Social Media & Text Analysis
lecture 1 - Introduction

CSE 5539-0010 Ohio State University
Instructor: Wei Xu
Website: socialmedia-class.org
Course Website

http://socialmedia-class.org/

Social media provides a massive amount of valuable information and shows us how language is actually used by lots of people. This course will give an overview of prominent research findings on language use in social media. The course will also cover several machine learning algorithms and the core natural language processing techniques for obtaining and processing Twitter data.

Instructor

Wei Xu is an assistant professor in the Department of Computer Science and Engineering at the Ohio State University. Her research interests lie at the intersection of machine learning, natural language processing, and social media. She holds a PhD in Computer Science from New York University. Prior to joining OSU, she was a postdoc at the University of Pennsylvania. She is organizing the ACL/COLING Workshop on Noisy User-generated Text, serving as a workshop co-chair for ACL 2017, an area chair for EMNLP 2016 and the publicity chair for NAACL 2016.

Time/Place

Fall 2017, CSE 5539-0010 The Ohio State University

Bolz Hall Room 318 | Tuesday 2:20PM – 4:10PM
dual-listed undergraduate and graduate course

[Office Hour] Dreese 495 | Tuesday 4:15PM – 5:15PM

Prerequisites

In order to succeed in this course, you should know basic probability and statistics, such as the chain rule of probability and Bayes’ rule. On the programming side, all projects will be in Python. You should understand basic computer science concepts (like recursion), basic data structures (trees, graphs), and basic algorithms (search, sorting, etc).

Course Readings

Various academic papers

Discussion Board

Piazza (TBA)
This is a special topic class

• hobby (not a mandatory course)
• but is lecture-based and project-based
• advanced and research-oriented
• but strong undergraduate students (sophomore, junior, senior) are encouraged to take this course
Who am I?
Wei Xu

- Assistant Professor in CSE at the Ohio State University
- Postdoctoral researcher at University of Pennsylvania
- PhD from New York University in Computer Science
- Research Areas:
  - Natural Language Processing
  - Social Media
  - Machine Learning
Why Social Media?
Vintage Social Media
so my plane just crashed...
pic.twitter.com/X51BLwa5PS

so yup pic.twitter.com/2WuLUWzpND
2014 Ukrainian Revolution
Impact

• Politics
• Business
• Socialization
• Journalism
• Cyber Bullying
• Rumors / Fake News
• Productivity
• Privacy
• Emotions
• …
• and our language (!)
Research Value

- In contrast to survey/self-report
- A probe to:
  - **real** human behavior
  - **real** human opinion
  - **real** human language use
- Easy to access and aggregate **a lot** of data
- thus **a lot** of information
Mood

https://liwc.wpengine.com/

Source: Golder & Macy. “Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures” Science 2011
“We found that individuals awaken in a good mood that deteriorates as the day progresses—which is consistent with the effects of sleep and circadian rhythm”

Source: Golder & Macy. “Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures” Science 2011

https://liwc.wpengine.com/
"We found that individuals awaken in a good mood that deteriorates as the day progresses—which is consistent with the effects of sleep and circadian rhythm."

"People are happier on weekends, but the morning peak in positive affect is delayed by 2 hours, which suggests that people awaken later on weekends."
Data Science

Source: Drew Conway
Data Science

- is the **practice** of:

  - asking question (formulating hypothesis)
  
  - finding and collecting the data needed (often big data)
  
  - performing statistical and/or predictive analytics (often machine learning)
  
  - discovering important information and/or insights
Data Science

• the infamous definition:

"A data scientist is a statistician who lives in San Francisco. Data Science is statistics on a Mac. A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician."
Marketing
User Profiling

Delighted I kept my Xmas vouchers - Happy Friday to me 😊 #shopping

Source: Volkova, Van Durme, Yarowsky, Bachrach
“Tutorial on Social Media Predictive Analytics” NAACL 2015
User Profiling

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

Source: Volkova, Van Durme, Yarowsky, Bachrach
“Tutorial on Social Media Predictive Analytics” NAACL 2015
User Profiling

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.

Source: Volkova, Van Durme, Yarowsky, Bachrach
“Tutorial on Social Media Predictive Analytics” NAACL 2015
User Profiling

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: newsscientist.com/article/dn2571
Health

Heart Disease Rates as Reported on Death Certificates

Heart Disease Rates as Predicted By Twitter

Source: World Well-Being Project @ University of Pennsylvania
Health

Hostility, Aggression

Hostile, Annoying

$r = .27$

Hate, Interpersonal Tension

Hates, Hate

$r = .21$

Boredom, Fatigue

Tired, Tiring

$r = .20$

Skilled Occupations

Conference, Leadership

$r = -.17$

Positive Experiences

Fabulous, Fantastic

$r = -.15$

Optimism

Overcome, Strength

$r = -.13$

Source: World Well-Being Project @ University of Pennsylvania
What is Natural Language Processing?
Sentiment Analysis

This nets vs bulls game is great
This Nets vs Bulls game is nuts
Wowsers to this nets bulls game
this Nets vs Bulls game is too live
This Nets and Bulls game is a good game
This netsbulls game is too good
This NetsBulls series is intense
Named Entity Recognition

Tim Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, Wei Xu
Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition
Machine Translation

Mingkun Gao, Wei Xu, Chris Callison-Burch. “Cost Optimization for Crowdsourcing Translation” In TACL (2014)
... the forced resignation of the CEO of Boeing, Harry Stonecipher, for ...

Harry Stonecipher

CEO, Boeing

In office

2003-2005

Maria Pershina, Bonan Min, Wei Xu, Ralph Grishman. “Infusion of Labeled Data into Distant Supervision for Relation Extraction” In ACL (2014)


Wei Xu, Ralph Grishman, Le Zhao. “Passage Retrieval for Information Extraction using Distant Supervision” In IJCNLP (2011)
Humanity’s Collective Knowledge is Locked in Text
Information Extraction

Text ➔ Structured Data
“Yess! Yess! Its official Nintendo announced today that they Will release the Nintendo 3DS in north America march 27 for $250”
“Yess! Yess! Its official Nintendo announced today that they Will release the Nintendo 3DS in north America march 27 for $250”
“Yess! Yess! Its official **Nintendo** announced today that they will release the **Nintendo 3DS** in **north America** **march 27** for **$250**”

<table>
<thead>
<tr>
<th>COMPANY</th>
<th>PRODUCT</th>
<th>DATE</th>
<th>PRICE</th>
<th>REGION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**PRODUCT RELEASE**
“Yess! Yess! Its official Nintendo announced today that they Will release the Nintendo 3DS in north America march 27 for $250”

<table>
<thead>
<tr>
<th>COMPANY</th>
<th>PRODUCT</th>
<th>DATE</th>
<th>PRICE</th>
<th>REGION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nintendo</td>
<td>3DS</td>
<td>March 27</td>
<td>$250</td>
<td>North America</td>
</tr>
</tbody>
</table>
### Information Extraction

**Samsung Galaxy S5** Coming to All Major U.S. Carriers Beginning April 11th

<table>
<thead>
<tr>
<th>COMPANY</th>
<th>PRODUCT</th>
<th>DATE</th>
<th>PRICE</th>
<th>REGION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samsung</td>
<td>Galaxy S5</td>
<td>April 11</td>
<td>?</td>
<td>U.S.</td>
</tr>
<tr>
<td>Nintendo</td>
<td>3DS</td>
<td>March 27</td>
<td>$250</td>
<td>North America</td>
</tr>
</tbody>
</table>
Information Extraction

**Samsung Galaxy S5 Coming to All Major U.S.**

- State of the art is maybe 80%, for single easy fields: 90%+
- Redundancy helps a lot!
- Much of human knowledge is waiting to be harvested from the Web!

<table>
<thead>
<tr>
<th>COMPANY</th>
<th>PRODUCT</th>
<th>DATE</th>
<th>PRICE</th>
<th>REGION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samsung</td>
<td>Galaxy S5</td>
<td>April 11</td>
<td>?</td>
<td>U.S.</td>
</tr>
<tr>
<td>Nintendo</td>
<td>3DS</td>
<td>March 27</td>
<td>$250</td>
<td>North America</td>
</tr>
</tbody>
</table>
Paraphrase

- cup
- word
- mug

- the king’s speech
- phrase
- His Majesty’s address

- sentence

... the forced resignation of the CEO of Boeing, Harry Stonecipher, for ...

... after Boeing Co. Chief Executive Harry Stonecipher was ousted from ...

Wuwei Lan, Wei Xu. “Neural Network Models for Paraphrase Identification, Semantic Textual Similarity, Natural Language Inference, and Question Answering” COLING (2018)

Wuwei Lan, Siyu Qiu, Hua He, Wei Xu. “A Continuously Growing Dataset of Sentential Paraphrases” EMNLP (2017)

Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. “Extracting Lexically Divergent Paraphrases from Twitter” In TACL (2014)

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, Colin Cherry. “Paraphrasing for Style” In COLING (2012)
Who is the CEO stepping down from Boeing?

... the forced resignation of the CEO of Boeing, Harry Stonecipher, for ...

... after Boeing Co. Chief Executive Harry Stonecipher was ousted from ...
Who is the CEO stepping down from Boeing?

... the forced resignation of the CEO of Boeing, Harry Stonecipher, for ...

... after Boeing Co. Chief Executive Harry Stonecipher was ousted from ...
Who is the CEO stepping down from Boeing?

... the forced resignation of the CEO of Boeing, Harry Stonecipher, for ...

... after Boeing Co. Chief Executive Harry Stonecipher was ousted from ...
Watson leverages multiple algorithms to perform deeper analysis

[Question]
In May 1898 Portugal celebrated the 400th anniversary of this explorer’s arrival in India.

[Supporting Evidence]
On the 27th of May 1498, Vasco da Gama landed in Kappad Beach

Stronger evidence can be much harder to find and score...
- Search far and wide
- Explore many hypotheses
- Find judge evidence
- Many inference algorithms
Watson leverages multiple algorithms to perform deeper analysis.

[Question]
In May 1898 Portugal celebrated the 400th anniversary of this explorer’s arrival in India.

[Supporting Evidence]
On the 27th of May 1498, Vasco da Gama landed in Kappad Beach.

Legend:
- Temporal Reasoning
- Statistical Paraphrasing
- GeoSpatial Reasoning

Stronger evidence can be much harder to find and score...
- Search far and wide
- Explore many hypotheses
- Find judge evidence
- Many inference algorithms

© 2014 IBM Corporation
(courtesy: Salim Roukos)
Natural Language Generation

who wants to get a beer?

want to get a beer?

who else wants to get a beer?

who wants to go get a beer?

who wants to buy a beer?

who else wants to get a beer?

trying to get a beer?

... (21 different ways)


Wei Xu, Alan Ritter, Ralph Grishman. “Gathering and Generating Paraphrases from Twitter with Application to Normalization” In BUCC (2013)
Data-Driven Conversation

• **Twitter:** ~ 500 Million Public SMS-Style Conversations *per Month*

• **Goal:** Learn conversational agents directly from massive volumes of data.
Data-Driven Conversation

- **Twitter:** ~ 500 Million Public SMS-Style Conversations *per Month*

- **Goal:** Learn conversational agents directly from massive volumes of data.
Noisy Channel Model

Input:

Who wants to come over for dinner tomorrow?
Noisy Channel Model

Input:

Who wants to come over for dinner tomorrow?

Output:

Yum! I
Noisy Channel Model

Input:

Who wants to come over for dinner tomorrow?

Output:

Yum! I want to
Noisy Channel Model

Input:

Who wants to come over for dinner tomorrow?

Output:

Yum! I want to be there
Noisy Channel Model

Input:

Who wants to come over for dinner tomorrow?

Output:

Yum! I want to be there tomorrow!
Another bizarre feature of our early prototype was its propensity to respond with “I love you” to seemingly anything. As adorable as this sounds, it wasn’t really what we were hoping for. Some analysis revealed that the system was doing exactly what we’d trained it to do, generate likely responses -- and it turns out that responses like “Thanks”, "Sounds good", and “I love you” are super common -- so the system would lean on them as a safe bet if it was unsure. Normalizing the
Another bizarre feature of our early prototype was its propensity to respond with “I love you” to seemingly anything. As adorable as this sounds, it wasn’t really what we were hoping for. Some analysis revealed that the system was doing exactly what we’d trained it to do, generate likely responses – and it turns out that responses like “Thanks”, "Sounds good", and “I love you” are super common -- so the system would lean on them as a safe bet if it was unsure. Normalizing the
Language Technology

making good progress

mostly solved

Sentiment analysis
- Best roast chicken in San Francisco!
- The waiter ignored us for 20 minutes.

Coreference resolution
- Carter told Mubarak he shouldn’t run again.

Word sense disambiguation (WSD)
- I need new batteries for my mouse.

 Parsing
- I can see Alcatraz from the window!

Machine translation (MT)
- The 13th Shanghai International Film Festival...

Information extraction (IE)
- You’re invited to our dinner party, Friday May 27 at 8:30

still really hard

Question answering (QA)
- Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase
- XYZ acquired ABC yesterday
- ABC has been taken over by XYZ

Summarization
- The Dow Jones is up
- The S&P500 jumped
- Housing prices rose
- Economy is good

Dialog
- Where is Citizen Kane playing in SF?
- Castro Theatre at 7:30. Do you want a ticket?
What will we cover in this class (and should you take it)?
What do you expect to learn

• Twitter API for obtaining Twitter data
• cutting edge research on:
  - Natural Language Processing (NLP)
  - Machine Learning
• useful NLP tools, especially for Twitter text
• basic machine learning algorithms:
  - Naïve Bayes, Logistic Regression
  - Probabilistic Graphical Models
  - Some deep learning basics
Guest Lectures

- At least one guest lecture from other NLP faculty members and/or industry, student researchers
Grading

• two programing assignments (30% individual)

• A 3rd assignment/research project (optional, 20% bonus)

• in-class presentation (20% group of two)

• paper summaries (20% individual, about 10 papers)

• several take-home Quizzes (15% individual)

• participation in class discussions (15%)
Grading

- two programming assignments (30% individual)
- in-class presentation (20% group of two)
- paper summaries (20% individual, about 10 papers)

- Grading on a 12-point scale — 10 for normal completion, 2 for going above and beyond. Final letter grade of the class will be graded on the curve.
Programming Assignments

• All in Python

• two programing assignments (30% — individual)
  1. Twitter’s Language Mix (on the course website now)
  2. Logistic Regression Algorithm (use Numpy package)

• a third assignment (optional — group recommended)
  3. Deep Learning Basics and Word2Vec
In-class Presentation

• a 12 minute presentation (20%)
  - A Social Media Platform
  - Or a research paper from NLP Researchers
  - Rehearse! We will use a timer as TED Talk
Survey a Social Media Platform, NLP Researcher or Dataset: In-class Presentation (20 points)

You will pair together (2 students) and give a 10-minute presentation (plus 2-minute Q&A) in class about a social media platform (an incomplete list here) or a paper from NLP researchers of your choice (an incomplete list of NLP groups here). You are also encouraged to find other NLP researchers that are not on this list through CS department homepages or top NLP conferences/journals (e.g. ACL, NAACL, TACL, EMNLP).

First, please sign up to pick a date you want to present, and pick a social media platform or a NLP researcher.

After your presentation in the class, upload your slides to OSU’s Carmen system. Your slides will be also published on this course website.

For NLP researchers, you may focus on

- Who: You are encouraged to consider NLP researchers who are current phd students and post-docs, as well as researchers in industrial labs. Summarize his/her career. How and why do they become successful?
- What: What research topics they are working on? What are they famous for? What does his/her first NLP paper look like? Present one of his/her important or recent work.

For social media platforms, you may focus on:

- Market: When it was founded, purchased, and etc?
- Interface: How people use it, and why?
- Software Development: Any API available?
- Academic Research: Any interesting studies? Any useful datasets?
- and any other things you think are important
# In-class Presentation

## 5539 Presentations (2019AU)

<table>
<thead>
<tr>
<th>Date</th>
<th>Name of NLP Researcher/Paper, or Social Media Platform/Dataset</th>
<th>Student Presentation Group #1</th>
<th>Student Presentation Group #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>8/22/2019</td>
<td>1st class - no student presentation</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>8/29/2019</td>
<td></td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>9/5/2019</td>
<td></td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>9/12/2019</td>
<td></td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>9/19/2019</td>
<td></td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>9/26/2019</td>
<td></td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>10/3/2019</td>
<td></td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>10/4/2019</td>
<td>Autumn Break</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>10/17/2019</td>
<td></td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>10/24/2019</td>
<td></td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>10/31/2019</td>
<td></td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>11/7/2019</td>
<td></td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>11/14/2019</td>
<td></td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>11/21/2019</td>
<td></td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>11/28/2019</td>
<td>Thanksgiving</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>12/4/2019</td>
<td>12/4 is the last day of classes</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>12/5/2019</td>
<td></td>
<td>----</td>
<td>----</td>
</tr>
</tbody>
</table>

Wei Xu  socialmedia-class.org
Quizzes

• several simple take-home quizzes
• hard-copy on paper
• will not be graded; but count for 10 points
• We have Quiz #0 today on class survey!
Paper Summaries

• roughly one paper assigned for reading per week
• about 10 papers in total
• allowed to skip two papers throughout the semester
• write a short summary between 100-200 words:
  - discuss positive aspects and limitations
  - suggest potential improvement or extensions
Paper Summaries

• Hal Daumé III's infamous NLP blog

P16-1009: Rico Sennrich; Barry Haddow; Alexandra Birch
Improving Neural Machine Translation Models with Monolingual Data

I like this paper because it has a nice solution to a problem I spent a year thinking about on-and-off and never came up with. The problem is: suppose that you're training a discriminative MT system (they're doing neural; that's essentially irrelevant). You usually have far more monolingual data than parallel data, which typically gets thrown away in neural systems because we have no idea how to incorporate it (other than as a feature, but that's blech). What they do here is, assuming you have translation systems in both directions, back translate your monolingual target-side data, and then use that faux-parallel-data to train your MT system on. Obvious question is: how much of the improvement in performance is due to language modeling versus due to some weird kind of reverse-self-training, but regardless the answer, this is a really cool (if somewhat computationally expensive) answer to a question that's been around for at least five years. Oh and it also works really well.
Research Project

• Optional

• Build a machine translation system and **web demo** that can transfer contemporary English text into Shakespearean style!
Stylistic Language Generation

Palpatine:

*If you will not be turned, you will be destroyed!*

*If you will not be turn’d, you will be undone!*

Luke:

*Father, please! Help me!*

*Father, I pray you! Help me!*

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, Colin Cherry. “Paraphrasing for Style” In COLING (2012)
Stylistic Language Generation

- Data and code:

  https://github.com/cocoxu/Shakespeare/

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, Colin Cherry. “Paraphrasing for Style” In COLING (2012)
Stylistic Language Generation

• It has yet become a popular student research project:

  - Stanford students: https://web.stanford.edu/class/cs224n/reports/2757511.pdf
  - University of Maryland students: http://xingniu.org/pub/styvar_emnlp17.pdf
  - CMU students: https://arxiv.org/abs/1707.01161

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, Colin Cherry. “Paraphrasing for Style” In COLING (2012)
Language Styles

she says

wonderfully delightfully beautifully fine well good nicely superbly

he says

Source: Daniel Preotuc-Pietro, Wei Xu and Lyle Ungar
“Discovering User Attribute Stylistic Differences via Paraphrasing” AAAI 2016
What will you get out of this class?

- Understanding of an emerging field of CS
- Programming and machine learning skills useful in industry companies and academic research
- Getting a taste of research and being prepared
Office Hour

- Have a question? Ask in/after class
- Or ask on Piazza discussion board
- Office Hour: TBA
Piazza Discussion Broad

How to Read a Technical Paper

One of you asked a good question -- "how to read a paper?".

In general, I think there is no single best way to read a paper -- it depends on. Many of you are writing very good and thoughtful reading notes in Carmen. We will discuss from time to time in the class, so hopefully you will learn from those discussions and from other people's thoughts.

That being said, Jason Esiner has written down some good advice on how to read a technical paper:
http://cs.jhu.edu/~jason/advice/how-to-read-a-paper.html

As I mentioned in class earlier, you may find other useful advice on Quora, and just by Googling it.

followup discussions

for lingering questions and comments

Start a new followup discussion

Compose a new followup discussion

Average Response Time... Special Mentions:
By Next Class:
- Sign up for in-class presentation
- HW#1 Twitter’s Language Mix