**Sentiment Analysis Notes**

**Packages**

<https://www.tidytextmining.com/sentiment.html>

Primary package is tidytext, which provides access to several sentiment lexicons. Three general-purpose lexicons are:

* AFINN from Finn Årup Nielsen,
* bing from Bing Liu and collaborators, and
* nrc from Saif Mohammad and Peter Turney.

Focusses on sentiment analysis at the word level. nrc has the advantage that it can operate in multiple languages (including all languages in our study).

Of these libraries, nrc feels like a strong contender due to the fact that it has dictionaries for over 100 languages (<http://www.saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm>). Though it should be mentioned that these dictionaries have been created using Google translate, so are likely to be only semi-reliable.

<https://www.red-gate.com/simple-talk/sql/bi/text-mining-and-sentiment-analysis-with-r/>

<https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html>?

Primary package is syuzhet. Seems to be able to process sentiment and emotional valence at the sentence level. Can also utilise all of the lexicons above. The default program uses Stanford’s coreNLP. Syuzhet works with any language that uses a latin character set. It is likely to perform best on fiction as it was trained using contemporary novels.

<https://uc-r.github.io/sentiment_analysis>

Basic sentiment analysis uses tidytext with same libraries as above. Larger sentiments analysed using sentimentr, which utilizes the Stanford CoreNLP.

<https://cran.r-project.org/web/packages/SentimentAnalysis/vignettes/SentimentAnalysis.html>

Lab on how to use SentimentAnalysis package (<https://cran.r-project.org/web/packages/SentimentAnalysis/SentimentAnalysis.pdf>). Package comes with 5 in-built dictionaries:

* Harvard-IV, a dictionary with a list of positive and negative words according to the psychological Harvard-IV dictionary as used in the General Inquirer software. This is a general-purpose dictionary developed by the Harvard University. (<https://rdrr.io/cran/SentimentAnalysis/man/DictionaryGI.html>; <http://www.wjh.harvard.edu/~inquirer/homecat.htm>; <http://www.textanalysis.info/pages/category-systems/general-category-systems/harvard-iv-dictionary.php>)
* Henry’s finance-specific dictionary (<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=933100>)
* Loughran-McDonald financial dictionary (<https://www.onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.2010.01625.x>)
* Uncertainty words from Loughran-McDonald
* Polarity words from data object key.pol which is by the package qdap (<https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html>; <http://trinker.github.io/qdapDictionaries/key.pol.html>; <https://dl.acm.org/doi/10.5555/1597148.1597269>)

Of these, I plan to look into Harvard-IV and qdap in more detail.

The package also offers the option to build custom dictionaries using labelled datasets. This uses LASSO regularization as a statistical approach to select relevant terms based on an exogenous response variable.

Overall Opinion

From what I can tell, the packages which allow for the most comprehensive contextual analysis (e.g. syuzhet, sentimentr), tend to rely on the Stanford CoreNLP toolkit. The paper describing this toolkit is here (<https://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf>). The sentiment analysis uses, “a compositional model over trees using deep learning (Socher et al., 2013). Nodes of a binarized tree of each sentence, including, in particular, the root node of each sentence, are given a senti-ment score.” The paper explaining this method is here (<https://nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf>), which I still need to have a read through (though one of the authors in Andrew Ng, so I imagine it’s good!).

A key issue I’ve noticed so far however is that the Standford CoreNLP focuses on English, Chinese and Arabic, but doesn’t offer sentiment analysis for German, Italian or Spanish. A summary it’s capacity in different languages is here (<https://stanfordnlp.github.io/CoreNLP/human-languages.html>). So, it may be worth opting for a simpler method which performs equally well across all 4 languages in our study. The nrc package outlined here (<https://www.tidytextmining.com/sentiment.html>), includes sentiment values for 13,901 words in each of the following languages: Arabic, Basque, Bengali, Catalan, Chinese\_simplified, Chinese\_traditional, Danish, Dutch, English, Esperanto, Finnish, French, German, Greek, Gujarati, Hebrew, Hindi, Irish, Italian, Japanese, Latin, Marathi, Persian, Portuguese, Romanian, Russian, Somali, Spanish, Sudanese, Swahili, Swedish, Tamil, Telugu, Thai, Turkish, Ukranian, Urdu, Vietnamese, Welsh, Yiddish, Zulu (<https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html>). In total it has dictionaries for over 100 languages (<http://www.saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm>). Though it should be mentioned that these dictionaries have been created using Google translate, so are likely to be only semi-reliable.

**Papers**

<https://www.sciencedirect.com/science/article/pii/S2214367X20301691#b0085> *A framework with efficient extraction and analysis of Twitter data for evaluating public opinions on transportation services*

Uses AFINN. Word by word sentiment.

<https://journals.sagepub.com/doi/epub/10.1177/1461444813480466> *Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens’ political preferences with an application to Italy and France*

Uses supervised machine learning.

<https://ieeexplore.ieee.org/document/8700266>

Cutting edge method, published 1/10/2020. It uses sentiment diffusion (the sentiment of both the original tweet and those of anyone who retweeted it) to predict a tweet’s overall sentiment. It also employs a range of other concepts, including structural virality, sentiment reversal (when the sentiment of the original tweet is different from its retweets) and diffusion patterns to refine its predictions. This method has shown a 5-8% increase in accuracy when predicting on a labelled dataset compared to the next best state-of-the-art model.

Conceptually and methodologically it’s far too complex for our timeline.

Papers which utilise NRC lexicon

<https://www.scilit.net/article/b2e23632050a883c5042b7f42563b55f> *Automatic Assessment of Performance of Hospitals using Subjective Opinions for Sentiment Classification*

<https://link.springer.com/article/10.1007/s12652-020-01726-4> *A review on recognizing depression in social networks: challenges and opportunities*

<https://thesai.org/Publications/ViewPaper?Volume=10&Issue=10&Code=IJACSA&SerialNo=47> *Lexicon-based Bot-aware Public Emotion Mining and Sentiment Analysis of the Nigerian 2019 Presidential Election on Twitter*

<https://www.mdpi.com/2076-3417/9/10/2035> *Using Social Media to Identify Consumers' Sentiments towards Attributes of Health Insurance during Enrollment Season*

<https://www.sciencedirect.com/science/article/pii/S1877050919318873> *Sentiment Analysis to Assess the Community's Enthusiasm Towards the Development Chatbot Using an Appraisal Theory*

<https://www.semanticscholar.org/paper/Lexicon-Based-Sentiment-Comparison-of-iPhone-and-Aoudi-Malik/d9ac143f8eaf9d843372753e2b172cc2a08190f2> *Lexicon Based Sentiment Comparison of iPhone and Android Tweets During the Iran-Iraq Earthquake*

<https://ieeexplore.ieee.org/document/7226110> *Exploiting Dependency Relations for Sentence Level Sentiment Classification using SVM*

<https://link.springer.com/chapter/10.1007/978-3-319-18117-2_7> *Feature Selection for Twitter Sentiment Analysis: An Experimental Study*