

Initial Exploratory Sentiment Analysis of the Reading Experience Database

Robin Davis

2012-04-27

LIS590TX

The data

- UK Open University project
- Collects all reading experiences of British subjects 1450-1945
- <http://www.open.ac.uk/Arts/reading/>
- 29,000 records
- 125 contributors
- Data kindly provided by Dr. E.G.C. King

Reading Experience:

Evidence:

Charlotte Bronte to James Taylor, 1 February 1851: 'Have you yet read Miss Martineau's and Mr Atkinson's new work "Letters on the Nature and Development of Man?" ... It is the first exposition of avowed Atheism and Materialism I have ever read ...'

Century: 1850-1899

Date: Between 1 Jan 1851 and 1 Feb 1851

Country: England

Time: n/a

Place: city: Haworth county: Yorkshire

Type of Experience (Reader):	silent	aloud	unknown
	solitary	in company	
	unknown		
	single	serial	unknown
Type of Experience (Listener):	solitary	in company	
	unknown		
	single	serial	unknown

Reader: Charlotte Bronte

Age: Adult (18-100+)

Gender: Female

Date of Birth: 21 Apr 1816

Socio-economic group: Professional / academic / merchant / farmer

Occupation: Writer

Religion: Church of England

Country of origin: England

Country of experience: England

Text Being Read:

Author: Harriet and H. G. Martineau and Atkinson

Title: Letters on the Laws of Man's Nature and Development

Genre: Other religious, Social Science

Form of Text: Print: Book

Publication: 1851

Provenance: unknown

Source Information:

Record ID: 4408

Dates of evidences

1900-1945	353
1850-1899	365
1800-1849	1138
1700-1799	357
1600-1699	19
1500-1599	9

Initial objectives

- Ascertain general sentiment level of a given text set, beginning with simple Bag of Words method
- Compare MPQA and ANEW lexicons as sentiment analysis tools

Tools: Python, Natural Language Tool Kit (NLTK), Excel

Lexicons

- MPQA Subjectivity Lexicon, aka OpinionFinder
 - Multi-Perspective Question Answering
 - <http://www.cs.pitt.edu/mpqa/>
- ANEW
 - Affective Norms for English Words
 - <http://csea.phhp.ufl.edu/media/anewmessage.html>

1. The first column contains a list of words that are present in the ANEW corpus but not in the OF corpus. These words are listed in a single column.

**Words in ANEW
but not OF**

2. The second column contains a list of words that are present in both the ANEW and OF corpora. These words are listed in a single column.

**Words in both
ANEW and OF**

3. The third column contains a list of words that are present in the OF corpus but not in the ANEW corpus. These words are listed in a single column.

**Words in OF
but not ANEW**

Preparation of lexicons

- Stem with PorterStemmer (NLTK)
- Dedupe stems
- Split into positive and negative word lists
 - For ANEW, split valence mean values into pos and neg

buoyant	posit
calm	posit
candid	posit
candor	posit
capabil	posit
capabl	posit

tortur	neg
torturd	neg
torturli	neg
totalitarian	neg
touchi	neg
tough	neg

Preparation of RED data

- Port deduped evidences from database into CSV
- Give each evidence a unique number
- Remove messy data like “”, [sic], , [italics]
- Split CSV into .txt files with unique number as filename
- Training set: 203 texts
- Full set: 14,073 texts

13547,"'While he read little but the Bible and religious periodicals, his son
13548,"""[William and Dorothy Wordsworth] probably read [the Decameron] toge
13549,"I liked my solitude, even tho? I had no one to say so to - & in spite o
13550,"I read half the 6th book of Antoninus today ? so I can?t say, after all
13551,"I finished the Endymion today. I do not admire it as a fine poem; but
13552,"'The question of conscience once arose when mother was reading """"Jess
13553,"Finished the Choepori, & began the Eumenides. Read more than 500 line
13554,"Witness statement in trial for theft: Mary Flint: """"...in consequenc
13555,"Witness statement in trial for burglary: James Gideon: """"On the 29th
13556,"'The propaganda of Robert Owen alone did not convert printer Thomas Fro
13557,"""""'Within the last month I have read Tristram Shandy, Brydone's Sicil
13558,"""""'Towards the end of his life, W[ordsworth] recalled that during his
13559,"""""'W[ordsworth] read the copy [of John Foxe, Acts and Monuments of Ma
13560,"'Garratt escaped [from factory life] to an evening course in English li
13561,"'As a seaman in the mid-1870s, Ben Tillett had not yet been exposed to
13562,"'[Mary Smith] found emancipation in Shakespeare, Dryden, Goldsmith and
13563,"'like the great man [Carlyle] himself, [Mary Smith] studied Fichte, Sch
13564,"H. J. Jackson on readers' responses in annotations to Samuel Johnson's
13565,"[Marginalia]: various annotations including text marks and numbers thro
13566,"H. J. Jackson notes """"'extra illustration"""" by Philip Gosse of his g
13567,"H. J. Jackson discusses John Horseman's annotations to, and insertions
13568,"'...an article of his in the Daily News on 21 November, blaming Liberal
13569,"'Fine writing and realism were what John Masefield was after in prose.
13570,"'Marjory Todd read [the books of Hesba Stretton, Mrs O.F. Walton and An
13571,"'I went through that extraordinary work of Lord Monboddo on the """"Ori
13572,"'Larpent listened while her husband and stepson read aloud to her from
13573,"'Along with her old school books [Maud Montgomery] read whatever she co

Steps to evaluate each evidence

For file in texts/:

For word in file:

Remove adverbs (words ending in -ly)

Stem word with PorterStemmer

If previous word is a negation word:

If word is in ANEW/MPQA-posList:

Add word after negation word to NegList

If word is in ANEW/MPQA-negList:

Add word after negation word to PosList

If previous word is not a negation word:

If word is in ANEW/MPQA-posList:

Add word to PosList

If word is in ANEW/MPQA-negList:

Add word to NegList

(Continued)

If $\text{len}(\text{PosList}) > \text{len}(\text{NegList})$:

 Add file to TotalPosList

If $\text{len}(\text{NegList}) > \text{len}(\text{PosList})$:

 Add file to TotalNegList

Else:

 Add file to TotalNeutList

10023	N	0	N	0	neut	0
10024	N	0	N	0	neut	0
10025	N	1	neg	2	N	5
10026	pos	6	N	0	N	1
10027	N	0	N	0	neut	0
10028	N	0	N	0	neut	0
10031	pos	1	N	0	N	2
10035	N	3	neg	4	N	1
10036	pos	1	N	0	N	0
10037	pos	2	N	0	N	1
10038	pos	2	N	1	N	0
10039	pos	3	N	0	N	0
10040	pos	2	N	0	N	2

Evaluation of test set

- First manually assigned each evidence a positive, negative, or neutral tag
- After running through evidences with MPQA and ANEW, compared to gold standard as a whole and as pos / neg / neut

Simple evaluation results

- ANEW:
 - negative = $3/24 = 12.5\%$
 - neutral = $30/75 = 40\%$
 - positive = $94/104 = 90.3\%$
 - total: 62.6% correct
- MPQA
 - negative $10/24 = 41.7\%$
 - neutral $28/75 = 36.3\%$
 - positive $85/104 = 80.8\%$
 - overall = 60.6% correct

203	N	0 N	0 neut	0 neg	
251	N	0 N	0 neut	0 neut	1
334	pos	1 N	0 N	0 neut	
432	N	3 N	3 neut	3 neut	1
470	N	0 N	0 neut	0 neut	1
585	pos	3 N	0 N	1 neut	
641	N	0 N	0 neut	0 neut	1
738	N	0 N	0 neut	0 neut	1
865	pos	15 N	6 N	7 pos	1
973	pos	7 N	4 N	6 pos	1
1220	pos	2 N	1 N	0 neut	
1239	N	1 neg	3 N	1 pos	
1430	pos	6 N	2 N	2 pos	1
1536	N	0 N	0 neut	2 pos	
1607	pos	4 N	1 N	1 pos	1
1923	pos	1 N	0 N	1 neut	
1961	N	1 neg	5 N	2 neut	
2109	N	3 neg	4 N	4 neg	1
2792	pos	4 N	0 N	2 pos	1
2918	N	0 neg	1 N	4 neg	1
3292	pos	8 N	6 N	7 neut	
3453	pos	2 N	0 N	0 pos	1
3464	pos	4 N	0 N	2 pos	1
3645	N	2 neg	3 N	2 neut	
3739	pos	6 N	4 N	4 pos	1
3870	pos	2 N	0 N	0 pos	1
4062	pos	8 N	3 N	5 pos	1
4112	pos	9 N	1 N	6 pos	1

Sentiment for full set

ANEW:

neutral: 4,676

negative: 983

positive: 8,160

MPQA:

neutral: 4,147

negative: 2,655

positive: 7,017

Words:

negative: 7,158

positive: 28,087

Words:

negative: 26,715

positive: 45,212

ANEW common positives

accept admir **admir** adult advantag adventur affect air aliv angel answer art astonish autumn bath **beauti bed** bird birthday bless blue **boy** bride bright brother
capabl chanc **charm** cheer **child** christma **church** citi cloth doud comedi comfort confid cook cottag coupl **curiou delight** desir devot **dinner** dog dream dress earth
easi eat educ eleg employ engag **enjoy** event **excel** excit exercis face fame famili famou fascin **father** field flower food free freedom **friend** game garden
gentl gift **girl** god gold **good** grate green happi health **heart** heaven holiday **home** honest hope **hous** human idea imagin
impress improv innoc inspir **interest** intim joy justic **kind** king knowledg lake leader **learn** leisur **letter** liberti **life**
live love magic mail **man** master memori memi mighti **mind** money moral **mother** mountan music **natur** nice nurs **opinion** palac
paradis parti passion peac **peopl** perfect **person** **pleasur** polit **power** **present** pretti pride priest profit **progress** proud qualiti **queen** red
refresh **respect** rich river romant safe saint satisfi save scholar secur sex **sleep** social song **spirit** spring star strong success sun surpris swift talent tast tender **thought**
thrill **travel** treasur treat tree trust **truth** victori virtu vision voyag water white **wife** wise **wit** woman **world** **young** youth

Many Eyes

MPQA common positives

abl abov accept accord acknowledg **admir** adventur affect agre **amus** appreci astonish attent attract **back beauti** bless bright care celebr
charm classic clear clever close comfort conclus consist **content continu** convict convinc correct court curiou **deal dear delight** deserv desir devot
divin dream easi educ effect eleg engag **enjoy entertain** equal event evid **excel excit express** extraordinari fair faith famou **fanci** fascin fashion favour
fine fit fond free **friend gener** geniu german gift glad **good** grace **great** greatest happi hardi **heart** heaven
hero histor **hope human** humour imagin import **impress** improv inclin **inform** inspir instruct intellig **interest** joy justic **kind larg** law
lead **learn** liberti **light live love** lover marvel **master memori merit mind** minist modern moment moral move **natur** nice nobl object
oblig **open** opportun **origin** paradis passion peac perfect **person play** poetic **polit** popular **power** practic **prais** prefer prepar press **pretti**
principl product **progress** promis proper provid purpos quit readi **real reason** recommend regard **remark respect** respons **rest** revolut rich romant rosi
satisfact season **select sens** sensibl sentiment simpl sound **spirit** stand star **state** strike strong sublim success suggest superior sweet swift talent **tast**
thought time togeth **treasur treat truth understand** univers valu valuabl virtu white worth youth

ANEW common negatives

absurd accid ach admir afraid agoni **alon** anger angri anguish annoy **anxiou** **armi** arrog bastard beast bee beggar bench blind bloodi **board** bomb bore broken
brutal burden burial **burn** cell cliff coars **cold** confus contempt **content** control corner corrupt crash **crime** crimin crisi crude cruel curtain **cut** damag
danger dark dead **death** defeat delay demon depress demelict desert despair despis destroy destruct detest devil dirti disappoint discourag
disgust distress disturb divorc **dread** dreari drown **evil** excus execut failur **fall** fals fat fatigu **fault** **fear** feebl fever fight **fire** flood foul fraud funer
glass good gossip grief guilti gun **habit** **hate** hatr headach hell heroin hide hit hooker horror hurt ignor **ill** imagin immor indiffer **inferior** injuri insan insol insult iron
irrit kick killer **lie** **lost** mad mind miseri **mistak** morbid **murder** neglect nervou **nonsens** **odd** offend outrag overwhelm **pain** pamphlet
piti plain poison poverti **prison** punish quarrel rage sat **regret** reject resent reserv revolt **ridicul** rotten rough rude **sad** scandal scom **seat** selfish
sever shadow shame **sick** **sin** sissi slave sleep slow **solemn** spider squar startl stiff storm stress **stupid** suicid suspici **terribl** thorn thought tobacco tomb
tragedi trash **troubl** unhappi upset useless **vaniti** violent **war** wast weari wick wound

MPQA common negatives

absurd abus adam affect afraid air antiqu anxiou argu argument asham attack **bad** battl beg black blind blood bore break burn busi censur cheap
close cold comedi comic **commonplac** complain concern confess confus contempt content controversi cri crime **critic** cross cut danger dark
dead **death** declin deep defect depress devil die differ difficult difficulti disappoint disgust dislik disturb dog **doubt** drama dread dull error evil
extrem fall fall fals **fanci** fault fear fell fiction fight **fine** fool forc forget grave **hard** harri hate horror humbl hurt idl ignor **ill** imposs indulg inferior
juvenil keen kill lack laugh **lectur** lie limit **littl** long lose lost low mad margin mere **mind** miser **miss** mistak
murder mysteri **nation** nonsens **object** obscur odd offend **opinion** opposit outsid **pain** pass peculiar piti plot **poor** prejudic press pretend pride
prison problem protest punch puzzl **quit** radic rank refus regret retir ridicul sad satir savag secret seriou **sermon** **sever** sharp shock sick sin
solemn sorri sorrow sound spite spot stern storm **strang** strike struck struggl stupid **subject** suffer superior suspect sympathi terribl throw tire tragedi
tri trial **troubl** understand utter vain vaniti vice violent vulgar **war** wast weak wearl whatev wick **wild** wors worst wound **wrong** wrought

Next steps

- Refine lexicons (or choose/create new one)
 - What counts as positive sentiment? Frequent readings? Belief in its truth? Enjoyment?
 - What counts as negative sentiment? What if a good book makes you feel sad?
 - Include British spellings (e.g., disfavour)
- Refine system
 - Classifier evaluation
 - Better distinguish negative sentiment
 - Phrase-level analysis and relevance identification
 - Set focus to areas in text that are around the title of the work, if multiple works are mentioned
 - Consider intensity in addition to polarity
 - adverbial intensification
 - i found ----miss---- sara hutchinson read coleridg christabel to johnni wordsworth she wa ----tire---- i read the greater part of it he wa ----**excess**---- ++++interest++++ espec with the first part

Next next step

- Perform analysis on subsets of data
 - E.g.: what is the historical opinion of Shakespeare's works?
 - For 292 Shakespeare records:
 - 164 positive (56.1%)
 - 41 negative (14.0%)
 - 87 neutral (29.8%)