Significance in NLP

• You develop a new method for text classification; is it better than what comes before?

• You’re developing a new model; should you include feature X? (when there is a cost to including it)

• You're developing a new model; does feature X reliably predict outcome Y?
Evaluation

• A critical part of development new algorithms and methods and demonstrating that they work
Metrics

- Evaluations presuppose that you have some metric to evaluate the fitness of a model.
  - Text classification: accuracy, precision, recall, F1
  - Phrase-structure parsing: PARSEVAL (bracketing overlap)
  - Dependency parsing: Labeled/unlabeled attachment score
  - Machine translation: BLEU, METEOR
  - Summarization: ROUGE
  - Language model: perplexity
Metrics

• Downstream tasks that use NLP to predict the natural world also have metrics:
  • Predicting presidential approval rates from tweets
  • Predicting the type of job applicants from a job description
  • Conversational agent
Classification

A mapping \( h \) from input data \( x \) (drawn from instance space \( \mathcal{X} \)) to a label (or labels) \( y \) from some enumerable output space \( \mathcal{Y} \)

\[
\mathcal{X} = \text{set of all documents} \\
\mathcal{Y} = \{\text{english, mandarin, greek, ...}\}
\]

\( x = \) a single document \\
\( y = \) ancient greek
# Multiclass confusion matrix

<table>
<thead>
<tr>
<th></th>
<th>Dem</th>
<th>Repub</th>
<th>Indep</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dem</strong></td>
<td>100</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td><strong>Repub</strong></td>
<td>0</td>
<td>104</td>
<td>30</td>
</tr>
<tr>
<td><strong>Indep</strong></td>
<td>30</td>
<td>40</td>
<td>70</td>
</tr>
</tbody>
</table>
# Accuracy

\[
\frac{1}{N} \sum_{i=1}^{N} I[\hat{y}_i = y_i]
\]

\[I[x] = \begin{cases} 
1 & \text{if } x \text{ is true} \\
0 & \text{otherwise}
\end{cases}
\]

<table>
<thead>
<tr>
<th></th>
<th>Dem</th>
<th>Repub</th>
<th>Indep</th>
</tr>
</thead>
<tbody>
<tr>
<td>True ((y))</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dem</td>
<td>100</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>Repub</td>
<td>0</td>
<td>104</td>
<td>30</td>
</tr>
<tr>
<td>Indep</td>
<td>30</td>
<td>40</td>
<td>70</td>
</tr>
</tbody>
</table>
Precision

Precision(Dem) =

\[
\frac{\sum_{i=1}^{N} I(y_i = \hat{y}_i = \text{Dem})}{\sum_{i=1}^{N} I(\hat{y}_i = \text{Dem})}
\]

Precision: proportion of predicted class that are actually that class.
Recall

Recall(Dem) =

\[
\frac{\sum_{i=1}^{N} I(y_i = \hat{y}_i = \text{Dem})}{\sum_{i=1}^{N} I(y_i = \text{Dem})}
\]

*Recall*: proportion of true class that are predicted to be that class.
F score

\[ F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \]
Ablation test

• To test how important individual features are (or components of a model), conduct an ablation test
  
  • Train the full model with all features included, conduct evaluation.
  
  • Remove feature, train reduced model, conduct evaluation.
## Ablation test

<table>
<thead>
<tr>
<th>Feature</th>
<th>Dev.</th>
<th>Test</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Our tagger, all features</strong></td>
<td>88.67</td>
<td>89.37</td>
<td></td>
</tr>
<tr>
<td>independent ablations:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- DISTSIM</td>
<td>87.88</td>
<td>88.31</td>
<td>-1.06</td>
</tr>
<tr>
<td>- TAGDICT</td>
<td>88.28</td>
<td>88.31</td>
<td>-1.06</td>
</tr>
<tr>
<td>- TWORTH</td>
<td>87.51</td>
<td>88.37</td>
<td>-1.00</td>
</tr>
<tr>
<td>- METAPH</td>
<td>88.18</td>
<td>88.95</td>
<td>-0.42</td>
</tr>
<tr>
<td>- NAMES</td>
<td>88.66</td>
<td>89.39</td>
<td>+0.02</td>
</tr>
<tr>
<td><strong>Our tagger, base features</strong></td>
<td>82.72</td>
<td>83.38</td>
<td></td>
</tr>
<tr>
<td><strong>Stanford tagger</strong></td>
<td>85.56</td>
<td>85.85</td>
<td></td>
</tr>
</tbody>
</table>

| Annotator agreement             | 92.2  |       |         |

Table 2: Tagging accuracies on development and test data, including ablation experiments. Features are ordered by importance: test accuracy decrease due to ablation (final column).

Gimpel et al. 2011, “Part-of-Speech Tagging for Twitter”
Significance

- If we observe difference in performance, what’s the cause? Is it because one system is better than another, or is it a function of randomness in the data? If we had tested it on other data, would we get the same result?

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Your work</td>
<td>58%</td>
</tr>
<tr>
<td>Current state of the art</td>
<td>50%</td>
</tr>
</tbody>
</table>
Hypotheses

<table>
<thead>
<tr>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>The average income in two sub-populations is different</td>
</tr>
<tr>
<td>Web design A leads to higher CTR than web design B</td>
</tr>
<tr>
<td>Self-reported location on Twitter is predictive of political preference</td>
</tr>
<tr>
<td>Your system X is better than state-of-the-art system Y</td>
</tr>
</tbody>
</table>
Null hypothesis

- A claim, assumed to be true, that we’d like to test (because we think it’s wrong)

<table>
<thead>
<tr>
<th>hypothesis</th>
<th>$H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>The average income in two sub-populations is different</td>
<td>The incomes are the same</td>
</tr>
<tr>
<td>Web design A leads to higher CTR than web design B</td>
<td>The CTR are the same</td>
</tr>
<tr>
<td>Self-reported location on Twitter is predictive of political preference</td>
<td>Location has no relationship with political preference</td>
</tr>
<tr>
<td>Your system X is better than state-of-the-art system Y</td>
<td>There is no difference in the two systems.</td>
</tr>
</tbody>
</table>
Hypothesis testing

• If the null hypothesis were true, how likely is it that you’d see the data you see?
Hypothesis testing

• Hypothesis testing measures our confidence in what we can say about a null from a sample.
Hypothesis testing

• Current state of the art = 50%; your model = 58%. Both evaluated on the same test set of 1000 data points.

• Null hypothesis = there is no difference, so we would expect your model to get 500 of the 1000 data points right.

• If we make parametric assumptions, we can model this with a Binomial distribution (number of successes in n trials)
Example

Binomial probability distribution for number of correct predictions in n=1000 with p = 0.5
Example

At what point is a sample statistic unusual enough to reject the null hypothesis?
Example

• The form we assume for the null hypothesis lets us quantify that level of surprise.

• We can do this for many parametric forms that allows us to measure $P(X \leq x)$ for some sample of size $n$; for large $n$, we can often make a normal approximation.
Z score

\[ Z = \frac{X - \mu}{\sigma / \sqrt{n}} \]

For Normal distributions, transform into standard normal (mean = 0, standard deviation =1)
Z score

510 correct
= z score 0.63

580 correct
= z score 5.06
Tests

- We will define “unusual” to equal the most extreme areas in the tails
least likely 10%

$< z = -1.65$

$> z = 1.65$
least likely 5%
least likely 1%
Tests

![Normal distribution curve with z-scores]

- 510 correct = z score 0.63
- 580 correct = z score 5.06

Tests
Tests

• Decide on the level of significance $\alpha$. \{0.05, 0.01\}

• Testing is evaluating whether the sample statistic falls in the rejection region defined by $\alpha$
• Two-tailed tests measured whether the observed statistic is different (in either direction)

• One-tailed tests measure difference in a specific direction

• All differ in where the rejection region is located; $\alpha = 0.05$ for all.
p values

A p value is the probability of observing a statistic at least as extreme as the one we did if the null hypothesis were true.

- Two-tailed test: $p\text{-value}(z) = 2 \times P(Z \leq -|z|)$

- Lower-tailed test: $p\text{-value}(z) = P(Z \leq z)$

- Upper-tailed test: $p\text{-value}(z) = 1 - P(Z \leq z)$
Errors

- Type I error: we reject the null hypothesis but we shouldn’t have.
- Type II error: we don’t reject the null, but we should have.
<table>
<thead>
<tr>
<th></th>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>“jobs”</td>
<td>is predictive of presidential approval rating</td>
</tr>
<tr>
<td>2</td>
<td>“job”</td>
<td>is predictive of presidential approval rating</td>
</tr>
<tr>
<td>3</td>
<td>“war”</td>
<td>is predictive of presidential approval rating</td>
</tr>
<tr>
<td>4</td>
<td>“car”</td>
<td>is predictive of presidential approval rating</td>
</tr>
<tr>
<td>5</td>
<td>“the”</td>
<td>is predictive of presidential approval rating</td>
</tr>
<tr>
<td>6</td>
<td>“star”</td>
<td>is predictive of presidential approval rating</td>
</tr>
<tr>
<td>7</td>
<td>“book”</td>
<td>is predictive of presidential approval rating</td>
</tr>
<tr>
<td>8</td>
<td>“still”</td>
<td>is predictive of presidential approval rating</td>
</tr>
<tr>
<td>9</td>
<td>“glass”</td>
<td>is predictive of presidential approval rating</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>“bottle”</td>
<td>is predictive of presidential approval rating</td>
</tr>
</tbody>
</table>
Errors

• For any significance level $\alpha$ and $n$ hypothesis tests, we can expect $\alpha \times n$ type I errors.

• $\alpha=0.01$, $n=1000 = 10$ “significant” results simply by chance
Multiple hypothesis corrections

• Bonferroni correction: for family-wise significance level $\alpha_0$ with $n$ hypothesis tests:

$$a \leftarrow \frac{\alpha_0}{n}$$

• [Very strict; controls the probability of at least one type I error.]

• False discovery rate
Confidence intervals

• Even in the absence of specific test, we want to quantify our uncertainty about any metric.

• Confidence intervals specify a range that is likely to contain the (unobserved) population value from a measurement in a sample.
**Confidence intervals**

Binomial confidence intervals (again using Normal approximation):

- $p =$ rate of success (e.g., for binary classification, the accuracy).
- $n =$ the sample size (e.g., number of data points in test set).
- $z_\alpha =$ the critical value at significance level $\alpha$.

- 95% confidence interval: $\alpha = 0.05$; $z_\alpha = 1.96$
- 99% confidence interval: $\alpha = 0.01$; $z_\alpha = 2.58$

$$p \pm z_\alpha \sqrt{\frac{p(1 - p)}{n}}$$
Issues

• Evaluation performance may not hold across domains (e.g., WSJ → literary texts)

• Covariates may explain performance (MT/parsing, sentences up to length n)

• Multiple metrics may offer competing results
Takeaways

• At a minimum, always evaluate a method on the domain you’re using it on.

• When comparing the performance of models, quantify your uncertainty with significant tests/confidence bounds.

• Use ablation tests to identify the impact that a feature class has on performance.
Ethics

Why does a discussion about ethics need to be a part of NLP?
Conversational Agents
Toxic generation

- Language models like GPT-{1,2,3} trained on toxic data (e.g., banned subreddits like /r/The_Donald or /r/WhiteRights) reproduce that toxicity in both prompted and unprompted generations.

Question Answering

According To Google, Barack Obama Is King Of The United States

Google Answers gets it wrong. Is this a Google Answers Bomb?

Bary Schwartz on November 25, 2014 at 6:04 pm

Google

King of United States

Web  Maps  Images  Shopping  Videos  More  Search tools

About 460,000,000 results (0.72 seconds)

All Hail King Barack Obama, Emperor Of The United States Of America!

All Hail King Barack Obama, Emperor Of The United States ...

www.breitbart.com/.../All-Hail-King-Barack-Obama-Emperor-Of...
Language Modeling
Ethics

• The decisions we make about our methods — training data, algorithm, evaluation — are often tied up with its use and impact in the world.

• NLP is now being used more and more to reason about human behavior.
Ethics

• Bias leading to allocational or representational harms.
• Privacy
• Exclusion
• Dual Use
Bias

- Allocational harms: automated systems allocate resources unfairly to different groups (access to housing, credit, parole).

- Representational harms: automated systems represent one group less favorably than another (including demeaning them or erasing their existence).

Adverse impact

“substantially different rate of selection in hiring, promotion, or other employment decision which works to the disadvantage of members of a race, sex, or ethnic group”

Uniform Guidelines on Employee Selection Procedures
Allocations

- Credit opportunities
- Assess to housing
- Job opportunities (LinkedIn, HR)
- Predictive policing

Not just categorical decisions, but e.g. advertising choices
Representations

- **Pleasant**: caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation.

- **Unpleasant**: abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, bomb, divorce, jail, poverty, ugly, cancer, evil, kill, rotten, vomit.

- Embeddings for African-American first names are closer to “unpleasant” words than European names (Caliskan et al. 2017)
• Sentiment analysis over sentences containing African-American first names are more negative than identical sentences with European names

Kiritchenko and Mohammad (2018), "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems"
• Toxicity detection systems score text with African-American English as more offensive

• Implicit negative perception of AAE → more AAE tweets are removed → users change language practices

Blodgett et al. (2020); Sap et al. (2019), “The risk of racial bias in hate speech detection”
Privacy
Congratulations!

The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.

On September 21, 2009 we awarded the $1M Grand Prize to team “BellKor’s Pragmatic Chaos”. Read about their algorithm, checkout team scores on the Leaderboard, and join the discussions on the Forum.

We applaud all the contributors to this

Privacy

• Large language models (e.g., GPT-3, BERT) can memorize training data, which is recoverable from it.

• Potential violations of confidential data (e.g., GMail messages) and contextual integrity (data being published in a way that violates a user’s expectations of use).

Carlini et al. (2020), “Extracting Training Data from Large Language Models”
Exclusion

- Focus on data from one domain/demographic
- State-of-the-art models perform worse for young (Hovy and Søgaard 2015) and minorities (Blodgett et al. 2016)
Exclusion

Table 3: Proportion of tweets in AAE- and white-aligned corpora classified as non-English by different classifiers. (§4.1)

<table>
<thead>
<tr>
<th></th>
<th>AAE</th>
<th>White-Aligned</th>
</tr>
</thead>
<tbody>
<tr>
<td>langid.py</td>
<td>13.2%</td>
<td>7.6%</td>
</tr>
<tr>
<td>Twitter-1</td>
<td>8.4%</td>
<td>5.9%</td>
</tr>
<tr>
<td>Twitter-2</td>
<td>24.4%</td>
<td>17.6%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parser</th>
<th>AA</th>
<th>Wh.</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>SyntaxNet</td>
<td>64.0 (2.5)</td>
<td>80.4 (2.2)</td>
<td>16.3 (3.4)</td>
</tr>
<tr>
<td>CoreNLP</td>
<td>50.0 (2.7)</td>
<td>71.0 (2.5)</td>
<td>21.0 (3.7)</td>
</tr>
</tbody>
</table>

Blodgett et al. (2016), "Demographic Dialectal Variation in Social Media: A Case Study of African-American English" (EMNLP)
Dual Use

• Authorship attribution (author of *Federalist Papers* vs. author of ransom note vs. author of political dissent)

• Fake review detection vs. fake review generation

• Censorship evasion vs. enabling more robust censorship
ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)

Ethics, Bias, and Fairness. This area includes work that analyzes, detects and mitigates stereotypical bias or offensive wording in language data as well as work discussing ethical concerns about NLP applications.
Activity

8.tests/ParametricTest

• Explore a simple hypothesis test checking whether the accuracy of a trained model for binary classification on your data is meaningfully different from a majority class baseline