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)
History of the Course

- Summer 2015, University of Pennsylvania
- Summer 2016, North American Summer School on Logic, Language, and Information (NASSLLI)
- Now, since Fall 2016, Ohio State University

Teaching Evaluation @ NASSLLI 2016

- Communication: 5.5 (this course)
- Preparation: 5.5 (this course)
- Interestingness: 5.5 (this course)
- Difficulty-level: 5.5 (this course)
- Good-fit: 6.0 (this course)
- Up-to-date: 5.0 (this course)
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?
Wee 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
We have a TA!
(supported by my research fund)
Pravar Mahajan

- 2nd year Masters student in CSE
- Research Intern, IBM Almaden Research Center
- Worked at Goldman Sacks; studied at IIT Madras
- top student in Fall 2016 class, recruited as RA
- Current research project:
  - Semantic Analysis of Hashtags
HashtagMaster

#songsonghaddafisitunes

Songs On Ghaddafis iTunes
Why Social Media?
Vintage Social Media
Broader Point of View

The Conversation
The Art of Listening, Learning, and Sharing

Source: http://www.conversationprism.com/
so my plane just crashed... pic.twitter.com/X51BLwa5PS

so yup pic.twitter.com/2WuLWzpND
Impact

- Politics
- Business
- Socialization
- Journalism
- Cyber Bullying
- Productivity
- Privacy
- Emotions
- ...
- and our language (!)
2014 Ukrainian Revolution
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

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

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

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

Source: Twitter Ads  https://www.youtube.com/watch?v=K8KJWoNk_Rg
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

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

- Hostility
- Aggression

$r = .27$

Hate, Interpersonal Tension

- Hate
- Interpersonal Tension

$r = .21$

Boredom, Fatigue

- Boredom
- Fatigue

$r = .20$

Skilled Occupations

- Skilled
- Occupations

$r = -.17$

Positive Experiences

- Positive
- Experiences

$r = -.15$

Optimism

- Optimism

$r = -.13$

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, Ritter, Wei Xu

Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition
Machine Translation

To the airport, please.

Bis zum Flughafen, bitte.
Information Extraction

... 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)
Paraphrase

Paraphrase

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

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


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

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

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

Legend
- Temporal Reasoning
- Statistical Paraphrasing
- GeoSpatial Reasoning
- Reference Text
- Answer

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

(courtesy: Salim Roukos)
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
- Reference Text
- Answer

Stronger evidence can be much harder to find and score...

- Search far and wide
- Explore many hypotheses
- Find judge evidence
- Many inference algorithms

(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)
Language Technology

making good progress

mostly solved

Spam detection
- Let’s go to Agra!
- Buy V1AGRA ...

Part-of-speech (POS) tagging
- ADJ
- ADJ
- NOUN
- VERB
- ADV
- Colorless green ideas sleep furiously.

Named entity recognition (NER)
- PERSON
- ORG
- LOC
- Einstein met with UN officials in Princeton

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

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

Summarization
- 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 programming assignments (45 pts/individual)
• A 3rd assignment/research project (optional, 20 bonus pts)
• in-class presentation (20 pts/group of two)
• paper summaries (20 points/individual, about 10 papers)
• several take-home Quizzes (10 points/individual)
• participation in class discussions (5 pts)
Programming Assignments

• All in Python

• two programing assignments (45 points — 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 10 minute presentation (20 points)
  - A Social Media Platform
  - Or a NLP Researcher
Quizzes

• several simple take-home quizzes (about 5 or 6)
• hard-copy on paper
• will not be graded; but count 10 points

• We have Quiz #1 today on pre-requirements!
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

I and my collaborators released the data and code:

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

she says

wonderfully delightfully beautifully fine well good nicely superbly

he says

Source: Daniel Preotşăc-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

• Summer internships?
Office Hour

• Have a question? Ask at/after each class

• Or ask on Piazza discussion broad

• Office hour — Tuesday 4:15-5:15pm (Dreese 495)
Piazza Discussion Broad

Read tips and tricks for a successful Piazza

Enroll your students
Paste email addresses below in any format. Or visit Manage Class page to upload your Class Signup Link.
Each will receive a welcome email.

Welcome to Piazza!
Piazza is a Q&A platform designed to get you great answers from classmates and instructors. We've put together the

Student Enrollment
0 enrolled

Are there TAs/other instructors in your course?
By Next Class:
- Hand in Quiz #1
- HW#0 Become a Twitter User