

Creating a Historical Sentiment Lexicon Using Johnson's Dictionary and HPC



Millie Cain
University of Leeds
ml22mc@leeds.ac.uk

Project Overview

My project has been to build a historical sentiment lexicon using Johnson's 18th century dictionary using Google's BERT model.

This model allowed the the 50,000 headwords, definitions and examples within the dictionary to calculate a sentiment score for each word. This is based on how positively or negatively each word is defined, as opposed to how the language would be perceived today.

This lexicon can then be applied to historical texts such as newspaper reviews and articles using HPC. While there have been some small-scale efforts to build lexica specific to datasets, this project represents a major step forward in making sentiment analysis historically rigorous.

Data & Methods

My key data source was the 18c dictionary, which I was luckily able to download completely into an Excel file, where I formed 3 columns (Word, definition and sentiment).

I cleaned this data by removing numerical symbols from definitions and excluding non-sentimental words, such as 'and'. For this, I used a Hugging Face BERT model, which I could use to categorise words to be removed, such as 'plant names' and 'conjunction words', before I did a final manual sense check.

Using LLMs proved to be a challenge as results were often incorrect/incomplete, so I preferred to use trained computing and manual methods. This process removed approximately 12,000 words. I kept a couple of non-sentimental words to help balance the scale for later training.

Data & Methods

Following the data cleansing, I categorised the remaining 38,000 words and grouped similar definitions or suffixes, ie: to abandon, abandon, abandoned, abandoning, all being scored the same, which saved lots of time.

A second round of training followed, which a specific code built for dilBERT with the support of my supervisors. This included multiple rounds of batch testing to fine-tune the model in preparation for the whole dictionary sequence.

The data had to be entirely exported into a CSV file and uploaded locally & into an open repository; this was my first experience using GitHub and shared file sourcing, as well as the final run on the HPC.

Finally, I tested further on 1755 texts, to a 92% accuracy.

Results

My results created a fully historically sound sentimental lexicon of the first English dictionary, which can now be further used on accessible texts, which I tested on to check the accuracy and compared with modern sentiment lexicons:

```
lexicon_testing.py
22 ):
26     if normalized_word and pd.notna(sentiment):
29
30     text = ""
31     Fair Venus so often was miss'd from the skies,
32     And Bacchus as frequently absent likewise,
33     That the synod began to enquire out the reason,
34     Suspecting the culprits were plotting of treason.
35     At length it was found they had open'd a ball,
36     At a place by the mortals call'd Bachelors' Hall;
37     Where Venus disclos'd ev'ry fun she could think of,
38     And Bacchus made nectar for mortals to drink of.
39     Jove highly displeased at such riotous doings
40     Sent time to reduce the whole building to ruins.
41     But time was so slack with his traces and dashes
42     That Jove in a passion consumed it to ashes
43     ""
```

```

Pondera'tion.,The act of weighing.,1.0,,1.054662,0.05466198921203613
Rapi'dly.,Swiftly; with quick motion.,2.0,,1.8882948,0.11170518398284912
DELI'RIUM.,Alienation of mind; dotage.,-2.0,,,-2.6930034,0.6930034160614014
To Re'lish.,1. To have a pleasing taste. 2. To give pleasure. 3. To have a flavour.,2.5,,2.597399,0.09739899635314941
Admi'ttable.,The person or thing which may be admitted.,-1.0,,1.6749948,2.6749948263168335
Adu'lterate.,1. Tainted with the guilt of adultery. 2. Corrupted with some foreign mixture.,-3.0,,,-2.8761523,0.1238477
Pu'zzle.,Embarrassment; perplexity.,-1.5,,,-2.5047796,1.004779577255249
COLE.,A general name for all sorts of,-2.75,,0.797383,3.5473830103874207
Ponti'fically.,In a pontifical manner.,1.75,,1.76358,0.013579964637756348
Rese'rve.,1. Store kept untouched. 2. Something kept for exigence. 3. Something concealed in the mind. 4. Exception; p
A'bdicative.,That which causes or implies an abdication.,-2.0,,,-2.546694,0.5466940402984619
Ra'pper.,One who strikes.,-2.0,,,-2.270299,0.27029895782470703
Pontificate.,Papacy; popedom.,2.0,,1.0296798,0.9703202247619629
Wrench.,"1. A violent pull or twist. 2. A sprain. 3. Wrenches, in Chaucer, signifies means, sleights, subtilties;",-3.
Acco'r'dant.,Willing; in a good humour.,2.75,,2.5692496,0.18075037002563477
To Pore.,To look with great intenseness and care; to examine with great attention.,1.5,,2.428592,0.9285919666290283
ACCO'UNT.,"1. A computation of debts or expences; a register of facts relating to money. 2. The state or result of a c
RE'LISH.,1. Taste; the effect of any thing on the palate; 2. Taste; small quantity just perceptible. 3. Liking; deligh
Go'ldhammer.,A kind of bird.,1.0,,1.0480369,0.048036932945251465
To Poke.,To feel in the dark; to search any thing with a long instrument.,1.25,,0.05168887,1.1983111314475536
Pu'nishableness.,The quality of deserving or admitting punishment.,-2.75,,,-1.501282,1.248718023300171
Po'nent.,Western.,1.0,,1.0641696,0.06416964530944824
Po'lish.,1. Artificial gloss; brightness given by attrition. 2. Elegance of manners.,2.5,,2.5942943,0.0942943096160888
Poly'theist.,One that holds plurality of gods.,1.5,,1.7440027,0.24400269985198975
To Mu'tilate.,To deprive of some essential part.,-2.5,,,-2.7362106,0.23621058464050293
To Po'lish.,To answer to the act of polishing; to receive a gloss.,2.5,,2.4843936,0.015606403350830078
To REFU'SE.,1. To deny what is solicited or required. 2. To reject; to dismiss without a grant.,-2.0,,,-2.6906152,0.690
Ponk.,A nocturnal spirit; a hag.,-2.0,,,-1.9613936,0.03860640525817871
To Recrui't.,1. To repair any thing wasted by new supplies. 2. To supply an army with new men.,1.5,,2.0976198,0.597619
Admira'tion.,1. Wonder; the act of admiring or wondering.,3.0,,2.537248,0.462752103805542
To ACCO'MMODATE.,"1. To supply with conveniencies of any kind. to adapt, to fit, to make consistent with.",2.5,,2.4766
Re'fuse.,Unworthy of reception; left when the rest is taken.,-2.0,,,-2.5082362,0.5082361698150635
Proleta'rian.,Mean; wretched; vile; vulgar.,-3.0,,,-2.878554,0.12144589424133301
```

Discussion

Overall, my approach was a lot of trial and error, mainly due to my lack of background knowledge. For this project, I used Python, HPC, and Anaconda for the first time.

I spent the majority of my time learning the methodology of coding and working with the HPC. I felt more intuitive using and working with the lexicon data, as it suits my degree and usual research interests.

I learnt lots from a language perspective about the details and specificity of language perception through time, and of course so much about computational methods, digital research and environments, but had so much wonderful support from the team.

Moving forwards, this project could be expanded to compare against current lexicons and on critical theory texts.

A moment to share my appreciation for my supervisors, Dr Emily Middleton, Dr Tianyi Pan, and the whole RSE team at Leeds for all of their patience, help and humour throughout the process!