Exploration of World War I Letters Through Text Mining

Ada Mainetti **Kevin Warnakulasuriya** **Harry Zhao**
Amsterdam University College

Abstract

The first World War was among the most significant events of the 20th century. An important form of communication between soldiers and their loved ones were written letters, many of which have since been preserved and digitized. We aim to analyze such letters using text mining techniques to understand wartime experiences.

1. Introduction

Handwritten letters have long been an object of interest in historical studies as a primary source for researchers to learn about the past through first-hand accounts. With the emergence of digital archives, many have been made publicly available on various platforms. As Yeung et. al. (2011) state, this opens up the opportunity to use computational methods to analyze the perception of the past through unprecedented large amounts of data , which could back the traditional manual work of investigating and analyzing historical sources. This paper aims to explore some of the databases that contain collections of letters sent by and delivered to soldiers during the First World War (WWI). These preserved letters offer a great opportunity to learn about how both soldiers and their loved ones back home perceived the war and how they went about their daily lives. We apply a variety of natural language processing techniques and models in order to answer our main research question: *How can computational methods be used to learn about the lived experiences and collective memories of WWI soldiers from transcribed letters?*

To address this research question, we make use of three common frameworks used in the discipline of natural language processing: descriptive statistics, topic modelling, and sentiment analysis. Even though these computational tools have gradually garnered more attention in the field of digital humanities, such methodology remains quite novel in the analysis of older, handwritten texts.

2. Data

Our data consisted of merging various datasets:
1: The 33 English letters subset of a dataset hosted on Kaggle, originally sourced from the British National Archives (Sylvain, 2019), written by various soldiers in either English or French (Sylvain, 2019). The letters contained metadata about each letter (author, date and place of writing, original source).
2: 367 WWI letters scraped and transcribed from the "My Dearest" project (Hargreaves, 2023). The correspondences from this database all concern one particular soldier, David Henry Taylor, and thus only capture the war from the perspective of David's family. There were five main categories of letters, based on David's circumstances or location at the time of writing: London (1915-16): 9 letters; France and Belgium (1917): 77 letters; Missing (1917): 13 letters; Prisoner of War (1917-18): 253 letters; Sheerness (1919): 15 letters.

Using the Beautiful Soup library, we implemented a Python program which reads in the letters from the website and formats them into another JSON file to facilitate our subsequent analysis.

2.1 Preprocessing

Having obtained our raw data, we first extracted only the English letters from the Kaggle dataset. Afterwards, we applied a basic pipeline involving the NLTK tokeniser and part-of-speech tagger, as well as the WordNet lemmatiser. During this step, we also decided to remove the stopwords specified by the NLTK library, in addition to taking out the 2% most frequent words from the Kaggle and "My Dearest" datasets respectively. The majority of the removed lemmas consisted mainly of non-informative function words ("go", "say", "quite"), as well as basic words pertaining to time ("1917", "day", "year"); letter-writing, and war. The names of consistently recurring individuals, like "David" or his sister "Ethel", were also omitted (Hargreaves, 2023).

3. Methodology

3.1 Descriptive Statistics

As a preliminary survey of our dataset, we computed a number of statistical measures of our preprocessed data. Firstly, we calculated the distribution of text sizes (total tokens), vocabulary sizes (unique tokens), and type-token ratio (TTR) over the entire dataset. Secondly, we extracted the 30 most and least frequent words

and bigrams from the remaining corpus after preprocessing, not only of the dataset as a whole but also for letters grouped by year.

We fit a scikit-learn TF-IDF (term frequency—inverse document frequency) vectoriser and extracted the words with highest TF-IDF scores in documents per year. This quantity is higher for terms which occur frequently but only in certain sections of a corpus, which makes it useful for identifying words that are characteristic of a particular set of documents.

3.2 Topic Modelling

Topic Modelling was performed using gensim's LDA model. Additional preprocessing was performed to meet the needs of the Topic Modelling. This included removing common words, adding bigrams and trigrams using gensim's bigram and trigram models. We tried to only retain nouns and adjectives, but the model's performance declined, so this was not ultimately implemented. The optimal topic number searched between 3 to 15 topics, with increments of 2.

3.3 Sentiment Analysis

We used 2 different pretrained sentiment analysis models. One of the models (BERT Multi-classification Sentiment Analysis) employs BERT whereas the other employs DistilBERT (DistilBERT-Base-Uncased for Sentiment Analysis). The BERT based model is trained on the 'GoEmotions' dataset and is able to predict 27 emotional categories as well as a neutral category. The DistilBERT model is trained on the 'Stanford Sentiment Treebank v2' and is able to predict either a positive or negative sentiment.

4. Results

4.1 Descriptive Statistics

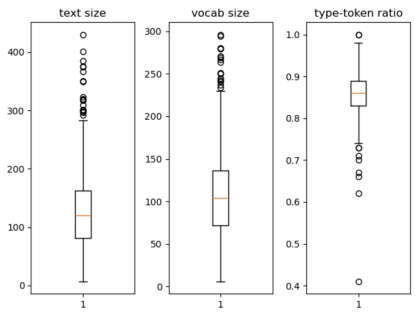

Figure 1: *Box plots illustrating the quartile distribution of text sizes, vocabulary sizes, and the type-token ratios in the preprocessed dataset.*

As Figure 1 shows, the median text size and vocabulary size are both around 100 words each. The type-token ratio is rather high, with a mean value of 0.86, which is likely due to our removal of highly frequent tokens. Nonetheless, this still suggests that soldiers used a diverse choice of words in their letters, which could indicate a wide variety of topics being discussed. Examining the shortest letters documented reveal that quite a number of them have been censored or delayed, with an example preprocessed text being "21 chance card 21 chance censor delay".

The most and least common words and bigrams reveal certain patterns in the text that can be qualitatively analysed. Besides non-specific words such as "since" or "always", many common words relate to healthcare and provisions ("red cross", "receive", "1 lb") or feelings of gratitude ("thank goodness", "thank indeed"). On the contrary, infrequent items generally fall under the category of war-related jargon ("non-fighting", "transfer regiment", "platoon company") and people or events on the front line ("Somme", "corporal J.", "B.E.F France", "might drop").

When performing the same analysis with letters divided by the year of writing, not much new information could be gained for 1914, due to a lack of data, as well as 1917-18, for the opposite reason. Since most letters are sent between these years, they are far too broad to label under a single theme. Distinguishing lexical fields do appear for the other years: letters often talk of destruction in 1915 ("fire", "bomb fell", "gun"); training and ceremonies in 1916 ("national anthem", "physical drill"); and demobilisation in 1917 ("demobilise", "false alarm").

Surprisingly, ranking the words based on the TF-IDF score proves less insightful than simply based on word frequency, for which reason they will not be discussed further.

4.2 Topic Modelling

Quantitative

Model performance was evaluated using coherence scores. We aimed to obtain a coherence score of between 0.4 and 0.6.

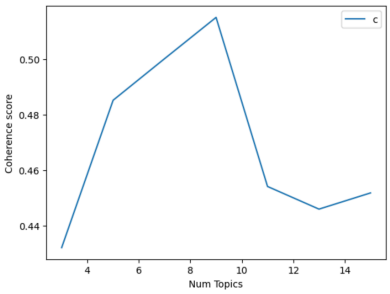

Figure 2: *Coherence scores vs number of topics*

The best model was the 4th one, with a coherence score of 0.515 and 9 topics, and coherence scores dropped off sharply after the number of topics hit the double digits.

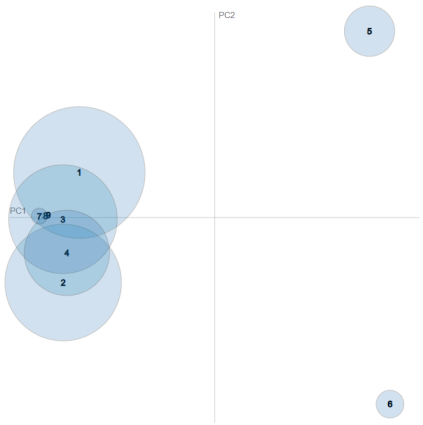

Figure 3: *pyLDAvis visualization of topics*

A few of the topics are extremely small, despite this being the optimal number of topics by coherence score, at least 3 seem small enough to be trivial. This is supported by the topic heatmap (figure 4), in which 4 of the 9 topics are salient. It is possible that selecting 9 topics created the conditions to filter miscellaneous words into smaller, trivial topics, leading to a higher coherence in the main 4 topics.

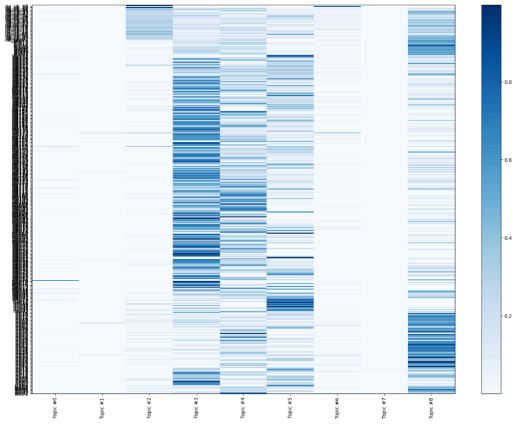

Figure 4: *Heatmap of Topic Clusters*

Qualitative

Here we analyze the topic clusters found. The full list of top 10 words for each topic is found in Appendix A. The top words of the two most common topics (3 and 4) contain words such as cold, monday, weekend, dinner, weather, friday, holiday, garden, life, with little to no mention of wartime affairs. This suggests that despite the war, the average letter writer was mainly writing or inquiring about matters of everyday life. Signs of the war show up in the following two of the 4 salient topics (8, 5), with top words related to war and movement. Among these we have mile, platoon, german, colonel, company, prisoner, wound, fire, water. Though interspersed with other unrelated top words, these topics show evidence of the battles taking place and how a common soldier might document them in correspondence. While, as stated, the bottom 5 topics contain a very small percentage of the total dataset, one stands out for having a very discernible theme. Topic 0 consists primarily of food related words, alongside “machine_gun”, “beg”, “ration”, and “cigarette”. It is likely that a small subset of the data talks very descriptively about food rations, while still showing signs of the war, with mentions of machine guns and rations. This level of detail may be due to their dissatisfaction with the rations, lack of other topics to discuss on quiet days, censorship, or a focus on more “menial” topics to not disturb loved ones with the reality of war.

4.3 Sentiment Analysis

We initially employed the BERT model to predict the sentiment of the letters. Our initial findings were not particularly interesting. As shown in Figure 5, most letters were classified as “neutral”.

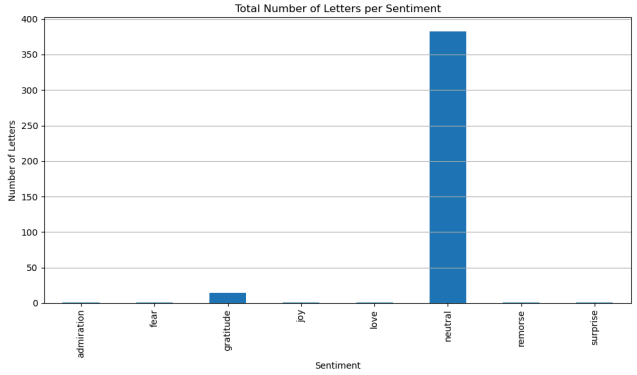

Figure 5*: Initial Sentiment Analysis Results*

However, when looking through the predictions we noticed that shorter letters had non-”neutral” predictions. Breaking up the letters into chunks to be able to analyse smaller sections of the letter and average these chunk level predictions allowed us to get a final letter level prediction. When determining the chunk size, we knew that we wanted to have as few "neutral" predictions as possible, so we ran a repeated analysis to find the chunk size that would do this. We were able to determine a chunk size of 130 words would result in the least number of "neutral" predictions. After running our analysis on the chunk level predictions we got the results outlined in Figure 6.

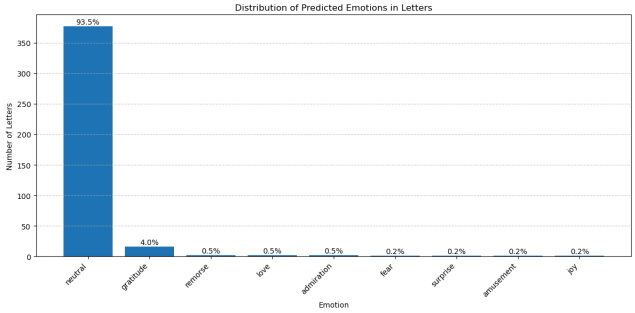

Figure 6: *Sentiment analysis after chunk level predictions.*

Once again we had an overwhelming number of "neutral" predictions, over 90%. The next highest emotion predicted was "gratitude" with 4%. Even though most letters were classified as "neutral", we still found it interesting that there were as many letters classified as "gratitude". Aside from the "neutral" predictions we found the sentiment of the letters to be notable.

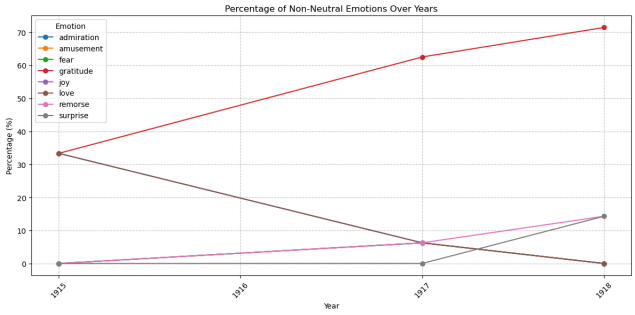

Figure 7: *Yearly letter sentiment analysis*

Figure 7 shows us that the number of letters containing gratitude increase as the war progresses, however, the number of letters containing love fall.

The results from the DistelBERT model better aligned with what we might have expected from our sentiment analysis. Figure 8 shows that the majority of the letters sent had a negative sentiment.

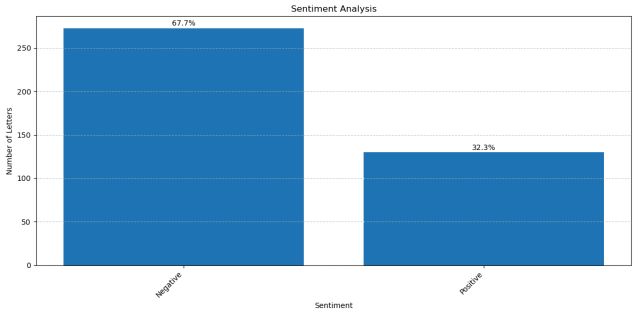

Figure 8*: DistelBERT output*

On a yearly basis (Figure 9) we can see that in the earlier stages of the war the majority of letters sent were positive, but as time progressed the negative letters superseded the number of positive ones, possibly showing increasing popular pessimism as the war dragged on.

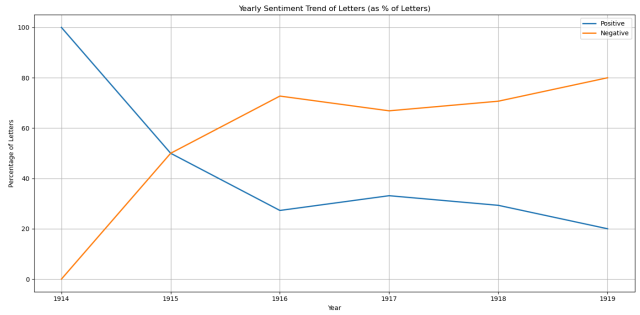

Figure 9: *Positive vs. Negative (Yearly)*

5. Conclusion

Our analysis shows that reasonably cohesive topics can be drawn from the dataset of letters. Though war undoubtedly constitutes one of the major topics, the largest ones pertain not to war itself, but to everyday, mundane life. Analyzing aspects such as letter lengths and type to token ratio also provided insight into aspects such as censors. It is also possible to some extent to track sentiments developing during the course of the war, though most letters seemed neutral, both gratitude and negative sentiments increased as time passed, showing increasing pessimism but also possibly appreciation during hard times.

6. Limitations and Reflections

The primary limitation we faced was insufficient data. Our data (and associated meta-data) proved insufficient to delve into significant in-depth analysis such as comparing topic models or sentiments not just overall, but also by year, author or author role. This limitation also impacted our ability to properly model "collective memories", as our data was primarily sourced from a singular family, and only reflected their own particular narrative and experiences. We suggest that applying a similar method to larger, more diverse datasets could yield better results about the perception of an event such as WWI. Similarly, studies have been done on topic evolution over time (Blei & Lafferty, 2006) (Wang & McCallum, 2006), but our dataset did not provide a sufficient enough distribution of data, which hindered our ability to paint a timeline of the war's evolution and to expect interesting results from year-based or other meta-data based topic modelling.

References

AssemblyAI. (2025). assemblyai/distilbert-base-uncased-sst2 · Hugging Face. Huggingface.co. https://huggingface.co/assemblyai/distilbert-base-uncased-sst2

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models (pp. 113–120). https://doi.org/10.1145/1143844.1143859

Ching-man Au Yeung and Adam Jatowt. 2011. Studying how the past is remembered: towards computational history through large scale text mining. Association for Computing Machinery, New York, NY, USA, 1231–1240. https://doi.org/10.1145/2063576.2063755

Lande, O. (2025). *Om Lande | bert-multiclassification-sentiment-analysis | Kaggle*. Kaggle.com. https://www.kaggle.com/models/omlande/bert-multiclassification-sentiment-analysis

Van Nuenen, T. (2022, March 30). Analyzing Reddit communities with Python — Part 5: topic modeling. *Medium*. https://tomvannuenen.medium.com/analyzing-reddit-communities-with-python-part-5-topic-modeling-a5b0d119add

Wang, X., & McCallum, A. (2006). Topics over time. https://doi.org/10.1145/1150402.1150450

Appendix A

List of top 10 words per topic found, along with tentative topic naming

Listed in decreasing order of salience.

Topic 3 (Mundane Life: Activities and Weather): cold, monday, balham, till, bazaar, late, guess, business, weekend, mrs., dinner, return, weather, mother, pleased, wait, friday, mind, reach, new

Topic 4 (Mundane Life: Object Oriented): large, use, side, holiday, however, door, window, although, mine, sit, list, garden, big, fact, three, life, able, pass, enclose, photo

Topic 8 (Warfront Organization): mile, platoon, pretty, along, road, order, german, whole, billet, fire, water, always, number, colonel, something, company, except, mess, however, sleep

Topic 5 (Injuries and Relationships): receive, july, prisoner, card, doctor, wound, son, child, 10, esther, best, show, sept, love, buy, dec, perhaps, picture, ear, address

Topic 2 (More Warfront Organization): section, let, lot, regiment, receive, officer, army, since, back, little, fine, chap, platoon, sincerely, put, bad, office, use, troop, shall

Topic 6 (Movement at War): guard, show, top, sand, give, charge, bag, high, 1st, alright, hun, week, thing, tent, aeroplane, best, try, lord_mayor, photo, manner

Topic 0 (Food and Rations): soap, milk, scenery, 1, tin, oat, biscuit, beef, machine_gun, ration, paul, beg, pudding, lb, jam, meat, ham, linn, bacon, cigarette

Topic 1 (Leisure at the Front): author, pair_stocking, lesson, jump, to-day, campbell, band, sing, remark, jan, burst, etc., alter, burn, decide, slowly, hundred, quarter, 0, lieutenant

Topic 7 (Travel): ticket, delay, pitch, £40, division, £20, residence, thankful, 3., nation, repeat, arthur, clapham, relate, christ, wish, calmly, pal, district, 25th