CIS 530: Computational Linguistics

MONDAYS AND WEDNESDAYS 1:30-3PM
3401 WALNUT, ROOM 401B
COMPUTATIONAL-LINGUISTICS-CLASS.ORG

PROFESSOR CALLISON-BURCH
Professor Callison-Burch (not Professor Burch)

Bachelors from Stanford
PhD from University of Edinburgh
6 years at Johns Hopkins University
Joined Penn faculty in 2013

I have been working in the field of NLP since 2000. In 2017, I was the general chair of the 55th meeting of the ACL.
Course Staff

Bhavna Saluja  
Gaurav Kumar  
Harry Zhang  
Liam Dugan  
Sihao Chen  
Tatiana Tsygankova  
Tyler Larkworthy
<table>
<thead>
<tr>
<th>Urdu</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>S → NP(^1) VP(^2)</td>
<td>NP(^1) VP(^2)</td>
</tr>
<tr>
<td>VP → PP(^1) VP(^2)</td>
<td>VP(^2) PP(^1)</td>
</tr>
<tr>
<td>VP → V(^1) AUX(^2)</td>
<td>AUX(^2) V(^1)</td>
</tr>
<tr>
<td>PP → NP(^1) P(^2)</td>
<td>P(^2) NP(^1)</td>
</tr>
<tr>
<td>NP → hamd ansary</td>
<td>Hamid Ansari</td>
</tr>
<tr>
<td>NP → nājib sdr</td>
<td>Vice President</td>
</tr>
<tr>
<td>V → namzd</td>
<td>nominated</td>
</tr>
<tr>
<td>P → kylje</td>
<td>for</td>
</tr>
<tr>
<td>AUX → taa</td>
<td>was</td>
</tr>
</tbody>
</table>
Hamid Ansari

was nominated for Vice President
Paraphrases

Differing **textual** expressions of the same meaning:

- cup $\leftrightarrow$ mug
- the king’s speech $\leftrightarrow$ His Majesty’s address
- $X_1$ devours $X_2$ $\leftrightarrow$ $X_2$ is eaten by $X_1$
- one JJ instance of NP $\leftrightarrow$ a JJ case of NP
riots were sparked by twelve of the cartoons that are offensive to the Islamic prophet.
Word Sense

microbe, virus, bacterium, germ, parasite
insect, beetle, pest, mosquito, fly
bother, annoy, pester
bug
microphone, tracker, mic, wire, earpiece, cookie
squealer, snitch, rat, mole
glitch, error, malfunction, fault, failure
## Semantic Relationships

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>twelve</td>
<td>12</td>
<td>equivalence</td>
</tr>
<tr>
<td>cartoons</td>
<td>illustrations</td>
<td>forward entailment</td>
</tr>
<tr>
<td>$\varepsilon$</td>
<td>in Denmark</td>
<td>reverse entailment</td>
</tr>
<tr>
<td>caused</td>
<td>prevented</td>
<td>negation</td>
</tr>
<tr>
<td>Europe</td>
<td>the middle East</td>
<td>alternation</td>
</tr>
</tbody>
</table>
The Gun Violence Database
Information Extraction

Three seconds. On a dashcam video clock, that's the amount of time between the moment when two officers have their guns drawn and the point when Laquan McDonald falls to the ground. The video, released to the public for the first time late Tuesday, is a key piece of evidence in a case that's sparked protests in Chicago and has landed an officer in handcuffs. The 17-year-old McDonald was shot 6 times on that day the video shows in October 2014. Chicago police Officer Jason Van Dyke was charged Tuesday with first-degree murder.

<table>
<thead>
<tr>
<th>Person #1014</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name</strong></td>
<td>Laquan McDonald</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Race</strong></td>
<td></td>
</tr>
</tbody>
</table>

Incident #1053

<table>
<thead>
<tr>
<th>City</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Date</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Shooter</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Victim</strong></td>
<td>McDonald</td>
</tr>
<tr>
<td><strong>Victim Killed</strong></td>
<td></td>
</tr>
</tbody>
</table>
What will you learn?

This will be a survey class in natural language processing

Focus will be programming assignments for hands-on learning

Topics will include things like
- Sentiment analysis
- Vector space semantics
- Machine translation
- Information extraction
Course textbook

Don’t buy this book!

The Authors are releasing free draft chapters of their updated 3rd edition.

https://web.stanford.edu/~jurafsky/slp3/

We will use the draft 3rd edition as our course textbook, along with required reading of research papers.
## Course Grading

- Weekly programming assignments
- Short quizzes on the assigned readings
- Self-designed final project
- No final exam or midterm
- All homework assignments can be done in pairs, except for HW1
- Final project will be teams of ~4-5 people
- 5 free late days for the term (1 minute - 24 hours = 1 day late)
- You cannot drop your lowest scoring homework
Text Classification and Sentiment Analysis

JURAFSKY AND MARTIN CHAPTER 4
Positive or negative movie review?

👎 unbelievably disappointing
👍 Full of zany characters and richly applied satire, and some great plot twists
👍 this is the greatest screwball comedy ever filmed
👎 It was pathetic. The worst part about it was the boxing scenes.
What is the subject of this article?

MEDLINE Article

MeSH Subject Category Hierarchy

- Antagonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology

...
Classify User Attributes Using Their Tweets

- Delighted I kept my Xmas vouchers - Happy Friday to me 😊 #shopping
- Yesterday’s look-my new obsession is this Givenchy fur coat! Wolford sheer turtleneck, Proenza skirt & Givenchy boots
- We’ve already tripled wind energy in America, but there’s more we can do.
- Two giant planets may cruise unseen beyond Pluto - space - June 2014 - New Scientist: newscientist.com/article/dn2571
Lexical Markers for Age

Slide from Svitlana Volkova
Lexical Markers for Political Preferences
Lexical Markers for Gender
Who wrote which Federalist papers?

1787-1788: anonymous essays try to convince New York to ratify U.S Constitution by Jay, Madison, Hamilton.

Authorship of 12 of the letters in dispute

1963: solved by Mosteller and Wallace using Bayesian methods
When a man unprincipled in private life, desperate in his fortune, bold in his temper... despotic in his ordinary demeanor — known to have scoffed in private at the principles of liberty — when such a man is seen to mount the hobby horse of popularity — to join in the cry of danger to liberty — to take every opportunity of embarrassing the government & bringing it under suspicion — to flatter and fall in with all the nonsense of the zealots of the day — It may justly be suspected that his goal is to throw things into confusion that he may ‘ride the storm and direct the whirlwind.’

—Alexander Hamilton, 1792
Text Classification

- Assigning subject categories, topics, or genres
- Spam detection
- Authorship identification
- Age/gender identification
- Language Identification
- Sentiment analysis
- ...
Sentiment Analysis

WHAT IS SENTIMENT ANALYSIS?
Input: "Spiraling away from narrative control as its first three episodes unreel, this series, about a post-apocalyptic future in which nearly everyone is blind, wastes the time of Jason Momoa and Alfre Woodard, among others, on a story that starts from a position of fun, giddy strangeness and drags itself forward at a lugubrious pace."

Output: positive (1) or negative (0)
Google Product Search

HP Officejet 6500A Plus e-All-in-One Color Ink-jet - Fax / copier / printer / scanner
$89 online, $100 nearby ★★★★★ 377 reviews

September 2010 - Printer - HP - Inkjet - Office - Copier - Color - Scanner - Fax - 250 shi

Reviews

Summary - Based on 377 reviews

<table>
<thead>
<tr>
<th>1 star</th>
<th>2</th>
<th>3</th>
<th>4 stars</th>
<th>5 stars</th>
</tr>
</thead>
</table>

What people are saying

ease of use

"This was very easy to setup to four computers."

value

"Appreciate good quality at a fair price."

setup

"Overall pretty easy setup."

customer service

"I DO like honest tech support people."

size

"Pretty Paper weight."

mode

"Photos were fair on the high quality mode."

colors

"Full color prints came out with great quality."
Twitter sentiment versus Gallup Poll of Consumer Confidence

Target Sentiment on Twitter

Type in a word and we'll highlight the good and the bad

"united airlines" Search Save this search

Sentiment analysis for "united airlines"

Sentiment by Percent

Negative (38%)

Positive (32%)

Sentiment by Count

0 5 10 15 20 25 30

Positive (11)

Negative (23)

jljacobson: OMG... Could @United airlines have worse customer service? W8g now 15 minutes on hold 4 questions about a flight 2DAY that need a human
Posted 2 hours ago

12345ciumsy6789: I hate United Airlines Ceiling!!! Fukn impossible to get my conduit in this damn mess! ?
Posted 2 hours ago

EMLandPRGbelgiu: EML/PRG fly with Q8 united airlines and 24seven to an exotic destination. http://t.co/Z9QIoAjF
Posted 2 hours ago

CountAdam: FANTASTIC customer service from United Airlines at XNA today. Is tweet more, but cell phones off now!
Posted 4 hours ago
Sentiment analysis has many other names

- Opinion extraction
- Opinion mining
- Sentiment mining
- Subjectivity analysis
Why sentiment analysis?

- **Movie**: is this review positive or negative?
- **Products**: what do people think about the new iPhone?
- **Public sentiment**: how is consumer confidence? Is despair increasing?
- **Politics**: what do people think about this candidate or issue?
- **Prediction**: predict election outcomes or market trends from sentiment
Scherer Typology of Affective States

**Emotion**: brief organically synchronized ... evaluation of a major event
  - angry, sad, joyful, fearful, ashamed, proud, elated

**Mood**: diffuse non-caused low-intensity long-duration change in subjective feeling
  - cheerful, gloomy, irritable, listless, depressed, buoyant

**Interpersonal stances**: affective stance toward another person in a specific interaction
  - friendly, flirtatious, distant, cold, warm, supportive, contemptuous

**Attitudes**: enduring, affectively colored beliefs, dispositions towards objects or persons
  - liking, loving, hating, valuing, desiring

**Personality traits**: stable personality dispositions and typical behavior tendencies
  - nervous, anxious, reckless, morose, hostile, jealous

Sentiment analysis is the detection of **attitudes**

“enduring, affectively colored beliefs, dispositions towards objects or persons”

1. **Holder (source)** of attitude
2. **Target (aspect)** of attitude
3. **Type** of attitude

From a set of **types**

- *Like, love, hate, value, desire,* etc.

Or (more commonly) simple weighted **polarity**:

- *positive, negative, neutral,* together with *strength*

From a **Text** containing the attitude

- Sentence or entire document
Simplest task:
  ◦ Is the attitude of this text positive or negative?

More complex:
  ◦ Rank the attitude of this text from 1 to 5

Advanced:
  ◦ Detect the target, source, or complex attitude types
Sentiment Analysis

A BASELINE ALGORITHM
Sentiment Classification in Movie Reviews

Polarity detection:
- Is an IMDB movie review positive or negative?

Data: *Polarity Data 2.0:*
- [http://www.cs.cornell.edu/people/pabo/movie-review-data](http://www.cs.cornell.edu/people/pabo/movie-review-data)


when _star wars_ came out some twenty years ago, the image of traveling throughout the stars has become a commonplace image. [...] when han solo goes light speed, the stars change to bright lines, going towards the viewer in lines that converge at an invisible point. Cool.

_october sky_ offers a much simpler image—that of a single white dot, traveling horizontally across the night sky. [...]

“snake eyes” is the most aggravating kind of movie: the kind that shows so much potential then becomes unbelievably disappointing.

it’s not just because this is a brian depalma film, and since he’s a great director and one who’s films are always greeted with at least some fanfare. and it’s not even because this was a film starring nicolas cage and since he gives a brauvara performance, this film is hardly worth his talents.
Baseline Algorithm (adapted from Pang and Lee)

1. **Tokenization**
2. **Feature Extraction**
3. **Classification using different classifiers**
   - Naïve Bayes
   - MaxEnt
   - SVM
   - CRF
   - Neural net
Sentiment Tokenization Issues

Deal with HTML and XML markup

Twitter mark-up (names, hash tags)

Capitalization (preserve for words in all caps)

Phone numbers, dates

Emoticons

Useful code:
- Christopher Potts sentiment tokenizer
- Brendan O’Connor twitter tokenizer

Potts emoticons

[<>]?  # optional hat/brow
[::=8]  # eyes
[\-o\*\']?  # optional nose
[\()\)\[\dDpP/:\}\{@\|\}\|]  # mouth
[\()\)\[\dDpP/:\}\{@\|\}\|]  # reverse orientation
[\-o\*\']?  # mouth
[::=8]  # optional nose
[<>]?  # eyes
[\()\)\[\dDpP/:\}\{@\|\}\|]  # optional hat/brow
Extracting Features for Sentiment Classification

How to handle negation
- I *didn’t* like this movie
  vs
- I really like this movie

Which words to use?
- Only adjectives
- All words
  - All words turns out to work better, at least on this data
Add NOT_ to every word between negation and following punctuation:

didn’t like this movie, but I

didn’t NOT_like NOT_this NOT_movie but I
Text Classification with Naïve Bayes

THE TASK OF TEXT CLASSIFICATION
Text Classification: definition

**Input:**
- a document $d$
- a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$

**Output:** a predicted class $c \in C$
Naïve Bayes

Intuition

Simple (”naïve”) classification method based on Bayes rule

Relies on very simple representation of document called a *bag of words*
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!
The bag of words representation

\[
\gamma(\ldots) = c
\]

<table>
<thead>
<tr>
<th>Word</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>seen</td>
<td>2</td>
</tr>
<tr>
<td>sweet</td>
<td>1</td>
</tr>
<tr>
<td>whimsical</td>
<td>1</td>
</tr>
<tr>
<td>recommend</td>
<td>1</td>
</tr>
<tr>
<td>happy</td>
<td>1</td>
</tr>
<tr>
<td>\ldots</td>
<td>\ldots</td>
</tr>
</tbody>
</table>
Bayes’ Rule Applied to Documents and Classes

For a document $d$ and a class $c$

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$
Naïve Bayes Classifier

\[ c_{MAP} = \arg\max_{c \in C} P(c | d) \]

\[ = \arg\max_{c \in C} \frac{P(d | c)P(c)}{P(d)} \]

\[ = \arg\max_{c \in C} P(d | c)P(c) \]

MAP is “maximum a posteriori” = most likely class

Bayes Rule

Dropping the denominator
Naïve Bayes Classifier

\[ c_{MAP} = \arg \max_{c \in C} P(d | c)P(c) = \arg \max_{c \in C} P(x_1, x_2, \ldots, x_n | c)P(c) \]

Document d represented as features x1..xn
Multinomial Naïve Bayes

Independence Assumptions

\[ P(x_1, x_2, \ldots, x_n \mid c) \]

**Bag of Words assumption:** Assume position doesn’t matter

**Conditional Independence:** Assume the feature probabilities \( P(x_i \mid c_j) \) are independent given the class \( c \).

\[ P(x_1, \ldots, x_n \mid c) = P(x_1 \mid c) \cdot P(x_2 \mid c) \cdot P(x_3 \mid c) \cdot \ldots \cdot P(x_n \mid c) \]
Multinomial Naïve Bayes Classifier

\[ c_{\text{MAP}} = \arg \max_{c \in C} P(x_1, x_2, \ldots, x_n | c) P(c) \]

\[ c_{\text{NB}} = \arg \max_{c \in C} P(c_j) \prod_{x \in X} P(x | c) \]
Problems: What makes reviews hard to classify?

Subtlety

Perfume review in *Perfumes: the Guide*:

“If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut.”

Dorothy Parker on Katherine Hepburn

“She runs the gamut of emotions from A to B”
Problems: What makes reviews hard to classify?

Thwarted Expectations and Ordering Effects

◦ “This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can’t hold up.”

◦ Well as usual Keanu Reeves is nothing special, but surprisingly, the very talented Laurence Fishbourne is not so good either, I was surprised.
Text Classification and Naïve Bayes

PARAMETER ESTIMATION AND SMOOTHING
Learning the Multinomial Naïve Bayes Model

First attempt: maximum likelihood estimates, which simply use the frequencies in the data

\[
\hat{P}(c_j) = \frac{\text{doccount}(C = c_j)}{N_{doc}}
\]

\[
\hat{P}(w_i | c_j) = \frac{\text{count}(w_i, c_j)}{\sum_{w \in V} \text{count}(w, c_j)}
\]
Create mega-document for topic $j$ by concatenating all docs in this topic

- Use frequency of $w$ in mega-document

$$
\hat{P}(w_i \mid c_j) = \frac{\text{count}(w_i, c_j)}{\sum_{w \in V} \text{count}(w, c_j)}
$$

fraction of times word $w_i$ appears among all words in documents of topic $c_j$
Problem with Maximum Likelihood

What if we have seen no training documents with the word $\text{fantastic}$ and classified in the topic $\text{positive (thumbs-up)}$?

\[
\hat{P}(\text{"fantastic" | positive}) = \frac{\text{count("fantastic", positive)}}{\sum_{w \in V} \text{count}(w, \text{positive})} = 0
\]

Zero probabilities cannot be conditioned away, no matter the other evidence!

\[
c_{\text{MAP}} = \arg\max_c \hat{P}(c) \prod_i \hat{P}(x_i | c)
\]
Laplace (add-1) smoothing for Naïve Bayes

\[ \hat{P}(w_i \mid c) = \frac{\text{count}(w_i, c) + 1}{\sum_{w \in V} (\text{count}(w, c) + 1)} \]

\[ = \frac{\text{count}(w_i, c) + 1}{\left( \sum_{w \in V} \text{count}(w, c) \right) + |V|} \]
Multinomial Naïve Bayes: Learning

- From training corpus, extract *Vocabulary*

Calculate $P(c_j)$ terms
- For each $c_j$ in $C$
  
  $\text{docs}_j \leftarrow$ all docs with class $= c_j$

  $P(c_j) \leftarrow \frac{|\text{docs}_j|}{|\text{total # documents}|}$

Calculate $P(w_k \mid c_j)$ terms
- $\text{Text}_j \leftarrow$ single doc containing all $\text{docs}_j$
- For each word $w_k$ in *Vocabulary*
  
  $n_k \leftarrow$ # of occurrences of $w_k$ in $\text{Text}_j$

  $P(w_k \mid c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha |\text{Vocabulary}|}$
Text Classification and Naïve Bayes

Precision, Recall, and the F Measure
The 2-by-2 contingency table

<table>
<thead>
<tr>
<th></th>
<th>correct</th>
<th>not correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>selected</td>
<td>tp</td>
<td>fp</td>
</tr>
<tr>
<td>not selected</td>
<td>fn</td>
<td>tn</td>
</tr>
</tbody>
</table>
Precision and recall

**Precision**: % of selected items that are correct

**Recall**: % of correct items that are selected

<table>
<thead>
<tr>
<th></th>
<th>correct</th>
<th>not correct</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>selected</strong></td>
<td>tp</td>
<td>fp</td>
</tr>
<tr>
<td><strong>not selected</strong></td>
<td>fn</td>
<td>tn</td>
</tr>
</tbody>
</table>

Precision = \( \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \)

Recall = \( \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \)
A combined measure: F

A combined measure that assesses the P/R tradeoff is F measure (weighted harmonic mean):

\[ F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \]

The harmonic mean is a very conservative average

People usually use balanced F1 measure
- i.e., with \( \beta = 1 \) (that is, \( \alpha = \frac{1}{2} \)):

\[ F1 = \frac{2PR}{P+R} \]
Text Classification and Naïve Bayes
Cross-Validation

Break up data into 10 folds
  - (Equal positive and negative inside each fold?)

For each fold
  - Choose the fold as a temporary test set
  - Train on 9 folds, compute performance on the test fold

Report average performance of the 10 runs
Development Test Sets and Cross-validation

Metric: P/R/F1 or Accuracy

Development test set
- avoid overfitting to the unseen test set
- Use dev set to select the “best” model
- **Cross-validation over multiple splits**
  - Handle sampling errors from different datasets
  - Compute pooled dev set performance
  - This way we can use all data for validation
NO CLASS ON MONDAY (MLK HOLIDAY)
FOR NEXT WEDNESDAY:
READ JURAFSKY AND MARTIN
CHAPTERS 2 & 4, AND THUMBS UP?
SENTIMENT CLASSIFICATION USING
MACHINE LEARNING TECHNIQUES
COMPLETE HOMEWORK 1 (ON YOUR
OWN).