sentibank: A Unified Resource of Sentiment Lexicons and Dictionaries

Nick Oh

Abstract

Sentiment analysis is critical across computational social science domains, but faces challenges in interpretability. Rule-based methods relying on expert lexicons enable transparency, yet applying them is hindered by resource fragmentation and lack of validation. This paper introduces sentibank, a large-scale unified database consolidating 15 original sentiment dictionaries and 43 preprocessed dictionaries, spanning 7 genres and 6 domains.

Motivation

1. Scattered Landscape and Formats

Various lexicons are *scattered across various sources*, such as GitHub repositories, appendices of publications, supplementary materials, and author/institutional websites. This fragmented distribution poses a significant challenge for researchers who seek to leverage sentiment analysis effectively. Furthermore, they are *distributed in diverse file formats* (e.g. .rdf, .xml, .py, .txt), necessitating the tedious process of exporting and importing data into a format compatible with the researcher's workflow.

2. Unrefined Data for Research

Numerous lexicons, including those that undergo peer review, frequently encounter challenges such as the *presence of duplicates accompanied by conflicting labels*. A substantial portion of existing dictionaries in sentibank (60%) required removal of duplicates, function words, and lexicons lacking substantive sentiment content.

Preprocessing

The core contribution of sentibank lies in *standardisation of sentiment dictionaries*, enabling rapid research utilisation. This saves researchers significant effort, allowing direct application of the sentiment data rather than expending time on formatting and quality checking diverse lexicons individually.

While some dictionaries required minimal changes (e.g. deduplication and minor formatting), ANEW, DAL, NoVAD, SentiWordNet, and WordNet-Affect underwent extensive preprocessing to mitigate ambiguities and inconsistencies arising from fuzzy or vector representations. The key objective was harmonising representations into well-defined, exclusive schemes.

Avaliable Dictionaries

Dictionary	Genre	Domain
AFINN	Social Media	General
Aigents+	Social Media	Cryptocurrency
ANEW	General (Standard English)	Psychology
Dictionary of Affect in Language (DAL)	Vernacular (Day-to-Day Expression)	General
Discrete Emotions Dictionary (DED)	News	Political Science
General Inquirer	General (Standard English)	Psychology, Political Science
Henry	Corporate Communication (Earnings Press Releases)	Finance
MASTER	Regulatory Filings (10-K)	Finance
Norms of Valence, Arousal and Dominance (NoVAD)	Vernacular (Day-to-Day Expression)	General, Psychology
OpinionLexicon	Reviews	Consumer Products
SenticNet	General (Standard & Non-Standard English)	General
SentiWordNet	General (Standard English)	General
SO-CAL	General (Standard & Non-standard English)	General
VADER	Social Media	General
WordNet-Affect	General (Standard English)	Psychology