Text Mining Report

Data

The data collected for this project consists of 101 opinion articles from ‘The Age’ (http://www.theage.com.au), an Australian newspaper. All articles are authored by Waleed Aly.

The web crawler used to scrape this content was adapted from that produced by myself, Denitsa Panova and Zsuzsa Holler as part of our homework assignments for ‘Text Mining for Social Sciences’. The general structure of the crawler is as follows. It takes the url of the first page of some search result, it finds the total number of pages relating to this search result, it scrapes the urls of the destination pages from each of these search result pages and finally scrapes the the html content from each one of these destination pages.

Specifically, for this project:

- The url of the first search page result that it takes is: http://www.theage.com.au/comment/by/waleed-aly
- Further result page urls are found to be of the form:
    - Page 2: http://www.theage.com.au/comment/by/waleed-aly?offset=20
    - Page 3: http://www.theage.com.au/comment/by/waleed-aly?offset=40, etc
- 20 article (destination) urls are present and scraped from each of the result pages (except the last result page, which potentially has less)
- Rather than saving the entire html from each destination page, only the following features are returned:
    - Article title
    - Article date
    - Number of comments reader have left on the page[1]
    - Article text
- The features of each article are saved as a string of text with identifiable separators.
- A random delay is included between scraping the html of each article to avoid being kicked off the site.
- Article logs are kept at two stages of the scraping process:
    1. Articles urls for which the html content is able to be scraped are saved to a “completed_html_articles” file and those for which it cannot be scraped are saved to a “rejected_html_articles” file.
    2. For articles whose html is correctly scraped, if the four features described above can be extracted, the article url is saved to a “completed_formatting_article_urls”

[1] The original plan was to scrape the number of facebook shares, however it was not possible to scrape this information. The number of comments was scraped as an alternative.

file. If any problems arise, the urls is saved to a “rejected_formatting_article_urls” file.

- Every 10 articles, the data is saved to disk.

During the scraping process, a substantial number of articles were rejected at the formatting stage. For the majority of articles this was due a problem with scraping the number of comments when the number of comments was zero. The crawler code was updated to deal with these instances and these articles were scraped in a second round.

After the scraping process the data was parsed and the document term matrix was constructed. Below is a histogram showing the distribution of word counts (stemmed and stopped) for the articles.

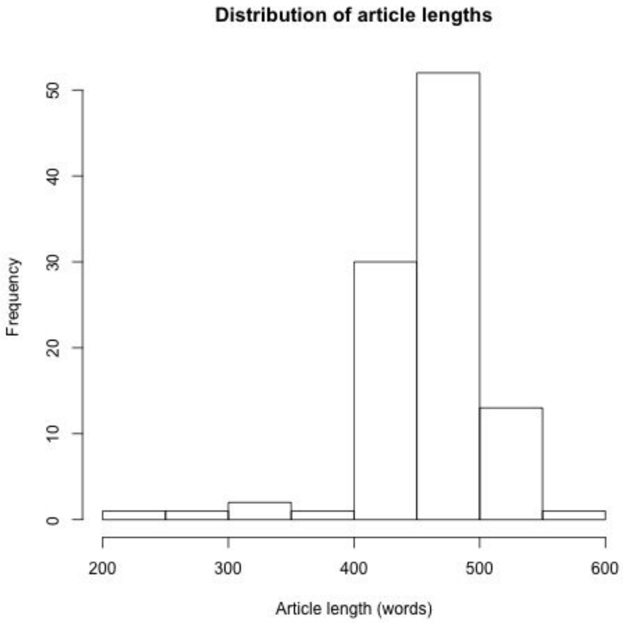


Figure 1. Histogram of article length. Words have been stemmed and stopped.

From the plot below of tokens, ranked by usage across the entire corpus, demonstrates the expected power law distribution.

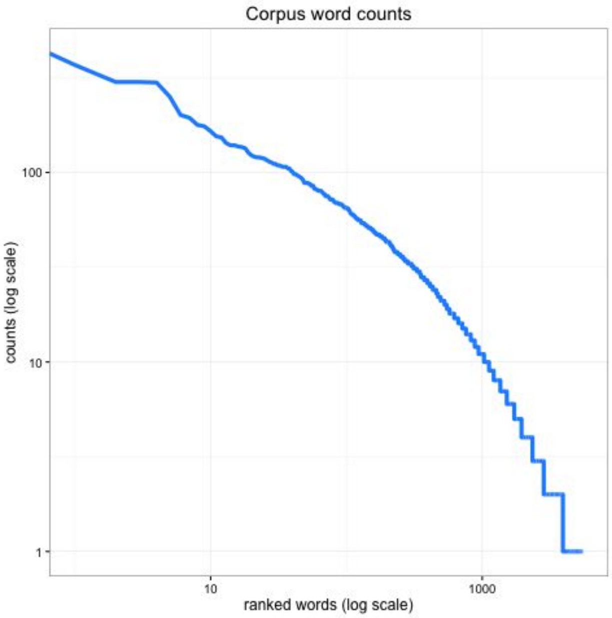


Figure 2. The plot above demonstrates the power law distribution of word counts across the corpus (Zipf's law). Few words have high frequency counts and many words have low frequency count. Words have been stemmed and stopped.

Question

In what is a tumultuous time for Muslims in Australia, many followers of the faith have been subject to prejudice and racism. Whilst the media is filled with reports of verbal attacks on women wearing hijabs, protests against the construction of mosques and islamic schools, and far right political parties garnering support, the Gold Logie for Best Personality on Australian Television (Gold Logie) was this year awarded to Waleed Aly, an openly Sunni Muslim television presenter.

The Gold Logie is voted for by the Australian public and is essentially an award for the most popular personality on television. Aly won the award for his role on the popular culture current affairs program 'The Project' but is also writes an opinion column for The Age newspaper. The question is: In a time when so much hatred is directed towards Muslim communities of Australia, what is Aly saying that is making him so popular? In this project I aim to investigate this question by analysing the text from opinion articles in The Age, authored by Aly and comparing these to a proxy for popularity.

It is important to note a couple of assumptions.

1. The topics Aly covers in his article are similar to those which he talks about on 'The Project'. It is possible that his opinion columns have a slightly more political stance than topics discussed on 'The Project' due to audience difference.

2. The number of comments from the public on each article on the The Age's website are a proxy for popularity. It is noted that this is perhaps not the best proxy since comments might not always be positive. However, I assume that more comments on a particular article lead to people being more aware about that article and more awareness of an article leads to it potentially being more popular.
3. Some variation occurs in his popularity.

Extracting Content

Text Cleaning

In order to create a corpus that could be used to create meaningful analysis the following techniques were applied to the text data in the order listed.

- Tokenization using the 'wordpunct_tokenizer' from the nltk package
- Token cleaning: accepting only tokens composed of completely alpha characters and having at least three characters
- Stopword removal using the 'stopwords.txt' file provide to us as part of the homework assignments for this course.
- Stemming using the Porter Stemmer
- Stems of length less than three were discarded since they were unintelligible.

Content Analysis Approaches

With the aim of discovering some basic relationships between sentiment, topics and popularity, the first approach considered was dictionary methods. The following dictionaries were used:

- AFINN - this dictionary consists of a list of English words rated for valence from minus five (negative) and plus five (positive).
- Politics (Harvard IV-4)
- Ethics (Harvard IV-4)

Since article lengths are clearly not equal and word counts follow a power law distribution, the dictionary scores for each article were calculated using the tf-idf matrix, rather than the document term matrix.

The second approach considered was Latent Dirichlet Allocation (LDA). This was used in order to identify which mixtures of topics are the more popular. LDA was implemented using the 'lda' package in python which uses collapsed Gibbs sampling.

Results

Dictionary Methods

AFINN: The histogram below shows that there are far more articles with negative sentiment than positive sentiment. This raises the question of what type of topics are covered in the articles and suggests that they are perhaps hard hitting topics. If articles with zero comments are ignored in the second plot, there appears to be a weak inverse relationship between article sentiment and the number of comments it generates. This suggests that readers are more moved by negative articles than positive ones.

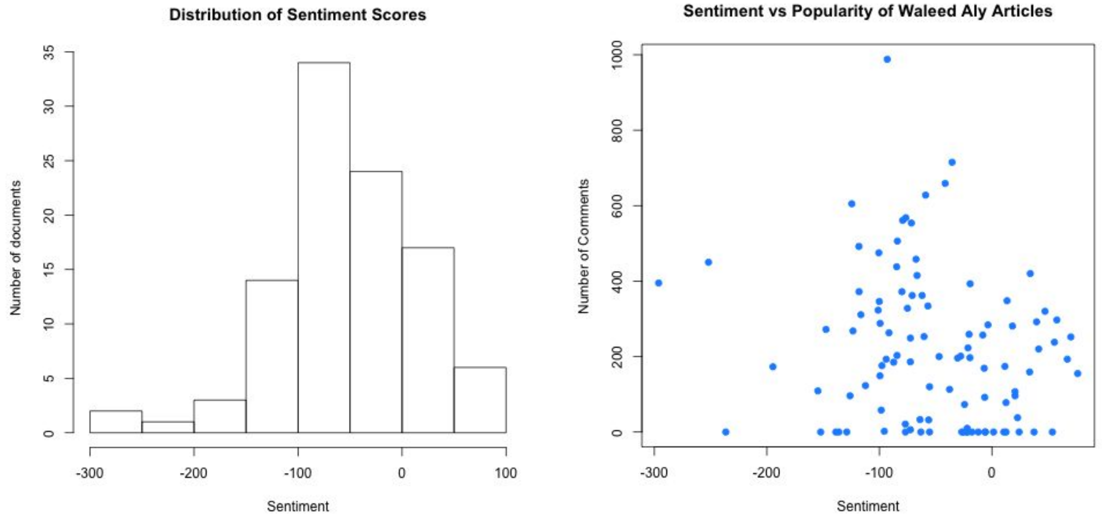


Figure 3. Histogram of article sentiment scores (left) and number of comments made on the article versus article sentiment (right).

POLITICS: Histogram shows us that there is a reasonable wide range of politics scores across the articles. If articles with zero comments are ignored, then there appears to be a weak inverse relationship between the level of politics in an article and the number of comments it generates. This suggests that articles which spend more time covering political issues, generate less interest in the audience and by extension are less popular.

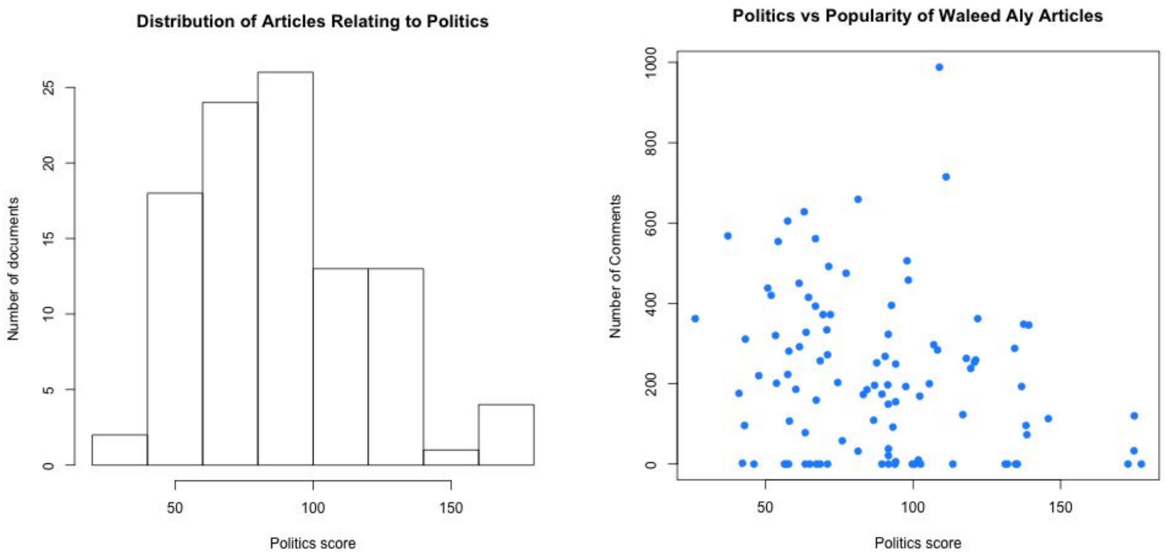


Figure 4. Histogram of article politics score (left) and number of comments made on the article versus article politics scores (right).

ETHICS: Histogram shows us that distribution of ethics scores across articles is skewed and that about half the articles have a very similar score to one another. If articles with zero comments are once again ignored, then there appears to be a very weak inverse relationship between the level of ethics in an article and the number of comments it generates.

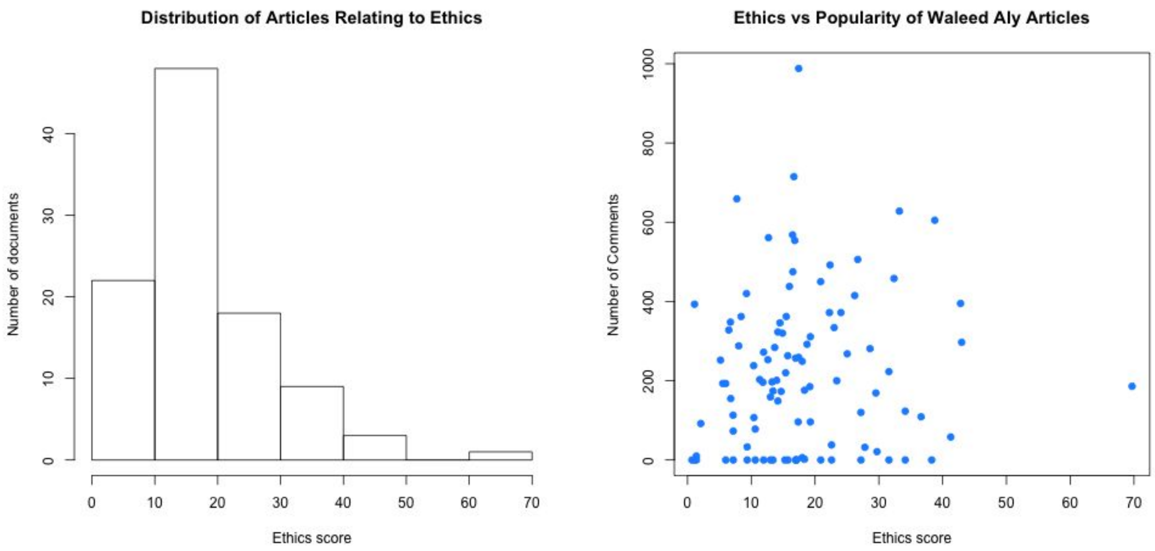


Figure 5. Histogram of article ethics score (left) and number of comments made on the article versus article ethics scores (right).

It is important to notice that very few articles with high numbers of comments have high politics or ethics scores. Perhaps audiences are more interested in articles that cover many topics, not just one.

LDA

Various different values of k (number of topics) were tested. The final value of k was determined by which value appeared to give the best separation into identifiable topics. The top five most probable words (stemmed) for each topic for different values of k are displayed in the following tables.

k	Top 5 words in each topic	Identified Topic
2	Topic 0: can power countri someth nation Topic 1: govern polit abbott will labor	- Domestic politics
3	Topic 0: peopl can simpli fact will Topic 1: power nation world state terror Topic 2: govern polit abbott labor year	Positive Terrorism Domestic politics
4	Topic 0: state nation terror militari want Topic 1: can world kind polit countri Topic 2: govern abbott polit labor year Topic 3: peopl fact thing point can	Terrorism/war World politics Domestic politics -
5	Topic 0: peopl simpli will fact can Topic 1: polit might power marriag week Topic 2: govern abbott polit labor will Topic 3: state nation terror war inde Topic 4: problem world can kind argument	Power of the people/positive - Domestic politics Terrorism/war -
6	Topic 0: can govern simpli question australia Topic 1: someth peopl terror problem australian Topic 2: state polit will right marriag Topic 3: world polit media cultur countri Topic 4: peopl mere polici nation death Topic 5: govern abbott labor polit year	Domestic politics 1 (domestic terrorism?) - World poltics - Domestic politics 2

Table 1. The top five most probable words in each topic for different values of k. I have tried to identify each topic.

It was decided that the most informative topic separation was achieved using three topics.

The next step was to try to identify which of the three topics generated the most comments. In order to calculate this, the number of comments from each article were assigned to the topics according to what percentage of that article is assigned to each topic.

Example

Article 1: "Who's in, who's out: Trump's version of modern society"
155 comments

	Topic 0	Topic 1	Topic 2	Total
Topic allocations	0.296	0.406	0.298	1
Allocation of comments	45.9	62.9	46.2	155

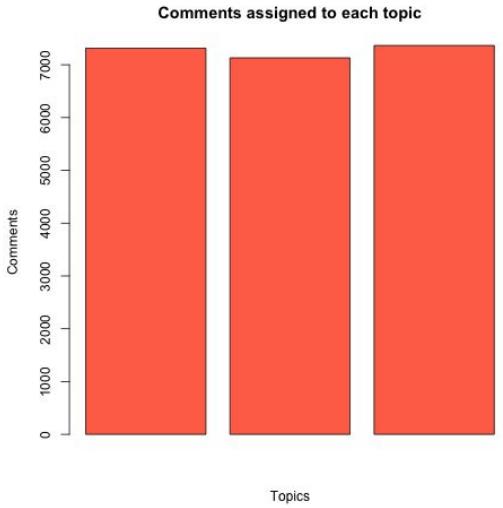


Figure 6. Allocation of comments from individual articles to topics. Topics from left to right are: Topic 0, Topic 1 and Topic 2.

As can be seen from the results shown in Figure 6, there is no clear separation between the topics in terms of how many comments they generate. Topic 2 generates slightly more comments than the other two topics.

Carrying out the same process for other values of k does not produce much better results. Refer to appendix for plots.

Conclusion

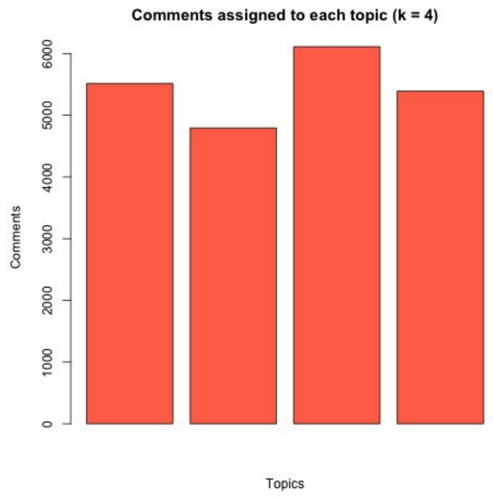
Whilst dictionary methods were able to identify some relationship between popularity and topics, LDA was only able to distinguish between topics but not make any inference on which ones are more popular. To really determine what is driving the popularity of Waleed Aly it is necessary to have a better measure of popularity (such as facebook likes) and a larger corpus. Also an analysis would be much clearer using transcripts from ‘The Project’. These could then be compared across presenters. However, one limitation of this idea as whole is that wit, a characteristic common to popular presenters is hard to capture using traditional text mining techniques.

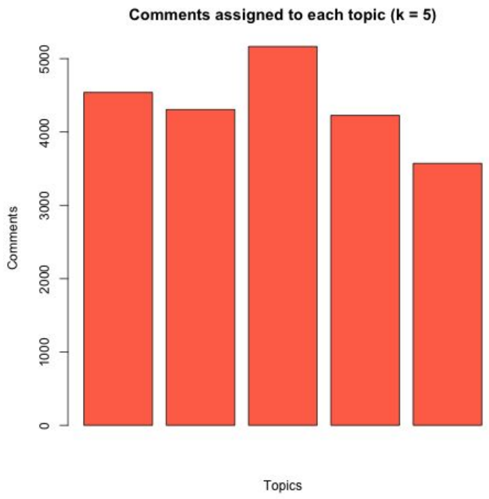
Code

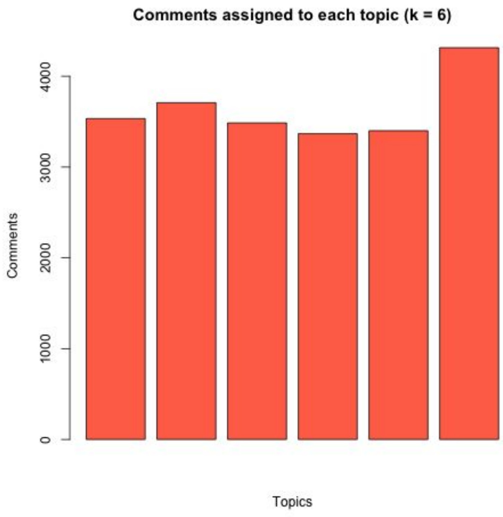
The code relating to this project can be found in the following github repository: https://github.com/aspeijers/text_mining_homework/tree/master/Project

Appendix

Comments per topic plots for other values of k.


Comments assigned to each topic (k = 4)
Comments
Topics


Comments assigned to each topic (k = 5)
Comments
Topics


Comments assigned to each topic (k = 6)
Comments
Topics