CS224C: NLP for CSS

Computational Basics

Diyi Yang
Stanford CS
Announcements

Please sign up for Presentation by this coming Tuesday (Apr 9th)
Without access to social media platform data, we risk being left in the dark

**Significance:**

Social media data are essential for studying human behaviour and understanding potential systemic risks. Social media platforms have, however, begun to remove access to these data. In response, other countries and regions have implemented legislation that compels platforms to provide researchers with data access. In South Africa, we have lagged behind the Global North when it comes to using platform data in our research and, given the recent access restrictions, we risk being left behind. In this Commentary, I call attention to this critical issue and initiate a conversation about access to social media data in South Africa.
Challenge Summary

1. The complexity of the theoretical issues confronting social science

2. The difficulty in obtaining the relevant observational data

3. The difficulty of manipulating large scale social organizationals experimentally

4. The complexity and difficulty in computationally, scientifically and rigorously modeling such problems and data
Computational Social Science in a nutshell

Data → Algorithm → Problem → Theory → Knowledge → Impact
Lecture Overview

- Classification
- Regression
- Clustering
- Big Data + CSS
Computational Framework

1. Formulate the research problem
2. Define population and sample
3. Collect the data
4. Do descriptive data analysis
5. Use appropriate statistical methods to solve the research problem
6. Report the results
Use of Classification or Regression

Two major uses of supervised classification/regression

**Prediction:**
Train a model on a sample of data \((x, y)\) to **predict** for some new data \(x'\)
In zero-shot or few-shot prompting, no training is needed!

**Interpretation or Explanation:**
Train a model on a sample of data \((x, y)\) to **understand the relationship** between \(x\) and \(y\)
Common Methods

Classification  Clustering  Regression
Classification

A **mapping** $h$ from input data $x$ (drawn from instance space $X$) to a label $y$ from some enumerable output space $Y$

- $X = \text{set of all documents}$
- $Y = \{\text{English, Mandarin, Greek, \ldots}\}$
- $x = \text{a single document}$
- $y = \text{ancient Greek}$
Reviews and Ratings

**IMDb Charts**

**IMDb Top 250 Movies**

(IMDb Top 250 as rated by regular IMDb voters.)

**Showing 250 Titles**

<table>
<thead>
<tr>
<th>Rank &amp; Title</th>
<th>IMDb Rating</th>
<th>Your Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The Shawshank Redemption (1994)</td>
<td>★★9.2</td>
<td>★★</td>
</tr>
<tr>
<td>2. The Godfather (1972)</td>
<td>★★9.2</td>
<td>★★</td>
</tr>
<tr>
<td>3. The Dark Knight (2008)</td>
<td>★★9.0</td>
<td>★★</td>
</tr>
<tr>
<td>5. 12 Angry Men (1957)</td>
<td>★★9.0</td>
<td>★★</td>
</tr>
</tbody>
</table>

---

**Reviewed: October 24, 2022**

Lovely little spot to spend some time. Very grateful for the clean, stylish room after travelling.

- Beautiful, spacious room - after 22 hours of travel and a botched flight, I cried with happiness when I arrive.
- Large, comfy bed - I didn't want to get out.
- Friendly, helpful staff - especially the bar staff.
- Accessible, enjoyable bar - open to late with a large selection of drinks.
- Lots of room to sit by the pool. With a spa too. As well as on the foreshore in hammocks. Very laidback, enjoyable environment. I happily spent the day here, relaxing before a friends wedding.
- I loved walking along the foreshore footpath, following the shoreline and walking past the other hotels. As I walked, I did plenty of people watching.

- The beach wasn't an inviting swim, though a beautiful backdrop - which is not a fault of the hotel's. But flagging in case you're romanticising a beach swim, the hotel pool is better.
- Watch: hidden costs. This might be normal/acceptable in non-Australian countries but I was caught off guard. There's the room cost, then there's the taxes (which booking.com tends to include in their final price), and THEN the hotel has a 'resort fee'. Which allows for 'amenities' access - which I find a bit "on the nose", the "amenities" is what you automatically have access to when you book a room... but I guess some countries/states prefer a staggered bill...
Some Text Classification Applications

<table>
<thead>
<tr>
<th>Task</th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language identification</td>
<td>text</td>
<td>{English, Mandarin, Greek, …}</td>
</tr>
<tr>
<td>Spam classification</td>
<td>email</td>
<td>{spam, not spam}</td>
</tr>
<tr>
<td>Authorship attribution</td>
<td>text</td>
<td>{jk rowling, james joyce, …}</td>
</tr>
<tr>
<td>Genre classification</td>
<td>novel</td>
<td>{detective, romance, gothic, …}</td>
</tr>
<tr>
<td>Sentiment classification</td>
<td>text</td>
<td>{positive, negative, neutral, mixed}</td>
</tr>
</tbody>
</table>
Lots of Model Choices

- Decision Trees
- Logistic Regression
- Random Forests
- Neural Nets (e.g., BERT)
- Support Vector Machine
- Prompting LLMs
Model Differences

Binary Classification

One out of 2 labels applied to a given x

Multiclass Classification

One out of N labels applied to a given x

Multilabel Classification

Multiple labels apply to a given x
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!
Beyond the Bag of Words

Some linguistic phenomena require going beyond the bag-of-words:

‣ That’s not bad for the first day
‣ This is not the worst thing that can happen
‣ It would be nice if you acted like you understood
‣ This film should be brilliant. The actors are first grade. Stallone plays a happy, wonderful man. His sweet wife is beautiful and adores him. He has a fascinating gift for living life fully. It sounds like a great plot, however, the film is a failure.
Signals in Text Beyond Words

Emojis: 🙏🤷🥰

Special characters: ' }{ [ ] # @ ! * < > ~

Out of vocabulary words: icebucketchallenge, wowwwwww

URLs: https://www.nytimes.com/

Typos or spelling errors: typs, tpos, …

Social media features: @user, RT, #hashtags

Slang words: chill, slay, sick …
Signals in Text Beyond English

Language identification has very high accuracy for long texts, but struggles with social media (short informal) text

Code switching:  I have 2 friends due estudiaron la contabilidad
Applying Text Classification

The “raw” form of text is usually a sequence of characters. Converting this into a meaningful feature vector $x$ requires a series of design decisions, such as tokenization, normalization, and filtering.

We do not cover this in class. Check out this after class if needed.
Vocabulary Size Filtering

A small number of word types accounts for the majority of word tokens.

The number of parameters in a classifier usually grows linearly with the size of the vocabulary. It can be useful to limit the vocabulary, e.g., to word types appearing at least x times, or in at least y% of documents.
## Experiment Setup (Train/Test Split)

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Training</th>
<th>Development Or Validation</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>80%</td>
<td>10%</td>
<td>10%</td>
</tr>
<tr>
<td>Purpose</td>
<td>Training Models</td>
<td>Model Selection (e.g., parameters)</td>
<td>Evaluation (You should never look at it until the very end)</td>
</tr>
</tbody>
</table>
Evaluating Your Classifier

Goal is to predict future performance, on unseen data.

It is hard to predict the future.

Do not evaluate on data that was already used …

For training
For hyperparameter selection
For selecting the classification model or model structure
For making preprocessing decisions, such as vocabulary selection.
Beyond Right and Wrong

For any label, there are two ways to be wrong:

- **False positive**: the system incorrectly predicts the label
- **False negative**: the system incorrectly fails to predict the label.

Similarly, there are two ways to be right:

- **True positive**: the system correctly predicts the label
- **True negative**: the system correctly predicts that the label does not apply to it.
Accuracy

\[
\text{accuracy} = \frac{TP + TN}{TP + FP + FN + TN}
\]

The problem with accuracy is rare labels.

Consider a system for detecting tweets written in Telugu. 0.3% of Tweets are written in Telugu. A system that always says “Not Telugu” is 99.7% accurate.
Recall

Recall = \frac{TP}{TP + FN}

Recall is the fraction of positive instances which were correctly classified. The “never Telugu” classifier has zero recall. An “always Telugu” classifier would have perfect recall.
Precision

Precision = \frac{TP}{TP + FP}

Precision is the fraction of positive predictions that were correct. The “never Telugu” classifier has precision 0/0. An “always Telugu” classifier would have precision p=0.003, which is the rate of Telugu tweets in the dataset.
Combining Recall and Precision

In binary classification, there is an inherent tradeoff btw recall and precision.

The correct navigation of this tradeoff is problem-specific!

For a preliminary medical diagnosis, we might prefer high recall. False positives can be screened out later.

The “beyond a reasonable doubt” standard of U.S. criminal law implies a preference for high precision.
Combining Recall and Precision

In binary classification, there is an inherent tradeoff btw recall and precision.

The correct navigation of this tradeoff is problem-specific!

If recall and precision are weighted equally, they can be combined into a single number called F-measure

\[ F = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \]
Evaluating Multi-Class Classification

Recall and precision imply binary classification: each instance is either positive or negative.

In multi-class classification, each instance is positive for one class, and negative for all other classes.
Evaluating Multi-Class Classification

Two ways to combine performance across classes:

**Macro F-measure:** compute the F-measure per class, and average across all classes. This treats all classes equally, regardless of their frequency.

**Micro F-measure:** compute the total number of true positives, false positives, and false negatives across all classes, and compute a single F-measure. This emphasizes performance on high-frequency classes.
Comparing Classifiers

Suppose you and your friend build classifiers to solve a problem:

- You classifier $C_1$ get 82% accuracy
- You friend’s classifier $C_2$ get 73% accuracy

Will $C_1$ be more accurate in the future?

- What is the test set had 10000 examples?
- What is the test set had 11 examples?
Getting Labels

Text classification relies on large datasets of labeled examples. There are two main ways to get labels:

**Metadata** sometimes tell us exactly what we want to know: Did the Senator vote for a bill? How many stars did the reviewer give? Was the request for free pizza accepted?

Other times, the labels must be annotated, by experts or by “crow-workers”
Let’s build a hate speech classifier for X

What do we need?

1. **Why** do we need to set it up as a classification task?
2. **What** counts as a hate speech?
3. How can we get the **ground truth** for a tweet?
4. How many **data points** do we need?
5. How can we **evaluate** this system?
6. Does the system really **work**?
Let’s build a hate speech classifier for X

A tweet is offensive if it
1. uses a sexist or racial slur.
2. attacks a minority.
3. seeks to silence a minority.
4. criticizes a minority (without a well founded argument).
5. promotes, but does not directly use, hate speech or violent crime.
6. criticizes a minority and uses a straw man argument.
7. blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims.
8. shows support of problematic hash tags. E.g. “#BanIslam”, “#whoriental”, “#whitegenocide”
9. negatively stereotypes a minority.
10. defends xenophobia or sexism.
11. contains a screen name that is offensive, as per the previous criteria, the tweet is ambiguous (at best), and the tweet is on a topic that satisfies any of the above criteria.

“Annotated 16,914 tweets,
3,383 of that for sexist content sent by 613 users,
1,972 for racist content sent by 9 users
11,559 for neither sexist or racist, sent by 614 users”

<table>
<thead>
<tr>
<th></th>
<th>char n-grams</th>
<th>+gender</th>
<th>+gender +loc</th>
<th>word n-grams</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>73.89</td>
<td>73.93</td>
<td>73.62*</td>
<td>64.58</td>
</tr>
<tr>
<td>Precision</td>
<td>72.87%</td>
<td>72.93%</td>
<td>72.58%</td>
<td>64.39%</td>
</tr>
<tr>
<td>Recall</td>
<td>77.75%</td>
<td>77.74%</td>
<td>77.43%</td>
<td>71.93%</td>
</tr>
</tbody>
</table>
Let’s build a hate speech classifier for X

Is this classifier good?

Figure 1: Phrases in African American English (AAE), their non-AAE equivalents (from Spears, 1998), and toxicity scores from PerspectiveAPI.com. Perspective is a tool from Jigsaw/Alphabet that uses a convolutional neural network to detect toxic language, trained on crowdsourced data where annotators were asked to label the toxicity of text without metadata.

Racial Bias in Hate Speech and Abusive Language Detection Datasets

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thomas Davidson</td>
<td>Department of Sociology</td>
<td><a href="mailto:trd54@cornell.edu">trd54@cornell.edu</a></td>
</tr>
<tr>
<td></td>
<td>Cornell University</td>
<td></td>
</tr>
<tr>
<td>Debasmita Bhattacharya</td>
<td>Department of Computer Science</td>
<td><a href="mailto:db758@cornell.edu">db758@cornell.edu</a></td>
</tr>
<tr>
<td></td>
<td>Cornell University</td>
<td></td>
</tr>
<tr>
<td>Ingmar Weber</td>
<td>Qatar Computer Research Institute</td>
<td><a href="mailto:i.weber@hbku.edu.qa">i.weber@hbku.edu.qa</a></td>
</tr>
</tbody>
</table>

The Risk of Racial Bias in Hate Speech Detection

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maarten Sap</td>
<td>Paul G. Allen School of Computer Science &amp; Engineering, University of Washington, Seattle, USA</td>
</tr>
<tr>
<td>Dallas Card</td>
<td>Machine Learning Department, Carnegie Mellon University, Pittsburgh, USA</td>
</tr>
<tr>
<td>Saadia Gabriel</td>
<td>Allen Institute for Artificial Intelligence, Seattle, USA</td>
</tr>
<tr>
<td>Yejin Choi</td>
<td>Allen Institute for Artificial Intelligence, Seattle, USA</td>
</tr>
<tr>
<td>Noah A. Smith</td>
<td>Allen Institute for Artificial Intelligence, Seattle, USA</td>
</tr>
</tbody>
</table>

Exploring the Role of Grammar and Word Choice in Bias Toward African American English (AAE) in Hate Speech Classification

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camille Harris</td>
<td>Georgia Institute of Technology</td>
<td><a href="mailto:charris38@gatech.edu">charris38@gatech.edu</a></td>
</tr>
<tr>
<td>Matan Halevy</td>
<td>Georgia Institute of Technology</td>
<td><a href="mailto:matan@gatech.edu">matan@gatech.edu</a></td>
</tr>
<tr>
<td>Ayazna Howard</td>
<td>The Ohio State University</td>
<td></td>
</tr>
<tr>
<td>Amy Bruckman</td>
<td>Georgia Institute of Technology</td>
<td></td>
</tr>
<tr>
<td>Diyi Yang</td>
<td>Georgia Institute of Technology</td>
<td></td>
</tr>
</tbody>
</table>
Let’s build a hate speech classifier for X

Hate speech goes beyond explicit usage of keywords

**Hate Speech**

**Explicit** Hate Speech
(intentional)

stereotyping,
racism, sexism,
targeted threats

**Implicit** Hate Speech

“[IDENTITY] destroy everything they touch”

“White revolution is the only solution.”

**Latent Hatred: A Benchmark for Understanding Implicit Hate Speech**

Mai ElSherief
Caleb Ziems
David Muchinski
Vaishnavi Anupindi
Jordyn Seybolt
Munmun De Choudhury
Diyi Yang

1 UC San Diego, 2 Georgia Institute of Technology
malsherief@ucsd.edu
cziems, dmuchinski13, vanupindi3@gatech.edu
jseybolt3, munmund, dyang888@gatech.edu
Let’s build a hate speech classifier for X

Hate speech goes beyond explicit usage of keywords

Hate Speech

**Explicit** Hate Speech

*intentional*
- stereotyping,
- racism, sexism,
- targeted threats

**Implicit** Hate Speech

*intentional*
- coded language, colloquialisms, connotations, dog whistles, entity framing, hidden threats, idioms, inferiority assumptions, irony, metaphors, presuppositions, symbolic language, ...
Let’s build an implicit hate speech classifier for X

**Incitement of Violence** (Somerville, 2011; Assembly, 1966)

**Inferiority Language** (Nielsen, 2002; Kennedy et al., 2018)

**Irony** (Waseem and Hovy, 2016; Justo et al., 2014)

**Stereotypes** (Warner and Hirschberg, 2012)

**Threats and Intimidation** (Sanguinetti et al., 2018)

**White Grievance** (Berbrier, 2000; Bloch et al., 2020; Miller-Driss, 2020)
Let’s build an implicit hate speech classifier for X

5M tweets from prominent hate groups

Filter out tweets with explicit hate

Annotation with good agreement Fleiss’ $\kappa = 0.6$
Let’s build an implicit hate speech classifier for X

1st stage fine-tune on existing hate speech datasets via multi-task learning

2nd stage fine-tune on Implicit Hate, with auto-augmentation
Let’s build an implicit hate speech classifier for X

Accuracy

 Implicit Hate or Not  Implicit Hate Categories

Hate Sonar  Google Perspective API  BERT  Ours

25.6% Δ  63.7  78.3  80

2.2% Δ  62.9  65.8

4.6% Δ

**p<0.01: significantly better over BERT**
Recognizing A Classification Problem and Its Complexities

Can you formulate your question as a choice among some possible classes?

Can you create (or find) labeled data that marks that choice for a bunch of examples? Can you make that choice?

Can you create features that might help in distinguishing those classes?

Is the classification enough to capture the nuances?
Regression
Regression

A mapping from input data $x$ (drawn from instance space $X$) to a point $y$ in $R$

$R$: the set of real numbers

$x =$ the empire state building

$y =$ 17444.5625"

Slide content credit to David Bamman
Linear Regression

Suppose we have \( n \) data points. For each data point \( i \), we observe 

\[(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)\]

Linear regression states that 

\[
\hat{y}_i = \sum_{i=1}^{F} x_i \beta_i
\]
Regression for Social Sciences

LINEAR REGRESSION

LINEAR REGRESSION EVERYWHERE
Polynomial Regression

\[ \hat{y}_i = \sum_{i=1}^{F} x_i \beta_{a,i} + \sum_{i=1}^{F} x_i^2 \beta_{b,i} \]

\[ \hat{y}_i = \sum_{i=1}^{F} x_i \beta_{a,i} + \sum_{i=1}^{F} x_i^2 \beta_{b,i} + \sum_{i=1}^{F} x_i^3 \beta_{c,i} \]
Nonlinear Regression

Support vector machines (regression)
Neural Networks

...
Number of Parameters

\[ \hat{y}_i = \sum_{i=1}^{F} x_i \beta_{a,i} \]

\[ \hat{y}_i = \sum_{i=1}^{F} x_i \beta_{a,i} + \sum_{i=1}^{F} x_i^2 \beta_{b,i} \]

\[ \hat{y}_i = \sum_{i=1}^{F} x_i \beta_{a,i} + \sum_{i=1}^{F} x_i^2 \beta_{b,i} + \sum_{i=1}^{F} x_i^3 \beta_{c,i} \]

Slide content credit to David Bamman
Overfitting

Memorizing the nuances (and noise) of the training data that prevents generalizing to unseen data
Sources of Error

**Bias:** Error due to mis-specifying the relationship between input and output

*Too few parameters, or the wrong kinds*

**Variance:** Error due to sensitivity to random fluctuations in the training data. If you train on different data, do you get radically different predictions?

*Too many parameters*
Low variance

Low bias

High variance

High bias

Slide content credit to David Bamman

Image from Flach 2012
Regression analysis is a very useful tool for social sciences

✦ Understand the relationship between variables, adjusting for other potential confounders

✦ Predict the value of one variable based on others
In Other Terminology

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Dependent Variable

Independent Variable

Independent Variable

Intersect
How good is the Fit?

Mean squared error (MSE) \[
\frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2
\]

Mean absolute error (MAE) \[
\frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i|
\]
How good is the “fit”?

Sum of the squares total (SST): total variability about the mean
\[ \sum (Y - \bar{Y})^2 \]

Sum of the squared error (SSE): variability about the regression line
\[ \sum (Y - \hat{Y})^2 \]

Sum of the squares due to regression (SSR): total variability that is explained by the model
\[ \sum (\hat{Y} - \bar{Y})^2 \]
Coefficient of Determination $r^2$

The proportion of the variability explained by regression model

$$r^2 = \frac{SSR}{SST}$$
Recommendations for Building Regression Models

A high $r^2$ is desired with a reasonable set of variables

When more variables get added to the model, $r^2$ usually increases.

Thus, adjusted $r^2$ is often used to account for the number of variables

Independent variables might contain **duplicated** information

**Colinear** if two variables are correlated

**Multicolinearity** if more than two variables are correlated - this will make the interpretation of regression coefficient problematic
Let’s predict tie strength on Facebook

1. **Why** is this a regression task?
2. **What** is tie strength?
3. How can we get the **ground truth**?
4. How to get **data**?
5. How can we **evaluate** it?
6. Does the system really **work**?

https://murraydare.co.uk/marketing-theory/strong-weak-ties
Let’s predict tie strength on Facebook

Mark Granovetter introduced the concept of tie strength in 1973 “The Strength of Weak Ties”

The strength of a tie is a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie

Let’s predict tie strength on Facebook

How strong is your relationship with this person?
barely know them ———————————————————— we are very close

How would you feel asking this friend to loan you $100 or more?
would never ask ———————————————————— very comfortable

How helpful would this person be if you were looking for a job?
no help at all ———————————————————— very helpful

How upset would you be if this person unfriended you?
not upset at all ———————————————————— very upset

If you left Facebook for another social site, how important would it be to bring this friend along?
would not matter ———————————————————— must bring them
Let’s predict tie strength on Facebook

What features we could use to predict self-reported tie strength?

<table>
<thead>
<tr>
<th>Predictive Intensity Variables</th>
<th>Distribution</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wall words exchanged</td>
<td></td>
<td>9549</td>
</tr>
<tr>
<td>Participant-initiated wall posts</td>
<td></td>
<td>55</td>
</tr>
<tr>
<td>Friend-initiated wall posts</td>
<td></td>
<td>47</td>
</tr>
<tr>
<td>Inbox messages exchanged</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>Inbox thread depth</td>
<td></td>
<td>31</td>
</tr>
<tr>
<td>Participant’s status updates</td>
<td></td>
<td>80</td>
</tr>
<tr>
<td>Friend’s status updates</td>
<td></td>
<td>200</td>
</tr>
<tr>
<td>Friend’s photo comments</td>
<td></td>
<td>1352</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Intimacy Variables</th>
<th>Distribution</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant’s number of friends</td>
<td></td>
<td>729</td>
</tr>
<tr>
<td>Friend’s number of friends</td>
<td></td>
<td>2050</td>
</tr>
<tr>
<td>Days since last communication</td>
<td></td>
<td>1115</td>
</tr>
<tr>
<td>Wall intimacy words</td>
<td></td>
<td>148</td>
</tr>
<tr>
<td>Inbox intimacy words</td>
<td></td>
<td>137</td>
</tr>
<tr>
<td>Appearances together in photo</td>
<td></td>
<td>73</td>
</tr>
<tr>
<td>Participant’s appearances in photo</td>
<td></td>
<td>897</td>
</tr>
<tr>
<td>Distance between hometowns (mi)</td>
<td></td>
<td>8182</td>
</tr>
<tr>
<td>Friend’s relationship status</td>
<td></td>
<td>32% married</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Duration Variable</th>
<th>Distribution</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Days since first communication</td>
<td></td>
<td>1328</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reciprocal Services Variables</th>
<th>Distribution</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Links exchanged by wall post</td>
<td></td>
<td>668</td>
</tr>
<tr>
<td>Applications in common</td>
<td></td>
<td>18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Structural Variables</th>
<th>Distribution</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of mutual friends</td>
<td></td>
<td>206</td>
</tr>
<tr>
<td>Groups in common</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>Norm. TF-IDF of interests and about</td>
<td></td>
<td>73</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Emotional Support Variables</th>
<th>Distribution</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wall &amp; inbox positive emotion words</td>
<td></td>
<td>197</td>
</tr>
<tr>
<td>Wall &amp; inbox negative emotion words</td>
<td></td>
<td>51</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Social Distance Variables</th>
<th>Distribution</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age difference (days)</td>
<td></td>
<td>5995</td>
</tr>
<tr>
<td>Number of occupations difference</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>Educational difference (degrees)</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Overlapping words in religion</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Political difference (scale)</td>
<td></td>
<td>4</td>
</tr>
</tbody>
</table>
Let’s predict tie strength on Facebook

The model’s Adjusted R2 values for all five dependent variables, broken down by the model’s three main terms.

Modeling interactions between tie strength dimensions results in a substantial performance boost.

The model performs best on Loan $100? and How strong?, the most general question.
Let’s predict tie strength on Facebook

<table>
<thead>
<tr>
<th>Top 15 Predictive Variables</th>
<th>β</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Days since last communication</td>
<td>-0.76</td>
<td>453</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Days since first communication</td>
<td>0.755</td>
<td>7.55</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Intimacy × Structural</td>
<td>0.4</td>
<td>12.37</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Wall words exchanged</td>
<td>0.299</td>
<td>11.51</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Mean strength of mutual friends</td>
<td>0.257</td>
<td>188.2</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Educational difference</td>
<td>-0.22</td>
<td>29.72</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Structural × Structural</td>
<td>0.195</td>
<td>12.41</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Reciprocal Serv. × Reciprocal Serv.</td>
<td>-0.19</td>
<td>14.4</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Participant-initiated wall posts</td>
<td>0.146</td>
<td>119.7</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Inbox thread depth</td>
<td>-0.14</td>
<td>1.09</td>
<td>0.29</td>
</tr>
<tr>
<td>Participant’s number of friends</td>
<td>-0.14</td>
<td>30.34</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Inbox positive emotion words</td>
<td>0.135</td>
<td>3.64</td>
<td>0.05</td>
</tr>
<tr>
<td>Social Distance × Structural</td>
<td>0.13</td>
<td>34</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Participant’s number of apps</td>
<td>-0.12</td>
<td>2.32</td>
<td>0.12</td>
</tr>
<tr>
<td>Wall intimacy words</td>
<td>0.111</td>
<td>18.15</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

The fifteen predictive variables with highest standardized beta coefficients.

The two Days since variables have large coefficients because of the difference between never communicating and communicating once.

The utility distribution of the predictive variables forms a power-law distribution: **with only these fifteen variables, the model has over half of the information it needs to predict tie strength.**
Let’s predict tie strength on Facebook

Don’t forget error analysis

**rating: 0; prediction: 0.44**

I don’t know why he friended me. But I’m easy on Facebook, because I feel like I’m somehow building (at least a miniscule amount of) social capital, even when I don’t know the person. We went to the same high school and have a few dozen common friends. We’ve never interacted with each other on Facebook aside from the friending.

**rating: 0.96; prediction: 0.47**

This friend is very special. He and I attended the same high school, we interacted a lot over 3 years and we are very very close. We trust each other. My friend are I are still interacting in ways other than Facebook such as IM, emails, phones. Unfortunately, that friend and I rarely interact through Facebook so I guess your predictor doesn’t have enough information to be accurate.

**rating: 0.6; prediction: 0.11**

Ah yes. This friend is an old ex. We haven’t really spoken to each other in about 6 years, but we ended up friending each other on Facebook when I first joined. But he’s still important to me. We were best friends for seven years before we dated. So I rated it where I did (I was actually even thinking of rating it higher) because I am optimistically hoping we’ll recover some of our “best friend”-ness after a while. Hasn’t happened yet, though.