2.2 Sentiment Classification

Edwin Simpson

Department of Computer Science,

University of Bristol, UK.

Different Ways to Model Sentiment

- Positive vs. negative sentiment (attitude):

👍 → 👎

★ → ★★★★★

Different Ways to Model Sentiment

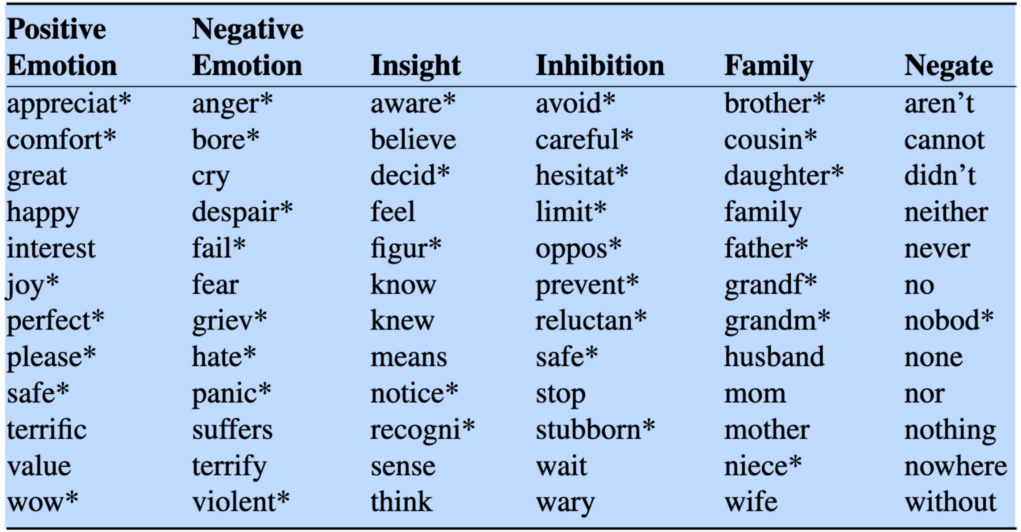
Affective states	Descriptors
Emotion	😡🤗😧😨
Mood	😃😀🙁☹️
Interpersonal stance	Supportive, mocking, …
Personality traits	Nervous, reckless, jealous, …

Modifying Naïve Bayes for Sentiment Analysis

- **Binary NB:** clipping word counts to 1 with each document;
- **Negation**:
  - *The movie was enjoyable* vs. *The movie was* ***not*** *enjoyable;*
  - First, detect negations using a regular expression to find *not, no, never, n't;*
  - Then, replace these phrases with a new token, *NOT_enjoyable;*
  - Can't handle more complex phrases like *no plot twists or great scenes*

Lexicons

- Some words indicate a particular sentiment or affective state.
- Our training set may omit many sentiment or affective state words we'll see in testing.
- A lexicon is a hand-crafted list of words with a specific sentiment or connotation



Positive Emotion	Negative Emotion	Insight	Inhibition	Family	Negate
appreciat*	anger*	aware*	avoid*	brother*	aren't
comfort*	bore*	believe	careful*	cousin*	cannot
great	cry	decid*	hesitat*	daughter*	didn't
happy	despair*	feel	limit*	family	neither
interest	fail*	figur*	oppos*	father*	never
joy*	fear	know	prevent*	grandf*	no
perfect*	griev*	knew	reluctan*	grandm*	nobod*
please*	hate*	means	safe*	husband	none
safe*	panic*	notice*	stop	mom	nor
terrific	suffers	recogni*	stubborn*	mother	nothing
value	terrify	sense	wait	niece*	nowhere
wow*	violent*	think	wary	wife	without

Figure 20.6 Samples from 5 of the 73 lexical categories in LIWC (Pennebaker et al., 2007). The * means the previous letters are a word prefix and all words with that prefix are included in the category.

Figures from Chapter 20, Speech and Language Processing, 3rd edition draft, Jurafsky & Martin (2020).

Using Lexicons with Naïve Bayes

- Augment the bag of words with new features, *IN_POS_LEXICON* and *IN_NEG_LEXICON*
- These features count occurrences of words in the positive and negative lexicons
- Add *n* occurrences of *IN_POS_LEXICON* to the bag of words, where *n* is the number of words in the positive lexicon.
- By including lexicon features, we use prior knowledge to supplement a lack of training data.

Additional Classification Features

Feature types	Example Task	Example Feature
Pre-defined phrases (in text body or subject line)	Spam filter	*Online pharmaceuticals*
Bigrams (two-word sequences)	Sentiment analysis	*Well written*
Character n-grams (two-character sequences)	Language identification	*Nya, cz, th*
Average word length > 4	Authorship attribution	See Brinegar (1963)

Brinegar, C. S. Mark Twain and the Quintus Curtius Snodgrass letters: A statistical test of authorship. *Journal of the American statistical Association* 58.301 (1963): 85-96.

Summary

- Tasks like sentiment analysis benefit from task-specific adaptations and features
- Detecting negation is important for classifier performance
- Lexicons provide additional information to make up for training set deficiencies
- Consider n-gram features as well as single tokens.