Life still goes on: 
Analysing Australian WW1 Diaries through Distant Reading

Ashley Dennis-Henderson, Matthew Roughan, Lewis Mitchell, Jonathan Tuke
ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS)
School of Mathematical Sciences, The University of Adelaide
{ashley.dennis-henderson,matthew.roughan,lewis.mitchell,симон.туке}@adelaide.edu.au

Abstract

An increasing amount of historic data is now available in digital (text) formats. This gives quantitative researchers an opportunity to use distant reading techniques, as opposed to traditional close reading, in order to analyse larger quantities of historic data. Distant reading allows researchers to view overall patterns within the data and reduce researcher bias. One such data set that has recently been transcribed is a collection of over 500 Australian World War I (WW1) diaries held by the State Library of New South Wales. Here we apply distant reading techniques to this corpus to understand what soldiers wrote about and how they felt over the course of the war. Extracting dates accurately is important as it allows us to perform our analysis over time, however, it is very challenging due to the variety of date formats and abbreviations diarists use. But with that data, topic modelling and sentiment analysis can then be applied to show trends, for instance, that despite the horrors of war, Australians in WW1 primarily wrote about their everyday routines and experiences. Our results detail some of the challenges likely to be encountered by quantitative researchers intending to analyse historical texts, and provide some approaches to these issues.

1 Introduction

World War I (WW1) was a defining event of the 20th century, and impacted millions worldwide. Researchers have studied the war, especially the experiences of those on the front lines. Primarily, this has been done through close reading of primary sources such as diaries and letters. However, recent advances in computational methods to analyse large text corpora offers the opportunity to analyse sources such as these through distant reading. Distant reading involves the application of mathematical and computational techniques from natural language processing (NLP) to perform statistical analysis of text (Jänicke et al., 2015). Distant reading has several advantages, including the ability to analyse large quantities of data and see overall patterns as well as the reduction of researcher bias. Further, distant and close reading can be combined such that interesting patterns found through distant reading can be more closely examined using close reading to determine why they occur. This work aims to use distant reading to understand what Australian soldiers went through and how they felt over the course of WW1, by analysing a unique historical data set: a large collection of transcriptions of Australian soldiers’ diaries, held by the State Library of New South Wales. To our knowledge this paper represents the first NLP analysis of this data set.

This research takes advantage of the fact that diaries contain temporal information. However, extracting dates is a difficult task due to the varying manner in which dates can be written. This is further complicated by the desire to focus on the dates on which entries were written and not dates mentioned within the entries as these may refer to times and events from outside the war or at least out of the context of the current entry. In order to extract and clean dates we use a combination of regular expressions and optimisation.

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.
Once dates were extracted we were able to apply topic modelling and sentiment analysis as a function of time. We are able to detect topics corresponding to particular developments of the war, and the associated sentiment for those periods. Further, we show that the diarists wrote more about everyday experiences, e.g., the time of day and meals, than they did about training and battles. This might be surprising as the war was one of the most traumatic events of the twentieth century and conventional historical narratives concentrate on the pain and suffering of the soldiers. However, in the diaries, we see the war’s participants adapting their everyday lives to their circumstances, and in fact their overall sentiment across the war is surprisingly positive.

2 Corpus

We focus on Australian WW1 diaries held by the State Library of New South Wales. After the war ended, the European War Collecting Project was created by Principal Librarian William Ifould and the trustees of the Library (State Library of New South Wales, 2019). Their aim was to collect documents, including diaries, letters, war narratives, memoirs and photographs, which gave the experiences and personal feelings of those who served. In total, this collection has 966 documents, 557 of which are non-empty war diaries. A complete breakdown of the collection is given in Table 1. Since collecting these documents, the library has scanned them and used crowd sourcing to transcribe them, giving researchers access to digital (text) versions of the documents.

| Type           | Number | # Pages | # Words    | # Authors |
|----------------|--------|---------|------------|-----------|
| Diary          | 577    | 60,004  | 9,266,353  | 236       |
| Letter         | 183    | 18,497  | 3,029,163  | 141       |
| Letter-Diary   | 22     | 3,955   | 639,184    | 16        |
| War Narrative  | 32     | 2,370   | 624,618    | 28        |
| Other          | 152    | 10,418  | 2,159,348  | 111       |
| **Total**      | **966**| **95,244**| **15,718,666**|           |

Table 1: The number of each type of document in the NSW State Library Collection, along with the number of pages, words and authors. The “Other” category includes documents such as telegrams, photos, postcards, scrapbooks, and newspaper clippings. Note, there are a total of 577 diaries in this collection, however, 20 of the transcribed diaries were empty, and so we only analyse 557 diaries.

These documents can be individually downloaded from the State Library’s website (State Library of New South Wales, 2020), however, we obtained the corpus directly from the Library. Before this corpus could be used it went through a variety of cleaning steps. First, the raw data was converted to a single text file per document, and a metadata table was created by using regular expressions to extract information from the document titles. Then dates had to be extracted so that we could perform analysis over time. Raw dates were extracted using regular expressions, however, several issues were found in these raw dates requiring us to clean them through optimisation. More information regarding this is given in Section 4.1. Finally, we changed all text to lowercase, removed numbers and punctuation, singularised words, converted abbreviations to the full word, and for topic modelling we removed stop words. Converting to lowercase, singularising words and converting abbreviations were all done to ensure that the various versions of a word are considered as the same word when performing our analysis. For example, “kill”, “killing”, and “Kills” all have the same base word: “kill”. The stop words we removed were based on the stop_words data set in the tidytext package (Silge and Robinson, 2006) in R. Figure 1 shows the number of words in our diary corpus per month after this process. Unsurprisingly, the majority of entries were written between August 1914 and December 1919 — Britain, and consequently
Australia joined WW1 on August 4, 1914, and although the armistice was signed in November 1918, it took some time for the more than 100,000 Australian soldiers still in the field to be repatriated.

![Number of words written in our entire diary collection per month. The majority of entries are written between August 1914 and December 1919, however there were some entries as late as 1923.](image)

**Figure 1:** Number of words written in our entire diary collection per month. The majority of entries are written between August 1914 and December 1919, however there were some entries as late as 1923.

3 Related Work and Background

3.1 Analysis of Historic Documents

Our corpus has previously been studied by Caulfield (2013) and Cochrane (2015). However, their analysis was based on close reading of a small subsection of the diaries. As far as we are aware, distant reading techniques have not previously been applied to this corpus. However, distant reading techniques have been broadly applied to other historic documents. For example, Boschetti et al. (2014) used computational techniques to analyse Italian war bulletins as part of the *Memories of War* project and Ahmad et al. (2012) developed a tool to map spelling from medieval documents to modern spellings, amongst numerous other examples. Analysis of diaries presents an additional challenge as to use the important temporal data, we must extract a large number of dates.

3.2 Analysis Techniques

Topic modelling is based on the idea that documents are made up of a series of topics, which in turn are a probability distribution over words (Steyvers and Griffiths, 2007). Currently, the primary method to perform topic modelling is LDA (Latent Dirichlet Allocation) which was initially introduced by Blei et al. (2003). For a description of the mathematics behind LDA please see Blei et al. (2003). Sentiment analysis aims to determine the attitude or emotion of the author towards the content of the text. An overview of sentiment analysis can be found in Pang and Lee (2008) or Taboada (2016). An example of the use of sentiment analysis can be seen in Burghardt et al. (2019) who applied sentiment analysis to the plays of G. E. Lessing. Additional details of our use of these approaches will be provided in the following section.

4 Methods

4.1 Date Extraction

Extracting accurate dates from the diaries is important as we wish to perform our analysis over time. However, this is difficult due to the many ways in which dates are written. Raw dates were extracted
using regular expressions, attempting to account for the various date formats, possible abbreviations of month and day of the week names, punctuation, and that some dates were written in French. After extracting these raw dates three main issues were discovered. First many dates were missing the month or year values, as from a human perspective it is not necessary to include this information if it was included in a previous date. In these diaries 13.91\% of dates were missing the month and 53.76\% were missing the year. Second, diarists sometimes wrote the wrong date, either due to not knowing the exact date or accidentally writing down the wrong day/month. The final issue is that we only want to extract the dates when the entries were written. However, regular expressions will also pick up dates of events mentioned within an entry as well as strings that look like dates such as 1st battalion, neither of which we wish to focus on here.

We overcame these issues by creating an optimisation program which outputs dates as close as possible to the true date by (i) keeping the dates close to their raw extracted version; (ii) keeping them close to the previous date in sequence; (iii) maintaining the sequence of dates; and (iv) keeping them in the range determined by the known start and end dates of the diary. The optimisation can also exclude dates that appear out of sequence, presenting them as references. We will provide code to perform this task on request.

4.2 Topic Modelling

In this paper we focus on using LDA (Latent Dirichlet Allocation) to perform topic modelling. This model was implemented using the topicmodels package (Grüen and Hornik, 2011) in R, using Gibbs Sampling with 10 topics and a randomly chosen seed of 1915.

The number of topics used was chosen based on four methods, by Arun et al. (2010), Cao et al. (2008), Deveaud et al. (2014), and Griffiths and Steyvers (2004), which were implemented using the ldatuning package (Nikita, 2020) in R. The results from each method are given in Figure 2. Based on this, we find that the optimal number of topics for Griffiths2004 is 8 or more, for CaoJuan2009 is 17 or more, for Arun2010 is 6, 7 or 10, and for Deveaud2014 is 8 - 12. We chose to use 10 topics since this falls in the range of best parameters for three of the methods.

![Figure 2: Results found by applying the four methods for determining the number of topics created by Arun et al. (2010), Cao et al. (2008), Deveaud et al. (2014), and Griffiths and Steyvers (2004). We chose 10 topics as it falls in the optimal range for three approaches.](image)

4.3 Sentiment Analysis

There are three general categories of sentiment analysis: dictionary based methods (DBMs), supervised learning methods, and unsupervised learning methods (Reagan et al., 2017). We focus on DBMs as they can be applied to corpora where there is no previous known information regarding the sentiment. DBMs compare the terms within the corpus with a dictionary of terms with known sentiment values. Let $f^T(w)$
be the frequency of word $w$ in text $T$, and $s_D(w)$ be the sentiment of word $w$ in dictionary $D$, then the average sentiment of the text is given by (Reagan et al., 2017)

$$s_{TD} = \frac{\sum_{w \in D} s_D(w) f_T(w)}{\sum_{w \in D} f_T(w)}.$$ (1)

For our analysis we tested the following dictionaries: AFINN, ANEW, Hulii, Loughran-Mcdonald, NRC, SenticNet, SentiWordNet, and Syuzhet. These dictionaries primarily come from the lexicon package (Rinker, 2018) in R. The two dictionaries not available through this package are AFINN, which was accessed using the tidytext package (Silge and Robinson, 2006), and ANEW which was obtained from Andrew Reagan’s GitHub folder: https://github.com/andyreagan/labMT-simple/tree/master/labMTsimple/data/ANEW. We can consider the percentage of unique words in our diaries which appear in the sentiment dictionaries, and compare this to the Brown Corpus, a standard corpus in NLP analysis. The Brown Corpus contains 1,006,770 words, including 45,215 unique words, from a collection of documents printed in the United States in 1961 (Francis and Kucera, 1971). The words contained in the Brown Corpus were obtained using the zipfR package (Evert and Baroni, 2007). Table 2 gives the number of words and possible sentiment values each dictionary has as well as the percentage of unique words in our diaries and the Brown Corpus which appear in the dictionaries. We note that approximately twice as many unique words from the Brown Corpus are covered by these dictionaries. This is despite the fact that our diary corpus contains more unique words (84,955 words) than the Brown Corpus does. This is likely because none of these dictionaries were created for wartime text.

| Dictionary          | Num. Words | Sentiment Values | WW1 Diaries | Brown Corpus |
|---------------------|------------|------------------|-------------|--------------|
| AFINN               | 2,477      | $(-5, 5)$        | 2.03        | 4.35         |
| ANEW                | 1,034      | $(1, 9)$         | 1.14        | 2.12         |
| Hulii               | 6,874      | $\{-1, 0, -1.05, -1, -2\}$ | 5.09        | 9.86         |
| Lougran-Mcdonald    | 2,702      | $\{-1, 1\}$     | 1.59        | 3.80         |
| NRC                 | 5,468      | $\{-1, 1\}$     | 5.49        | 10.09        |
| SenticNet           | 23,626     | $(-1, 1)$        | 14.71       | 26.67        |
| SentiWordNet        | 20,093     | $(-1, 1)$        | 7.83        | 14.01        |
| Syuzhet             | 10,738     | $(-1, 1)$        | 8.43        | 16.85        |

Table 2: The number of words and possible sentiment values in each of the eight sentiment dictionaries, as well as the percentage of unique words in the diaries and the Brown Corpus which appear in each dictionary. We can see that SenticNet provides the broadest coverage.

In order to compare the results from different dictionaries they are required to be on the same rating scale. As five of the dictionaries are already on the scale $(-1, 1)$ we chose to convert the others to this. AFINN and ANEW were converted to this scale using the formula:

$$x_{new} = \left( \frac{\max_{new} - \min_{new}}{\max_{old} - \min_{old}} \right) \left( x_{old} - \min_{old} \right) + \min_{new},$$ (2)

where $x_{old}$ and $x_{new}$ are the old and new value, respectively, $[\min_{old}, \max_{old}]$ is the old value range, and $[\min_{new}, \max_{new}]$ is the new value range. The Hulii lexicon was converted to the range $(-1, 1)$ by converting any word with a sentiment score of $-2$ to a score of $-1$. This could be done as a score of $-2$ was given to phrases that are always negative, e.g., “too much fun” (Rinker, 2019).

For both topic modelling and sentiment analysis we considered a “document” to be all of the diary entries written in a particular month.
When graphing our results we have applied a rolling mean using the `rollmean` function in the `zoo` package (Zeileis and Grothendieck, 2005) in R, with a rolling window of \( k = 5 \), in order to smooth out noise in the data. Due to the lack of data in 1923, as seen in Figure 1, it is not possible to calculate this rolling mean and hence, results for this year are not included.

5 Analysis

5.1 Topic Modelling

The most probable words for each of our 10 topics are shown in Appendix A. Based on the most probable words, we selected names for each of our topics, hence our topics are: Everyday Life, War at Sea, Egypt, Gallipoli, In the Trenches (Beginning), In the Trenches (Middle), In the Trenches (End), White Christmas, After the Armistice, and Home Again. Note, the most probable words in all three In the Trenches topics are regarding battles, the Western Front and the Middle East. Hence, we differentiate these topics as beginning, middle and end, based on where they peak in Figure 3. The proportion of each of these topics is shown as a function of time in Figure 3.

![Figure 3: The proportion of each topic obtained from our LDA model, over time. Note that a rolling mean with \( k = 5 \) has been applied to each point.](image)

Based on the most probable words as well as when the topics peak in Figure 3, several of these topics relate to specific developments of the war. War at Sea corresponds to the Australian occupation of German New Guinea and the sinking of the German raider the Emden. Egypt corresponds to the training of Australian troops on the outskirts of Cairo and battles around Egypt and the Suez Canal. Gallipoli corresponds to the Gallipoli Campaign, which for many Australian soldiers was their first experience in battle. The three In the Trenches topics cover the period when Australians were fighting on the Western Front and in the Middle East. The peaks in these three topics most likely correspond to specific battles such as the Battle of Romani (August 1916), the Second Battle of Arras (April-May 1917), the Battle of Jerusalem (November-December 1917), and the Battle of Hamel (July 1918). After the Armistice corresponds to the period after the armistice was signed in November 1918 that Australian soldiers had to wait before being sent home. For some soldiers it took up to a year to be repatriated and in this time they travelled around France and Britain as well as receiving vocational training from the AIF (Australian Imperial Force) (DVA, 2020). We also have two more general topics. Everyday Life is consistently the most prominent topic until December 1919. This topic includes words related to everyday things such as the time of day and meals. This shows that whilst the diarists did write about war related things, such as training and battles, they primarily focused on their ordinary day-to-day activities. After 1919 the Home
Again topic becomes most prominent. This is expected as this topic contains words related to being back in Australia, such as “mum”, “dad” and “shopping”, and corresponds to when the soldiers would have returned home.

5.2 Sentiment Analysis

Figure 4 gives the sentiment scores for our diaries over time for the eight sentiment dictionaries we considered as well as the average over these dictionaries. From this graph we first note that five of the dictionaries: AFINN, Huliu, Loughran-Mcdonald, NRC, and Syuzhet, follow the same general pattern. Further, SenticNet and SentiWordNet have a similar trend. Based on Table 2 we know that ANEW covers the least amount of words in our corpus, whilst SenticNet and SentiWordNet cover the most. This shows that our analysis is dependent on the words covered in the dictionaries. We also observe more variability in our sentiment scores in the first half of 1914 and from 1920 onwards. This would be due to only having a small amount of data for those periods as seen in Figure 1.

In the next section we compare our average sentiment curve with our topic model to understand why the sentiment peaks and dips at certain times.

\[\text{Figure 4: Sentiment scores over time for the eight dictionaries: AFINN, ANEW, Huliu, Loughran-Mcdonald, NRC, SenticNet, SentiWordNet, and Syuzhet, as well as the average of these dictionaries. Note, that before graphing we have applied a rolling mean, with } k = 5, \text{ to each of the dictionaries.}\]

5.3 Topic Modelling and Sentiment Analysis

The average sentiment curve shown in Figure 4 has several peaks and dips in sentiment. We investigate what these correspond to by comparing our sentiment with our topic model. Due to the variability in individual sentiment dictionaries prior to August 1914 and after December 1919 we do not consider these periods. Further, we exclude the Everyday Life and Home Again topics as they are prominent over large periods of time and hence are not likely to contribute to particular peaks and dips in sentiment. Figure 5 gives the comparison between topic probabilities and average sentiment scores.

In Figure 5 we note there are peaks in sentiment corresponding to peaks in the Egypt and After the Armistice topics, whilst there are dips in sentiment corresponding to the peaks in the Gallipoli and White Christmas topics. When arriving in Egypt for training, the soldiers would most likely have been excited about being in a new country and be keen to prove themselves in battle. This, combined with the fact that whilst in Egypt the men were able to take small trips into Cairo and around the pyramids, would lead to a more positive sentiment for that period. Contrary to this, the Gallipoli campaign would have been
the first battle experience for many of the soldiers leading to a more negative sentiment. Thomas Munro writes:

“It is an awful sight to see the dead and wounded on both sides, lying out and being walked on, no possibilitie [sic] to bring them in or bury them. Some of our men have been out there a month and are still there. The stench would knock you down.”

One of the top 40 most probable words in the White Christmas topic is “miserable”, suggesting that the cold weather lead to many having a more negative sentiment. Through close reading of the diaries over the months surrounding January 1917 we find several negative comments regarding the cold and wet weather. For instance, Langford Colley-Priest writes

“Raining heavily all day which made the conditions more miserable. The mud & slush is terrible.”

Further, whilst some men had a good Christmas, others didn’t. The contrast between these Christmas’ are seen in the following quotes:

“Christmas dinner and tea were very merry, the rations being supplemented by a lot of luxuries ... also by plum pudding ... ”, Hector McLean

“Cold, miserable & hungry, we filed up to the cook house for our “Christmas dinner” of Bully beef Stew and biscuits [sic], as our rations were not yet to hand and our Christmas comforts were delayed somewhere.”, Tom Taylor

It is not surprising that the sentiment rose after the armistice was signed. This rise in sentiment is further strengthened by the fact whilst waiting to be repatriated back to Australia soldiers spent their time travelling around France and Britain, and attending sport matches and plays (DVA, 2020).

Overall, our average sentiment during the war is always slightly positive which contradicts the typical perception of the war as horrific experience. This is most likely because the diarists predominantly wrote about everyday activities, which unlike battles, are not necessarily negative.
6 Conclusion and Future Work

This research aimed to analyse Australian WW1 diaries in order to determine what the soldiers wrote about and how they felt over the course of the war. Through the application of distant reading techniques we have seen that we can analyse large amounts of data to determine trends. Interestingly, while many people typically think of the war as a horrific experience we find that the diarists primarily wrote about their day-to-day activities. As such the diaries had an overall slightly positive sentiment, which is consistent with the positivity bias seen across human languages (Dodds et al., 2015), but is surprising for this particular corpus.

We focused on DBMs for sentiment, and found that the dictionaries used covered less of the diaries than standard texts such as the Brown Corpus. This suggests that DBMs may not be the most accurate method for determining sentiment in WW1 diaries and as such in the future we will investigate other sentiment analysis techniques such as embedding-based methods to determine if they are more applicable. Further, in the future we will write a paper detailing the difficulties with date extraction, as well as our approach and the accuracy of our method, as this is not a trivial issue.

Acknowledgements

We acknowledge the State Library of New South Wales for providing the data which made this research possible.

References

Mushtaq Ahmad, Stefan Gruner, and Muhammed Tanvir Afzal. 2012. Computational Analysis of Medieval Manuscripts: A New Tool for Analysis and Mapping of Medieval Documents to Modern Orthography. *Journal of Universal Computer Science*, 18(20):2750–2770.

R Arun, V Suresh, C.E Veni Madhavan, and M Narasimha Murty. 2010. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In M.J. Zaki, J.X. Yu, B Ravindran, and V Pudi, editors, *Advances in Knowledge Discovery and Data Mining, Part I*, pages 391 – 402. Springer, Berlin. Heidelberg.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Manuel Burghardt, Christian Wolff, and Thomas Schmidt. 2019. Toward Multimodal Sentiment Analysis of Historic Plays: A Case Study with Text and Audio for Lessing’s Emilia Galotti. In *4th Conference of the Association Digital Humanities in the Nordic Countries*, Copenhagen.

Juan Cao, Tian Xia, Jintao Li, Yongdong Zhang, and Sheng Tang. 2008. A density-based method for adaptive LDA model selection. *Neurocomputing*, 72(7-9):1775 – 1781.

Michael Caulfield. 2013. *The unknown Anzacs: the real stories of our national legend: told through the rediscovered diaries and letters of the Anzacs who were there*. Hachette Australia, Sydney.

Peter Cochrane. 2015. ‘Diamonds of the Dustheap’: Diaries from the First World War. *Humanities Australia: The Journal of the Australian Academy of the Humanities*, (6):22 – 33.

Romain Deveaud, Eric SanJuan, and Patrice Bellot. 2014. Accurate and effective Latent Concept Modeling for ad hoc information retrieval. *Document Numerique*, 17(1):61–84.

Peter Sheridan Dodds, Eric M Clark, Suma Desu, Morgan R Frank, Andrew J Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M Kloumann, James P Bagrow, et al. 2015. Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, 112(8):2389–2394.

DV A. 2020. Repatriation of Australians in World War I. *DVA (Department of Veterans’ Affairs) Anzac Portal*.

Stefan Evert and Marco Baroni. 2007. ziptR: Word Frequency Distributions in R. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Sessions*, Prague.
W Francis and H Kucera. 1971. Brown Corpus Manual.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America, 101:5228–5235.

Bettina Grün and Kurt Hornik. 2011. topicmodels: An R Package for Fitting Topic Models. Journal of Statistical Software, 40(13):1–30.

S Jänicke, G Franzini, M F Cheema, and G Scheuermann. 2015. On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. In R Borgo, F Ganovelli, and I Viola, editors, Eurographics Conference on Visualization (EuroVis).

Murzintcev Nikita. 2020. Idatuning: Tuning of the Latent Dirichlet Allocation Models Parameters.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(2):1–135.

Andrew J Reagan, Christopher M Danforth, Brian Tivnan, Jake Ryland Williams, and Peter Sheridan Dodds. 2017. Sentiment analysis methods for understanding large-scale texts: a case for using continuum-scored words and word shift graphs. EPJ Data Science, 6(1).

Tyler Rinker. 2018. lexicon: Lexicon Data.

Tyler Rinker. 2019. Package ‘lexicon’.

Julia Silge and David Robinson. 2006. tidytext: Text Mining and Analysis Using Tidy Data Principles in R. JOSS, 1(3).

State Library of New South Wales. 2019. Personal diaries and letters from the First World War.

State Library of New South Wales. 2020. Diarists from World War I.

Mark Steyvers and Tom Griffiths. 2007. Probabilistic Topic Models. In T Landauer, D McNamara, S Dennis, and W Kintsch, editors, Handbook of latent semantic analysis, pages 427–448. Lawrence Erlbaum Associates Publishers.

Maite Taboada. 2016. Sentiment Analysis: An Overview from Linguistics. Annual Review of Linguistics, 2(1):325–347, 1.

Achim Zeileis and Gabor Grothendieck. 2005. zoo: S3 Infrastructure for Regular and Irregular Time Series. Journal of Statistical Software, 14(6):1–27.

**Appendix A Topics**

Tables 3 - 12 give the 54 most probable words for each of the topics found using topic modelling.
| rank | term   | beta | rank | term  | beta | rank | term   | beta |
|------|--------|------|------|-------|------|------|--------|------|
| 1    | day    | 0.0213 | 19   | fine   | 0.0033 | 37   | troop  | 0.0022 |
| 2    | night  | 0.0132 | 20   | till   | 0.0032 | 38   | battalion | 0.0022 |
| 3    | morning | 0.0113 | 21   | evening | 0.0032 | 39   | horse   | 0.0022 |
| 4    | time   | 0.0095 | 22   | dinner | 0.0031 | 40   | company | 0.0022 |
| 5    | left   | 0.0070 | 23   | received | 0.0030 | 41   | train   | 0.0022 |
| 6    | afternoon | 0.0069 | 24   | water  | 0.0029 | 42   | french  | 0.0021 |
| 7    | pm     | 0.0057 | 25   | weather | 0.0028 | 43   | boy     | 0.0021 |
| 8    | camp   | 0.0050 | 26   | brigade | 0.0027 | 44   | australian | 0.0021 |
| 9    | letter | 0.0048 | 27   | hospital | 0.0026 | 45   | duty     | 0.0021 |
| 10   | arrived | 0.0047 | 28   | found  | 0.0025 | 46   | heavy    | 0.0021 |
| 11   | mile   | 0.0043 | 29   | town   | 0.0025 | 47   | returned | 0.0021 |
| 12   | home   | 0.0043 | 30   | passed | 0.0025 | 48   | breakfast | 0.0020 |
| 13   | tea    | 0.0042 | 31   | usual  | 0.0025 | 49   | station  | 0.0020 |
| 14   | hour   | 0.0041 | 32   | bed    | 0.0024 | 50   | tonight  | 0.0020 |
| 15   | round  | 0.0039 | 33   | cold   | 0.0024 | 51   | party    | 0.0019 |
| 16   | officer | 0.0038 | 34   | lot    | 0.0024 | 52   | war      | 0.0019 |
| 17   | line   | 0.0037 | 35   | parade | 0.0024 | 53   | called   | 0.0019 |
| 18   | leave  | 0.0035 | 36   | light  | 0.0023 | 54   | beautiful | 0.0019 |

Table 3: Top 54 terms for the *Everyday Life* topic with their probabilities.

| rank | term   | beta | rank | term  | beta | rank | term   | beta |
|------|--------|------|------|-------|------|------|--------|------|
| 1    | ship   | 0.0157 | 19   | port  | 0.0043 | 37   | ashore | 0.0024 |
| 2    | sydney | 0.0092 | 20   | deck  | 0.0040 | 38   | drill  | 0.0024 |
| 3    | german | 0.0088 | 21   | harbour | 0.0040 | 39   | crew   | 0.0024 |
| 4    | captain | 0.0074 | 22   | emden | 0.0038 | 40   | flag   | 0.0023 |
| 5    | officer | 0.0073 | 23   | naval  | 0.0035 | 41   | herbertshohe | 0.0023 |
| 6    | boat   | 0.0073 | 24   | administrator | 0.0034 | 42   | colombo | 0.0023 |
| 7    | board  | 0.0067 | 25   | force  | 0.0034 | 43   | commander | 0.0023 |
| 8    | lieutenant | 0.0065 | 26   | horse  | 0.0033 | 44   | australia | 0.0023 |
| 9    | island | 0.0064 | 27   | melbourne | 0.0031 | 45   | holme  | 0.0022 |
| 10   | troop  | 0.0058 | 28   | major  | 0.0030 | 46   | returned | 0.0022 |
| 11   | native | 0.0056 | 29   | government | 0.0029 | 47   | brigadier | 0.0021 |
| 12   | colonel | 0.0050 | 30   | cruiser | 0.0027 | 48   | sight  | 0.0020 |
| 13   | wireless | 0.0049 | 31   | station | 0.0027 | 49   | convoy | 0.0020 |
| 14   | message | 0.0048 | 32   | fleet  | 0.0027 | 50   | military | 0.0020 |
| 15   | company | 0.0045 | 33   | garrison | 0.0027 | 51   | signal | 0.0020 |
| 16   | rabaul | 0.0045 | 34   | steamer | 0.0026 | 52   | british | 0.0020 |
| 17   | received | 0.0044 | 35   | berrima | 0.0025 | 53   | war    | 0.0019 |
| 18   | sea    | 0.0044 | 36   | gun    | 0.0025 | 54   | prisoner | 0.0019 |

Table 4: Top 54 terms for the *War at Sea* topic with their probabilities.
| rank | term         | beta  | rank | term         | beta  | rank | term         | beta  |
|------|--------------|-------|------|--------------|-------|------|--------------|-------|
| 1    | cairo        | 0.0170| 19   | water        | 0.0040| 37   | serapeum     | 0.0022|
| 2    | canal        | 0.0131| 20   | troop        | 0.0039| 38   | oclock       | 0.0022|
| 3    | camp         | 0.0103| 21   | kebir        | 0.0039| 39   | trench       | 0.0022|
| 4    | horse        | 0.0098| 22   | regiment     | 0.0039| 40   | squadron     | 0.0021|
| 5    | parade       | 0.0097| 23   | train        | 0.0039| 41   | piastre      | 0.0021|
| 6    | ship         | 0.0093| 24   | sea          | 0.0038| 42   | maadi        | 0.0021|
| 7    | sand         | 0.0088| 25   | suez         | 0.0036| 43   | arab         | 0.0021|
| 8    | tent         | 0.0080| 26   | deck         | 0.0035| 44   | colombo      | 0.0021|
| 9    | desert       | 0.0074| 27   | heliopoli    | 0.0035| 45   | soldier      | 0.0020|
| 10   | native       | 0.0071| 28   | sydney       | 0.0033| 46   | colonel      | 0.0020|
| 11   | el           | 0.0066| 29   | pyramid      | 0.0033| 47   | mosque       | 0.0020|
| 12   | drill        | 0.0060| 30   | island       | 0.0030| 48   | infantry     | 0.0020|
| 13   | egypt        | 0.0049| 31   | harbour      | 0.0028| 49   | christmas    | 0.0020|
| 14   | boat         | 0.0046| 32   | hot          | 0.0027| 50   | wharf        | 0.0019|
| 15   | egyptian     | 0.0045| 33   | ashore       | 0.0027| 51   | fuller       | 0.0019|
| 16   | tel          | 0.0043| 34   | nile         | 0.0027| 52   | signalling   | 0.0019|
| 17   | camel        | 0.0043| 35   | ismailia     | 0.0024| 53   | marching     | 0.0019|
| 18   | alexandria   | 0.0040| 36   | port         | 0.0023| 54   | fatigue      | 0.0019|

Table 5: Top 54 terms for the Egypt topic with their probabilities.

| rank | term         | beta  | rank | term         | beta  | rank | term         | beta  |
|------|--------------|-------|------|--------------|-------|------|--------------|-------|
| 1    | turk         | 0.0194| 19   | hospital     | 0.0048| 37   | island       | 0.0029|
| 2    | trench       | 0.0188| 20   | turkish      | 0.0046| 38   | dug          | 0.0029|
| 3    | gun          | 0.0130| 21   | quiet        | 0.0046| 39   | alexandria   | 0.0028|
| 4    | shell        | 0.0120| 22   | hill         | 0.0041| 40   | landed       | 0.0028|
| 5    | wounded      | 0.0101| 23   | rifle        | 0.0041| 41   | hit          | 0.0028|
| 6    | ship         | 0.0093| 24   | cairo        | 0.0041| 42   | aeroplane    | 0.0028|
| 7    | fire         | 0.0085| 25   | killed       | 0.0040| 43   | fired        | 0.0028|
| 8    | enemy        | 0.0081| 26   | line         | 0.0039| 44   | anzac        | 0.0027|
| 9    | firing       | 0.0076| 27   | shot         | 0.0038| 45   | machine      | 0.0027|
| 10   | beach        | 0.0072| 28   | bullet       | 0.0037| 46   | warship      | 0.0027|
| 11   | boat         | 0.0065| 29   | bombardment  | 0.0035| 47   | sniper       | 0.0026|
| 12   | position     | 0.0064| 30   | ashore       | 0.0034| 48   | pm           | 0.0026|
| 13   | shrapnel     | 0.0059| 31   | heavy        | 0.0034| 49   | casualty     | 0.0026|
| 14   | attack       | 0.0057| 32   | water        | 0.0032| 50   | damage       | 0.0025|
| 15   | bomb         | 0.0055| 33   | landing      | 0.0032| 51   | harbour      | 0.0025|
| 16   | battery      | 0.0053| 34   | gully        | 0.0032| 52   | board        | 0.0024|
| 17   | sea          | 0.0049| 35   | troop        | 0.0032| 53   | aboard       | 0.0023|
| 18   | artillery    | 0.0049| 36   | lemno        | 0.0030| 54   | dead         | 0.0023|

Table 6: Top 99 terms for the Gallipoli topic with their probabilities.
| rank | term     | beta  | rank | term     | beta  | rank | term     | beta  |
|------|----------|-------|------|----------|-------|------|----------|-------|
| 1    | shell    | 0.0153| 19   | bombardment | 0.0052 | 37   | fire     | 0.0035 |
| 2    | trench   | 0.0153| 20   | marched   | 0.0050 | 38   | stunt    | 0.0034 |
| 3    | gun      | 0.0138| 21   | firing    | 0.0048 | 39   | evening  | 0.0033 |
| 4    | line     | 0.0131| 22   | battery   | 0.0044 | 40   | killed   | 0.0033 |
| 5    | fritz    | 0.0096| 23   | enemy     | 0.0042 | 41   | drill    | 0.0032 |
| 6    | german   | 0.0076| 24   | battalion | 0.0042 | 42   | london   | 0.0032 |
| 7    | wounded  | 0.0069| 25   | horse     | 0.0042 | 43   | o'Clock  | 0.0032 |
| 8    | artillery| 0.0069| 26   | machine   | 0.0041 | 44   | shelling | 0.0031 |
| 9    | front    | 0.0067| 27   | aeroplane | 0.0041 | 45   | fatigue  | 0.0031 |
| 10   | billet   | 0.0066| 28   | division  | 0.0040 | 46   | church   | 0.0031 |
| 11   | gas      | 0.0065| 29   | casualty  | 0.0040 | 47   | hut      | 0.0029 |
| 12   | camp     | 0.0065| 30   | attack    | 0.0039 | 48   | el       | 0.0029 |
| 13   | bomb     | 0.0064| 31   | fine      | 0.0039 | 49   | wet      | 0.0028 |
| 14   | mile     | 0.0059| 32   | parade    | 0.0038 | 50   | raining  | 0.0028 |
| 15   | plane    | 0.0058| 33   | position  | 0.0037 | 51   | wood     | 0.0028 |
| 16   | village  | 0.0057| 34   | albert    | 0.0037 | 52   | dug      | 0.0027 |
| 17   | heavy    | 0.0053| 35   | france    | 0.0035 | 53   | moved    | 0.0027 |
| 18   | road     | 0.0053| 36   | taube     | 0.0035 | 54   | tommy    | 0.0027 |

Table 7: Top 54 terms for the In the Trenches (Beginning) topic with their probabilities.

| rank | term     | beta  | rank | term     | beta  | rank | term     | beta  |
|------|----------|-------|------|----------|-------|------|----------|-------|
| 1    | road     | 0.0067| 19   | camel    | 0.0032| 37   | weather  | 0.0025|
| 2    | wrote    | 0.0057| 20   | raid     | 0.0031| 38   | shelling | 0.0023|
| 3    | fritz    | 0.0055| 21   | barrage  | 0.0031| 39   | ridge    | 0.0023|
| 4    | ypre     | 0.0053| 22   | boulogne | 0.0031| 40   | station  | 0.0023|
| 5    | gun      | 0.0048| 23   | wounded  | 0.0030| 41   | omer     | 0.0023|
| 6    | fine     | 0.0047| 24   | london   | 0.0030| 42   | lovely   | 0.0023|
| 7    | enemy    | 0.0047| 25   | walked   | 0.0029| 43   | deferred | 0.0023|
| 8    | brigade  | 0.0045| 26   | raining  | 0.0028| 44   | rain     | 0.0022|
| 9    | train    | 0.0045| 27   | letter   | 0.0028| 45   | report   | 0.0022|
| 10   | cold     | 0.0043| 28   | pt       | 0.0027| 46   | moved    | 0.0021|
| 11   | dinner   | 0.0041| 29   | sister   | 0.0027| 47   | stunt    | 0.0021|
| 12   | bomb     | 0.0040| 30   | paris    | 0.0027| 48   | book     | 0.0021|
| 13   | line     | 0.0039| 31   | plane    | 0.0026| 49   | battery  | 0.0021|
| 14   | hut      | 0.0038| 32   | farm     | 0.0026| 50   | dump     | 0.0020|
| 15   | lorry    | 0.0037| 33   | de       | 0.0026| 51   | lunch    | 0.0019|
| 16   | bailleul | 0.0037| 34   | miss     | 0.0026| 52   | battalion| 0.0019|
| 17   | shell    | 0.0034| 35   | machine  | 0.0025| 53   | division | 0.0019|
| 18   | fed      | 0.0033| 36   | messine  | 0.0025| 54   | le       | 0.0019|

Table 8: Top 54 terms for the In the Trenches (Middle) topic with their probabilities.
| rank | term     | beta  | rank | term     | beta  | rank | term     | beta  |
|------|----------|-------|------|----------|-------|------|----------|-------|
| 1    | fritz    | 0.0144| 19   | battery  | 0.0045| 37   | captured | 0.0027|
| 2    | gun      | 0.0128| 20   | amien    | 0.0043| 38   | forward  | 0.0026|
| 3    | line     | 0.0127| 21   | quiet    | 0.0042| 39   | move     | 0.0026|
| 4    | enemy    | 0.0106| 22   | moved    | 0.0042| 40   | american | 0.0025|
| 5    | shell    | 0.0095| 23   | somme    | 0.0040| 41   | wood     | 0.0024|
| 6    | front    | 0.0078| 24   | evening  | 0.0038| 42   | shelled  | 0.0024|
| 7    | plane    | 0.0072| 25   | trench   | 0.0038| 43   | hot      | 0.0024|
| 8    | village  | 0.0068| 26   | stunt    | 0.0036| 44   | advance  | 0.0024|
| 9    | road     | 0.0065| 27   | gas      | 0.0036| 45   | tank     | 0.0024|
| 10   | battalion| 0.0064| 28   | shelling | 0.0034| 46   | dug      | 0.0023|
| 11   | hun      | 0.0059| 29   | machine  | 0.0032| 47   | casualty | 0.0023|
| 12   | prisoner | 0.0059| 30   | viller   | 0.0032| 48   | lorry    | 0.0023|
| 13   | bomb     | 0.0053| 31   | dugout   | 0.0031| 49   | valley   | 0.0023|
| 14   | division | 0.0051| 32   | french   | 0.0030| 50   | aussie   | 0.0023|
| 15   | wounded  | 0.0047| 33   | heavy    | 0.0030| 51   | river    | 0.0022|
| 16   | position | 0.0047| 34   | barrage  | 0.0029| 52   | dump     | 0.0022|
| 17   | fine     | 0.0046| 35   | le       | 0.0028| 53   | night    | 0.0021|
| 18   | attack   | 0.0045| 36   | la       | 0.0027| 54   | kilo     | 0.0021|

Table 9: Top 54 terms for the In the Trenches (End) topic with their probabilities.

| rank | term     | beta  | rank | term     | beta  | rank | term     | beta  |
|------|----------|-------|------|----------|-------|------|----------|-------|
| 1    | cold     | 0.0429| 19   | parcel   | 0.0033| 37   | foggy    | 0.0022|
| 2    | snow     | 0.0260| 20   | fler     | 0.0032| 38   | albert   | 0.0022|
| 3    | mud      | 0.0150| 21   | camel    | 0.0030| 39   | harness  | 0.0022|
| 4    | christmas| 0.0140| 22   | stable   | 0.0029| 40   | thick    | 0.0022|
| 5    | hut      | 0.0075| 23   | rum      | 0.0029| 41   | ribemont | 0.0022|
| 6    | frost    | 0.0073| 24   | ration   | 0.0028| 42   | patient  | 0.0021|
| 7    | frozen   | 0.0071| 25   | freezing | 0.0028| 43   | delville | 0.0020|
| 8    | el       | 0.0070| 26   | miserable| 0.0026| 44   | thaw     | 0.0020|
| 9    | snowing  | 0.0064| 27   | frosty   | 0.0026| 45   | le       | 0.0020|
| 10   | fritz    | 0.0063| 28   | wind     | 0.0026| 46   | amien    | 0.0019|
| 11   | dugout   | 0.0060| 29   | rafa     | 0.0025| 47   | blyghty  | 0.0019|
| 12   | arish    | 0.0055| 30   | desert   | 0.0024| 48   | bazentin | 0.0018|
| 13   | wood     | 0.0053| 31   | taube    | 0.0024| 49   | hun      | 0.0018|
| 14   | ice      | 0.0052| 32   | mametx   | 0.0024| 50   | sleet    | 0.0018|
| 15   | foot     | 0.0051| 33   | walked   | 0.0024| 51   | needle   | 0.0017|
| 16   | blanket  | 0.0038| 34   | fricourt | 0.0024| 52   | ground   | 0.0017|
| 17   | bitterly | 0.0037| 35   | snowed   | 0.0023| 53   | cleaning | 0.0017|
| 18   | muddy    | 0.0033| 36   | dump     | 0.0022| 54   | headquater| 0.0017|

Table 10: Top 54 terms for the White Christmas topic with their probabilities.
| rank | term     | beta | rank | term     | beta | rank | term     | beta |
|------|----------|------|------|----------|------|------|----------|------|
| 1    | train    | 0.0084 | 19   | car      | 0.0035 | 37   | noon     | 0.0023 |
| 2    | boat     | 0.0069 | 20   | met      | 0.0034 | 38   | ashore   | 0.0023 |
| 3    | ship     | 0.0069 | 21   | troop    | 0.0034 | 39   | city     | 0.0023 |
| 4    | fine     | 0.0065 | 22   | person   | 0.0032 | 40   | cold     | 0.0022 |
| 5    | town     | 0.0060 | 23   | walked   | 0.0031 | 41   | picture  | 0.0022 |
| 6    | sea      | 0.0053 | 24   | lunch    | 0.0031 | 42   | le       | 0.0021 |
| 7    | london   | 0.0050 | 25   | bed      | 0.0030 | 43   | passed   | 0.0021 |
| 8    | evening  | 0.0049 | 26   | house    | 0.0029 | 44   | germany  | 0.0021 |
| 9    | home     | 0.0047 | 27   | board    | 0.0029 | 45   | charleroi| 0.0020 |
| 10   | hotel    | 0.0047 | 28   | dance    | 0.0029 | 46   | concert  | 0.0020 |
| 11   | deck     | 0.0046 | 29   | war      | 0.0029 | 47   | snow     | 0.0020 |
| 12   | pm       | 0.0043 | 30   | afternoon| 0.0029 | 48   | class    | 0.0020 |
| 13   | de       | 0.0043 | 31   | street   | 0.0028 | 49   | aboard   | 0.0020 |
| 14   | dinner   | 0.0041 | 32   | girl     | 0.0027 | 50   | armistice| 0.0020 |
| 15   | port     | 0.0041 | 33   | australia| 0.0025 | 51   | hut      | 0.0019 |
| 16   | walk     | 0.0038 | 34   | visited  | 0.0024 | 52   | billet   | 0.0019 |
| 17   | leave    | 0.0037 | 35   | office   | 0.0024 | 53   | lorry    | 0.0019 |
| 18   | paris    | 0.0035 | 36   | aussie   | 0.0024 | 54   | engine   | 0.0019 |

Table 11: Top 99 terms for the After the Armistice topic with their probabilities.

| rank | term     | beta | rank | term     | beta | rank | term     | beta |
|------|----------|------|------|----------|------|------|----------|------|
| 1    | home     | 0.0300 | 19   | sit      | 0.0044 | 37   | train    | 0.0028 |
| 2    | meet     | 0.0160 | 20   | elli     | 0.0042 | 38   | time     | 0.0026 |
| 3    | boat     | 0.0126 | 21   | tram     | 0.0041 | 39   | card     | 0.0026 |
| 4    | pm       | 0.0104 | 22   | dick     | 0.0040 | 40   | write    | 0.0025 |
| 5    | tea      | 0.0098 | 23   | miss     | 0.0040 | 41   | arrive   | 0.0025 |
| 6    | play     | 0.0092 | 24   | tickle   | 0.0039 | 42   | read     | 0.0022 |
| 7    | ring     | 0.0086 | 25   | wrote    | 0.0036 | 43   | spend    | 0.0022 |
| 8    | catch    | 0.0082 | 26   | dine     | 0.0035 | 44   | night    | 0.0021 |
| 9    | bed      | 0.0070 | 27   | drive    | 0.0035 | 45   | chat     | 0.0021 |
| 10   | manly    | 0.0063 | 28   | roy      | 0.0034 | 46   | dinner   | 0.0021 |
| 11   | mum      | 0.0058 | 29   | day      | 0.0033 | 47   | visit    | 0.0020 |
| 12   | dad      | 0.0054 | 30   | piano    | 0.0033 | 48   | sleep    | 0.0020 |
| 13   | walk     | 0.0052 | 31   | middle   | 0.0032 | 49   | cut      | 0.0019 |
| 14   | paddock  | 0.0050 | 32   | talk     | 0.0031 | 50   | lopped   | 0.0019 |
| 15   | town     | 0.0049 | 33   | otto     | 0.0030 | 51   | meeting  | 0.0019 |
| 16   | music    | 0.0046 | 34   | swim     | 0.0029 | 52   | dave     | 0.0019 |
| 17   | garden   | 0.0045 | 35   | rain     | 0.0029 | 53   | girl     | 0.0018 |
| 18   | george   | 0.0045 | 36   | stay     | 0.0029 | 54   | wharf    | 0.0018 |

Table 12: Top 54 terms for the Home Again topic with their probabilities.