaGillard/Gillard/Juliar/Julia/Jules
4	Shorten	Bill Shorten/Shorten
5	Labor Party	Labor Party/Labor/#ALP/ALP/the government/the govt
6	Greens	Greens/Milne/Bob Brown
7	Unions	unions/faceless men/AWU/HSU/Bill Ludwig/Ludwig/Paul Howes/Piggy Howes
the opposition		
8	Abbott	@TonyAbbotMHR/Tony Abbott/TonyAbbott/TAbbott/Abbott/Tony/TA/budgie smuggler/mad monk
9	Turnbull	Malcolm Turnbull/@TurnbullMalcolm/Turnbull
10	Coalition	Coalition/liberal/the opposition/Libs/#LNP/LNP
ideology		
11	Left Wing	left wing/far left/leftist/lefties/progressive/socialist/fabian/communist/class warfare
12	Right Wing	right wing/far right/conservative/tory/capitalist
'isms		
13	Racism	racism/muslim/islamophobia/islamisation/islam/jihadist/jihadi/shia/sunni
14	Sexism	misogyny/misogynist/sexist/cause of women/anti women/against women/gender
issues		
15	Border Protectio	border protection/border security/border/illegal immigrants/illegals/immigration/asylum seekers/asylum/economic refugees/refugees/refugee convention/people

	n	smugglers/smugglers/boats/deaths at sea/Indonesia/Christmas Island/Christmas/Nauru/Manus Island/Manus/PNG solution/PNG
16	Global Warming	global warming/climate change/climatechange/AGW/sea level/hurricane/drought
17	Pollution	polluters/pollute/carbon pollution/pollution
18	Environment	environmentalist/environment/conservationist/sustainability/sustainable
19	Censorship	censorship/censor/internet filter/ISP filter/freedom of speech/free speech
20	Internet Infrastructure	internet/National Broadband Network/Broadband/NBN/FTTN/FTTH/FraudBand
21	Education	education/schools/Gonski/BER/curriculum
22	Home Insulation	home insulation/homeinsulation/insulation/insulators/pink batts/pinkbatts/#pinkbat/batts/electrocution
23	National Disability Insurance	National Disability Insurance/disability insurance/disability/disabled/NDIS
energy		
24	Renewable Energy	green energy/renewable energy/clean energy/alternative energy/renewables/windmills/wind power/wind farm/wind turbines/solar/energy targets
25	Fossil Fuels	fossil fuel/coal seams/coal/natural gas/hydofracking/fracking/shale/CSG
26	Nuclear Power	nuclear power/nuclear
27	Commodities	commodities/commodity/mining
economy		
28	Economy	economy/deficit/austerity/razor gang/finance/debt/inflation/cost of living/prices/budget/surplus/fiscal stimulus/fiscal outlook/fiscal/dollar/GFC
taxation		
29	Carbon Tax/ETS	carbon tax/carbontax/emissions reduction/emissions trading/#ETS/ETS/feed in tariffs/tariffs

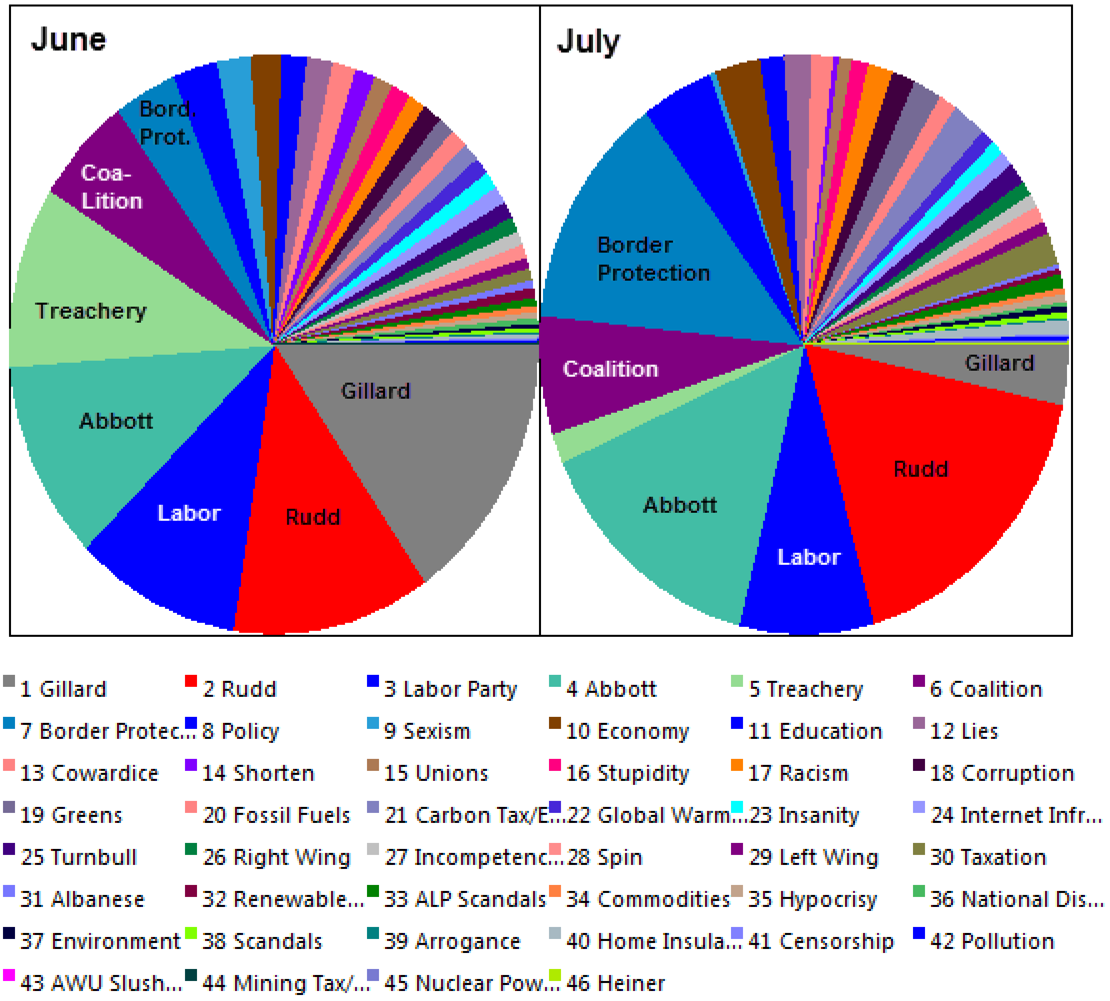
30	Mining Tax/MRRT	mining tax/MRRT
31	Taxation	taxation/taxpayers/tax
negative sentiment		
32	Lies	lying/lies/lie/deceive/deceit/fakers/fibs/snake oil
33	Spin	spin/slogans/politicise/pretence
34	Corruption	corruption/corrupt/fraud/sleaze/dishonest/stealing/steal/greed/shonky/crooks/criminals/thuggish/thugs/thieves/thief
35	Treachery	treachery/treacherous/back stab/backstab/back stabbing/backstabbing/stabbed/stabbing/knifing/knifed/plot/betrayal/betrayed/betray/spill/leadership coup/ousting/oust
36	Insanity	insanity/insane/nutter/crazy/lunatic/unstable/psychopath/psychotic/psycho/narcissist/delusional/delusion/egotist/egotistic/egotistical/egomaniac/ego/power mad/powermad/madness/deviant
37	Stupidity	stupidity/stupid/wanker/numpty/imbecilic/imbecile/zombie/clueless/moron/retarded/retard/idiotic/idiot/bogan
38	Incompetence	incompetent/mismanagement/dysfunctional/waste/inefficient/chaotic/chaos/destructive/inept
39	Cowardice	cowardice/coward/gutless/ticker/frightened/scared
40	Hypocrisy	hypocrisy/hypocrit/bigoted/bigot
41	Arrogance	arrogance/arrogant/smart arse/smartarse/smart ass/self indulgent/smugness/smug
scandals		
42	Scandals	scandalous/scandal
43	AWU Slush Fund	AWU slush fund/slush fund/Bruce Wilson
44	ALP Scandals	Peter Slipper/Slipper/Craig Thompson/Thompson/Eddie Obeid/Obeid
45	Heiner	Heiner
policy		
46	Policy	agenda/policies/policy

Note that many of the categories are now themes, such as corruption, racism, etc. The individual narratives which inform the themes have been generalised upwards.

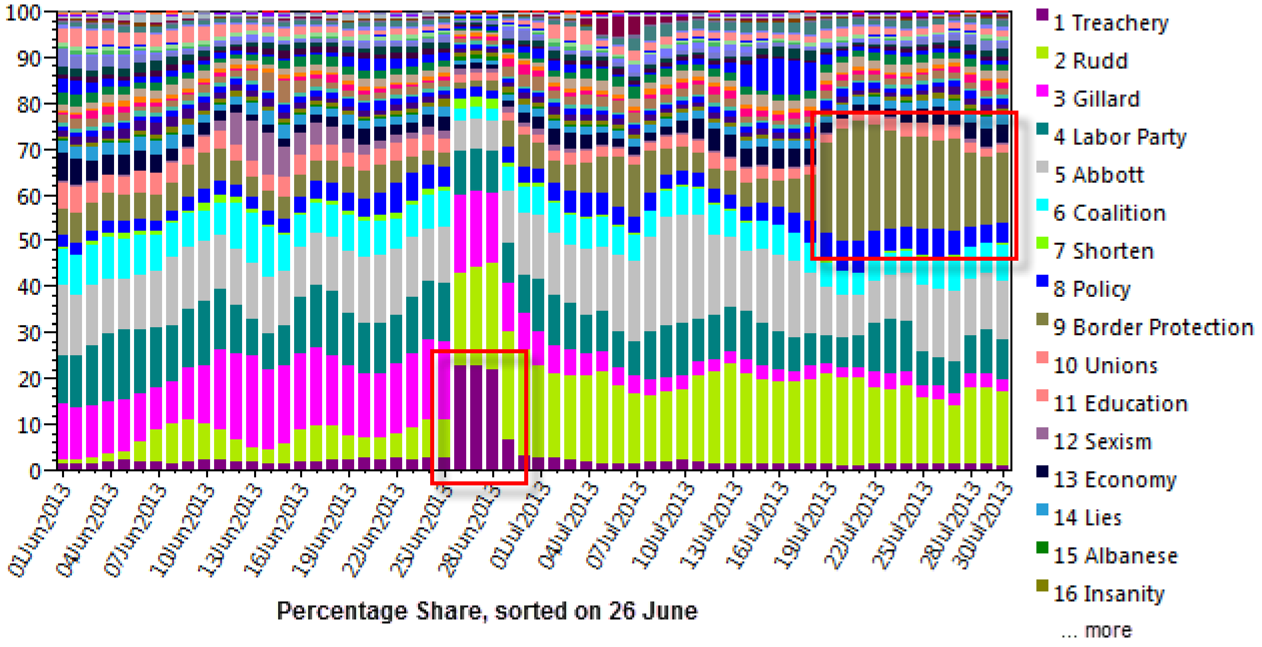
An area plot of the synonym/variant nets to codes 1=Rudd, 3=Gillard and 8=Abbott as percentages of their combined daily totals, gives the traditional advertising measure Share of Voice.



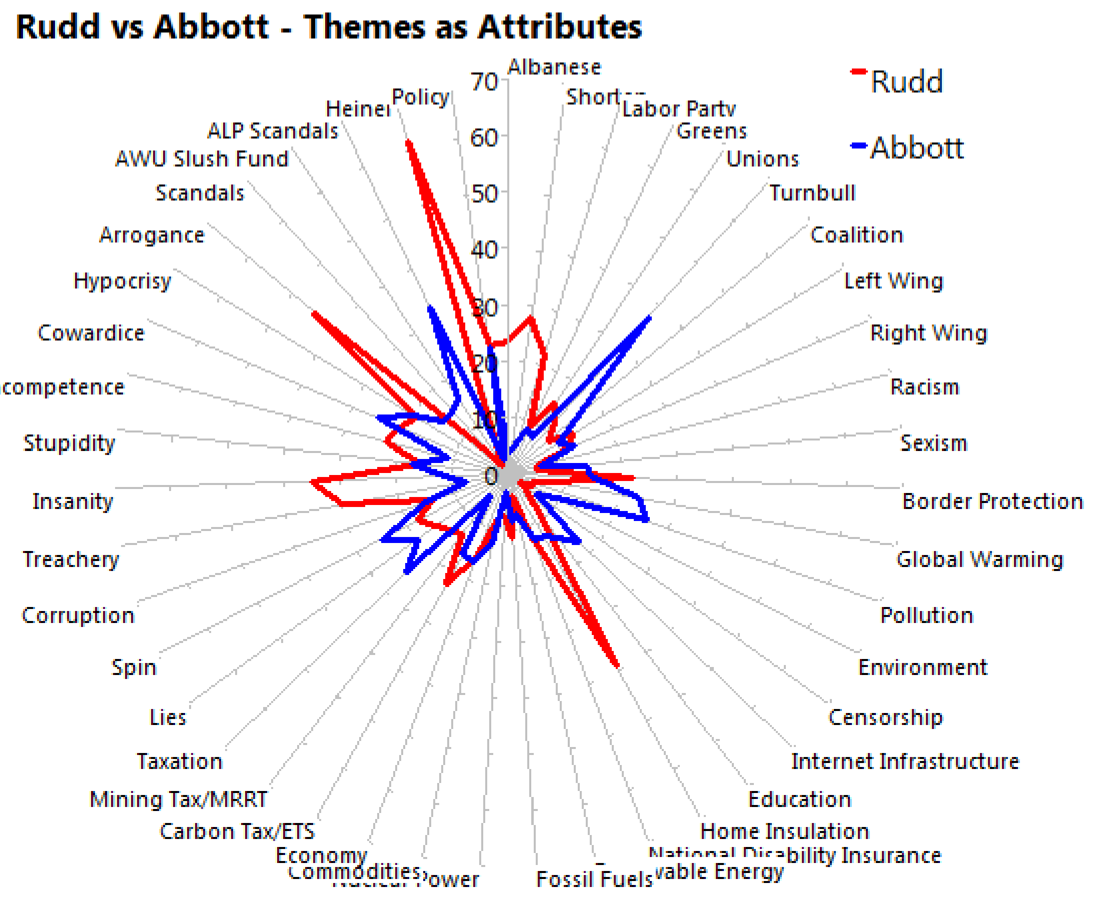
A comparison by pie charts on June and July shows Rudd's displacement of Gillard, Abbott and the Coalition's gradual incursion on Gillard/Rudd/Labor, treachery as the dominant theme in June (the leadership spill), and border protection as the dominant theme in July (when Rudd announced the PNG Solution).



A time series at 3 day moving average shows the big increase for Treachery at 26 June, and in Border Protection at 19 July, the PNG-Solution announcement.



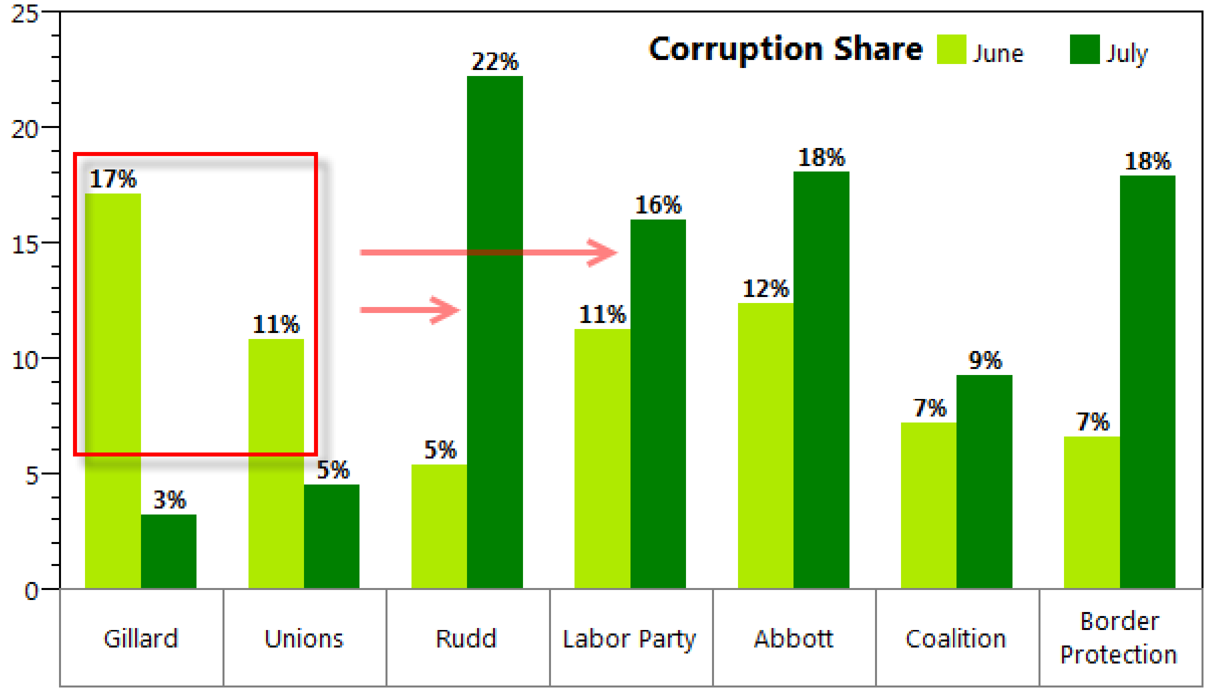
Considering the 46 categories as a type of image statement, a radar chart comparing Rudd vs Abbott shows quite distinct profiles:



Rudd is exclusively associated with Heiner (a child care sex-abuse scandal – Rudd is implicated in a cover-up). Home Insulation (#pinkbatts) is a scheme Rudd pushed in 2009 as a GFC stimulus measure. Too many installations were done badly, resulting in some deaths and an adverse Coroner's Report. Rudd is also high on Treachery, Insanity and Arrogance – three qualities often attributed to him by many of his former ministers and colleagues.

Abbott is less-mentioned, as would be expected for the Leader of the Opposition, so the percentages are a bit lower. Abbott spikes on Turnbull (former leader of the Opposition, defeated by Abbott, and the people's choice), Pollution (Abbott once famously said "climate change is crap", so this spike is not an endorsement), Education (conservative concerns regarding standards), Spin is slightly more attributed to Abbott than Rudd, the spike on Cowardice is from Abbott's refusal to participate in a short-notice public debate with Rudd. The spike on ALP Scandals is because it is his job to hammer the government on scandals.

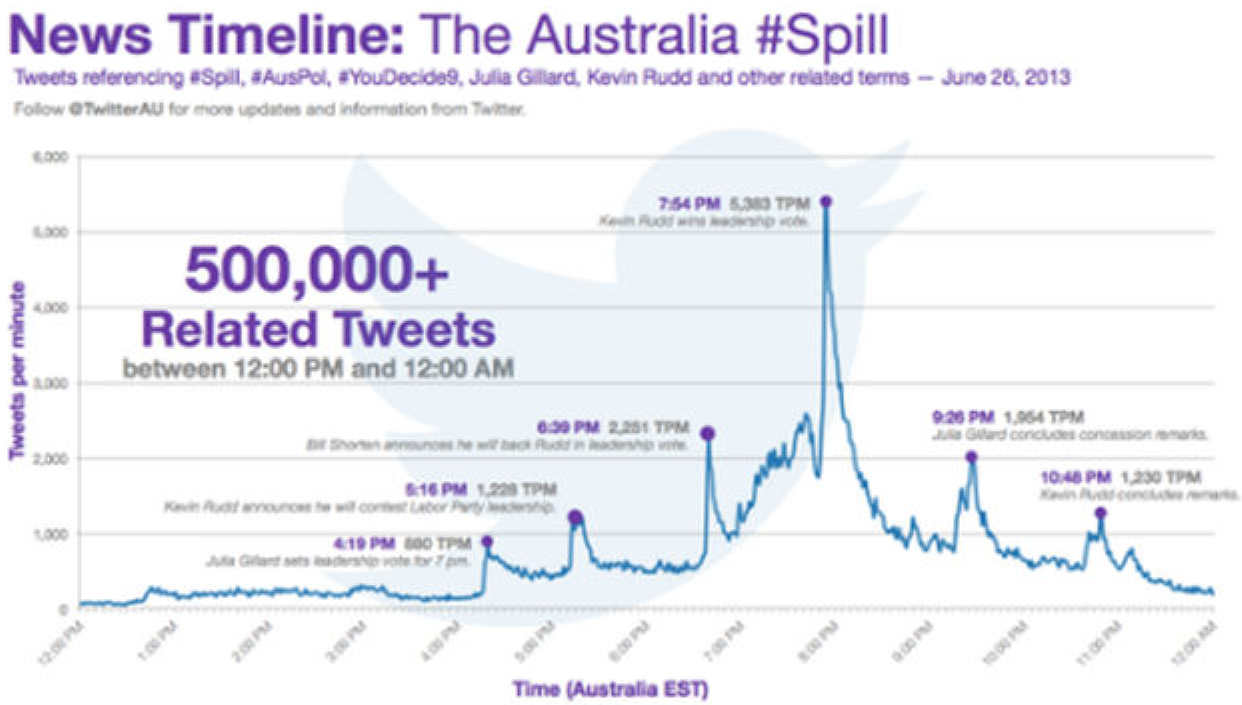
Filtering to Corruption shows a large shift in July. Most of the burden of Gillard's links to long-running union scandals has shifted to Rudd and the Labor Party. Abbott and the Coalition are also being more associated with corruption, but much of that is due to prosecuting the case against the government.



Border Protection is associated with Corruption via mentions of Indonesia and Papua New Guinea, both notoriously corrupt.

Comparing to Commercial and Public Search Tools

According to Topsy [8] there have been over six million tweets with the search term related to Australian elections over the last two years – too many for processing on a stand-alone PC using the techniques described here. There were over half a million on just 26 June, of which my search term captured 190,000.



Handling data of such magnitude over several years must remain at present with the commercial providers such as Topsy and Gnip, with their huge parallel processing and continuous indexing, but fine-detail analysis as conducted above using such tools is not feasible. The commercial providers rely on the Twitter search APIs, but the APIs do not support a search expression with more than 12 terms . The synonyms for Rudd already run to twelve terms:

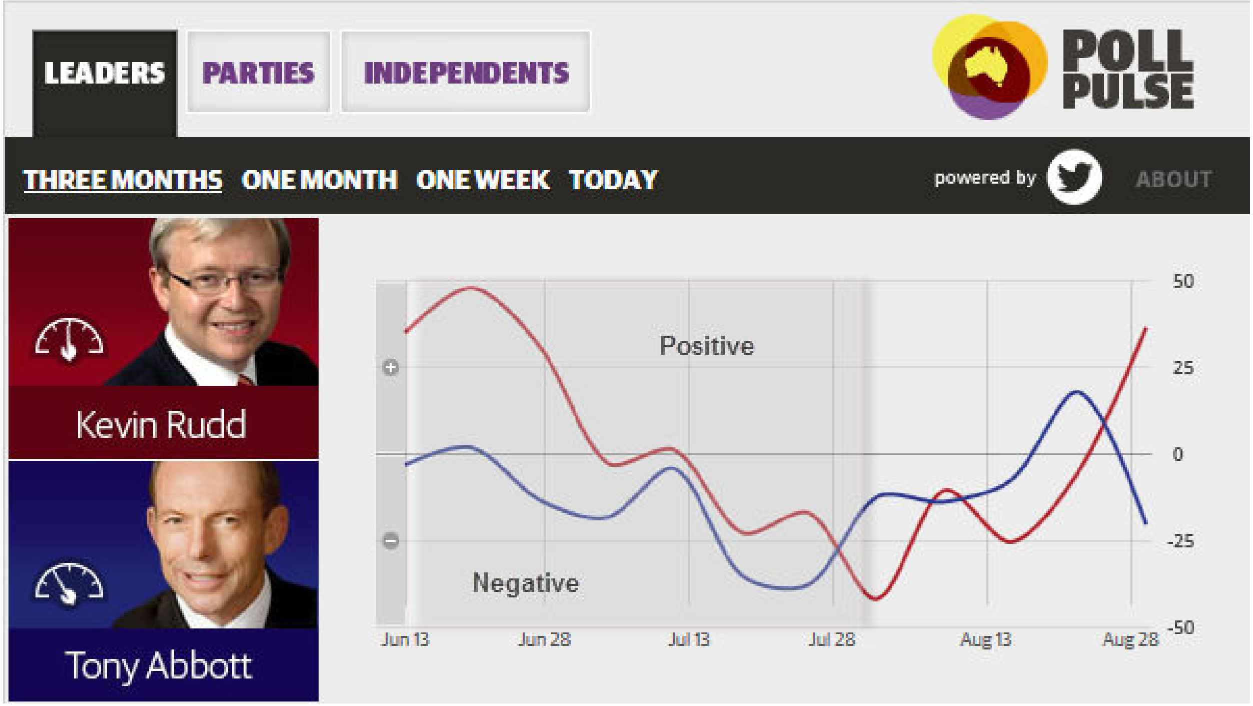
> Kevin Rudd/KevinRudd/Kevin13/Kevin747/Kevin/@KRuddMP/KRudd/Rudd/CrudDudd/KR/milky bar/messiah

Therefore ANDing this with a concept such as border protection

> border protection/border security/border/illegal immigrants/illegals/immigration/asylum seekers/asylum/economic refugees/refugees/refugee convention/people smugglers/smugglers/boats/deaths at sea/Indonesia/Christmas Island/Christmas/Nauru/Manus Island/Manus/PNG solution/PNG

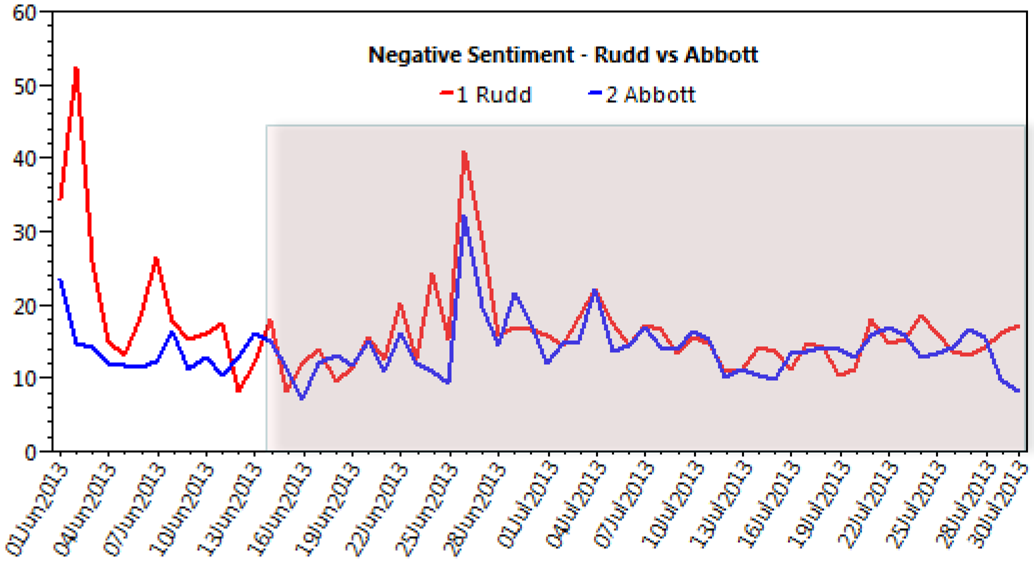
is not possible – and if it were, the results would be from exact string matches, not fuzzy ones. Furthermore, there is no transparency to the search process, and no way to ensure that percentages are on the assumed base. Was the full sample returned? How are derived variables calculated? Has smoothing been applied, and if so, by what algorithm? It is hard to impossible to validate results from a commercial search because both the data and the processing are opaque to the user.

As an example of these sorts of interpretative issues, consider Topsy's Sentiment score. The Topsy Sentiment Score is relayed through the News Ltd embedded application, *Poll Pulse*.



http://www.couriermail.com.au/news/special-features/ruddeffect-on-the-wane-as-abbott-retains-the-people8217s-trust/story-fnho52jo-1226683181964

How is the Sentiment Score calculated? What qualifies as positive or negative sentiment? The shaded region above corresponds to the Gnip data, and it shows increasing negativity for both Rudd and Abbott to the end of July. This bears little relationship to my chart of negative sentiment below, which is a plot of mentions of Rudd and Abbott filtered to lies, spin, corruption, treachery, insanity, stupidity, incompetence, cowardice, hypocrisy, and arrogance.



Which is correct? I can at least validate mine on the case data used, and from first principles if necessary.

Performance

The system presented here is feasible for up to a million or so tweets on an ordinary business Dell or similar. The machine used for the above is dual core, 4 gig RAM, nothing fancy, and no accelerations. Higher volumes may be possible using SSDs and other machine enhancements. The bottleneck is the Damerau-Levenshtein step on the tweet text, which for the above 46 categories on 114 megabytes of plain text, took about 15 hours to complete. Performance is linear to the number of individual target synonyms/variants. The number of category codes they are allocated over makes no material difference. Damerau-Levenshtein on the hashtags, a much smaller set of targets, completed in about 20 minutes.

The major time commitment from a human is in devising the target synonym and variants lists. For the Twitter data analysed above, that required several hours. For more routine applications of the technique, such as open-ended brand lists, preparing the targets is trivial.

References

[1] The original publication was

V. I. Levenshtein. *Binary codes capable of correcting deletions, insertions and reversals*. Doklady Akademii Nauk SSSR **163**(4) p845-848, 1965, also Soviet Physics Doklady **10**(8) p707-710, Feb 1966.

See also E. Ukkonen *On approximate string matching*. Proc. Int. Conf. on Foundations of Comp. Theory, Springer-Verlag, LNCS **158** p487-495, 1983.

[2] For a short description, walk-through and implementations in Java, C++ and VB.Net, see http://people.cs.pitt.edu/~kirk/cs1501/Pruhs/Spring2006/assignments/editdistance/Levenshtein%20Distance.htm). The algorithm is short and requires only casual programming skills to implement.

See also Berghel, Hal ; Roach, David : *An Extension of Ukkonen's Enhanced Dynamic Programming ASM Algorithm* http://berghel.net/publications/asm/asm.php

There is an interactive implementation of Damerau-Levenshtein at http://fuzzy-string.com/Compare/ for experimenting with different source and target strings.

[3] Bruns, Axel; Stieglitz, Stephan: *Quantitative Approaches to Comparing Communication Patterns on Twitter*, Journal of Technology in Human Services, Volume 30, Issue 3-4, 2012 page 20

[4] Gnip conducts historical searches which return as many matching records as possible over multiple social media platforms. The data is however expensive, with the minimum price set at $US600. Data is collected in JSON (JavaScript Object Notation) format, which can be a challenge to parse. See http://gnip.com/

ScraperWiki is cheap ($9.00 per month) or free (for lesser volumes) and near real-time, but cannot guarantee that a retrieval is complete. See https://scraperwiki.com/.

[5] A brief synopsis of recent Australian federal politics:

In 2007, Kevin Rudd, a traditional Labor Party outsider with no factional or union backing but substantial personal popularity, defeated the long-serving (12 years) conservative government of John Howard. Although enjoying a sustained honeymoon, his position weakened in 2010 in the aftermath of several botched stimulus programs, a poorly considered resources tax and a failure of nerve in pursuing an ETS. His enemies (primarily the traditional union base of the party and disaffected cabinet members) exploited this to have him overthrown and replaced by Julia Gillard, Australia's first female prime minister. Gillard immediately called an election which resulted in a hung parliament. Gillard governed with the support of the Greens, three independents and a majority of two. Gillard came to the job with an excellent reputation for ministerial efficiency, but soon became embroiled in controversy when her famous election promise 'there will be no carbon tax under a government I lead' was broken after a mere three months. Dogged by old scandals from her days as a union lawyer and multiple ministerial resignations, Gillard made serial policy errors, and the polls plummeted to lows of 40:60 after preferences. Meanwhile, Rudd fomented trouble from the back bench, launching two bids to regain his position. He was successful on 26 June 2013, due primarily to the Labor members' fear of losing their seats in a wipe-out under Gillard.

Tony Abbott became leader of the opposition on defeating Malcolm Turnbull in December 2009 over Turnbull's support for an ETS. Turnbull is very popular among the youth – supports a carbon dioxide tax, trendy image – but is not popular with the party faithful. Abbott, a conservative Catholic, is conversely popular within the party but not the people. He nevertheless won the September 2013 election comprehensively.

The major issue of the election was border protection. One of Rudd's first term actions was to dismantle the tough policies put in place by John Howard. The immediate effect was a huge escalation in illegal entries by boat from Indonesia.

Rudd, although not agreeing that he was wrong to relax border protection and irregular entry requirements, embarked on a new policy called the PNG-solution, ostensibly more punitive than anything the conservatives would ever consider. The deal was to give Papua New Guinea control over Australian foreign aid (previously tightly designated due to their chronic public sector corruption), and PNG takes the refugees permanently – they never get to Australia. This policy is predicated on the assumption (probably true) that life in PNG will be perceived as worse than in their country of origin.

A humorous account of this history done as a *Game of Thrones* parody is contained in the first several minutes of http://www.youtube.com/watch?v=QWU6tVxzO1I

A good sense of the fear and loathing from the progressive left with regard to Tony Abbott is encapsulated in this campaign ad by the Greens:

http://www.youtube.com/watch?feature=player_embedded&v=14vBe-PrJ-E

[6] http://pickeringpost.com

[7] See http://www.theguardian.com/world/2013/jul/08/tony-abbott-book-tour-expenses

Also http://www.dailytelegraph.com.au/news/nsw/opposition-leader-tony-abbott-sparks-storm-on-twitter-after-telling-bridie-jabour-to-calm-down/story-fni0cx12-1226676460305

which claims 2,000 tweets using #CalmDownBridie.

[8] http://topsy.com/ provides an annual subscription model, but it is priced at a corporate level, and requires a substantial budget.