CS 424P/ LINGUIST 287
Extracting Social Meaning and Sentiment

Dan Jurafsky and Chris Potts

Lecture 10: Wrap-up

General Take-Away Messages

- **Most sentiment papers use really dumb features**.
  - The field is young. There's lots of room for creativity!
- Consider a **broad spectrum of social meaning**: such papers are likely to make a valuable contribution.
- Sentiment is expressed w/words prosody, gesture…
    - **The best models will probably synthesize lots of modalities**.
- **Look at your data**, don't just run a classifier and report F-score.

Better Sentiment Lexicons

- **Build an effective sentiment lexicon** and it will get used *a lot*. Witness LIWC, SentiWordNet, Inquirer.
- **Words aren't enough**
  - Sentiment units don't necessarily align with word units; lexicons should contain phrases, constructions, etc.
- **Strings aren't enough.**
  - Neither are (string, category) pairs.
  - We might need the full richness of the context of utterance.
- **Sentiment is domain-dependent**

Better Sentiment Lexicons

- **Seed-sets:** not necessarily bad, but the field needs to move past hand-built seed-sets, because the nature of the seed-sets is often the most important factor in performance.
- **Graph-based approaches** are common, but there is **no consensus that they are the best.**

Sentiment Classification

- Sentiment seems to be harder than other classification problems; expect lower scores.
- Sentiment might be blended and continuous, making classification somewhat inappropriate.
- Feature presence seems to be better than feature frequency. (The opposite is generally true for other tasks.)
- Feature selection is important. It is worth finding a measure that maximizes information gain but also considers the strength of the evidence.
- Sentiment features are highly topic/context dependent.

Prosody

- Pitch and energy are easy
  - consider range or any other measure of variance in addition to max and min
  - remove outliers; can do this via quartiles, or SD, or dropping 10%, etc
  - A maxim from speech recognition: biggest reduction in error rate always comes from cleaning up training data.
- Duration/rate of speech
  - rate of speech is easier to compute.
  - pause length, number of pauses, or burstiness (variance of duration features)
- Jitter/Shimmer seem powerful, but may require cleaner speech

Disfluencies

- A big untapped feature set
- Consider the non-"uh/um" ones especially, uh/um don’t seem to have much social meaning
- Types:
  - restarts, repetitions, you know/i mean/like, word fragments

Flirtation, Dialogue

- Use Pennebaker or other lexicons only as a first pass to help you build a real features set;
- Don't rely on the Pennebaker names of classes; look at your data (i.e. at how the lexical cluster behaves).
  - "mean" is very rarely about meaning, nor "like" about preference.
- Dialogue features more important than lexical ones.
  - build dialogue features from the linguistics literature, or from other labeled dialogue corpora
- Think about speaker differences (personality, etc). NLP doesn't consider individual speakers nearly as much as it should.

Emotion

- You need to look at other languages if you are trying to make a general claim.
- But this is hard.
- There's very little work on lexical features in emotion detection.
  - Most people in this field seem to be from prosodic backgrounds.
  - So there’s an opening here for exciting research in lexical/dialogue features.

Deception

- Deception features are weak.
- Deception features are varied: lexical, intonational, physical.
- Commonly used domain-independent features:
  - 1st person pronouns, positive and negative words, exclusive terms, ``motion verbs''.
- Domain-dependent features are valuable, and the nature of the domain can affect the above profoundly.
- No single feature identifies deception. Perhaps clusters of them can (in context).

Medical

- It's not good enough to build a classifier that can find a disease or state if it can’t distinguish it from different diseases/states (i.e. a drunk detector is useless if it can't tell drunk from stressed)
- Lesson from Agatha Christie detector: many features are genre-specific, and even if the feature isn't, the threshold might be.
- “I" and "we" seem very useful but it's even better if you look at the data and convince yourself they worked for the right reason.

Politics

- Bias != subjectivity.
- Bias is in the eye of the beholder/classifier. Consider the observer's role.
- Some people are deceptive about their partisanship, but our algorithms have a chance to see through the subterfuge.
- Classification models benefit from social, relational information.
- Work in this area can inform journalism and public-policy debates.

Some ideas for future projects

- Detect power structures from conversation (Enron?)
- Paranoia detection, compulsiveness detection, autism spectrum, Aspergers.
  - perhaps build a corpus from online forums, autobiography by people with known diagnosis, etc
- Determine from chat how close friends people are.
- Tone-checking in email
- Build an effective sentiment lexicon. Perhaps begin with small high-precision seeds (LIWC, GI, etc), and then extend/adapt it in particular domains
- Mine interesting tasks from Scherer's 5 kinds of affective speech.
- Use disfluencies to build better drunk-detector

Some ideas for future projects

- feature presence versus feature counts
  - a function of data size?
  - a function of genre?
  - synonymy?
  - "how helpful is this review". does it correlate with large counts?
- using non-text structure in your data (e.g., graph structure) to cluster to do unsupervised lexicon generation
- humor detection, humor generation
- better features to describe the sentiment context of words
- compositionality of vector models of semantics
- politeness
  - vocatives for addressees in email
- web search with sentiment
- controversial topic detection
  - blocked wikipedia
  - wikitrust addon