

CAD 4773/6773 Social Media Mining (Fall'19)

Classes Monday and Wednesday, 2:00–3:15 p.m., room BSN 2205.

Instructor Giovanni Luca Ciampaglia <glc3@usf.edu>.

Office hours (Instructor) Wednesdays 4:00-5:30 p.m. (ENB 318).

TA TBD

Office hours (TA) TBD

Website <http://glciampaglia.com/teaching/cad4773-6773>.

Course description

Have you ever wondered how Facebook is able to tell who are the people you may know? When does Twitter know that a hashtag is trending? Why does Yelp highlight certain reviews but not others?

Social media are nowadays ubiquitous and seem to enter every aspect of our life. At the heart of many of these platforms are digital traces left by our social interactions, individual tastes and preferences, and collective activities, which are generated as part of the everyday operation of social media platforms. But vast as they may be, all these petabytes of data would be of little value without a way to extract information from them. The field of *social media mining* is concerned with turning these deposits of raw information into actionable knowledge and insight.

Course objectives & learning outcomes

This Special Topics course will have a strong applied flavor. All topics will be explored from a computational perspective. We will learn what type of social data are available through a number of platforms. Examples of these will include: Facebook, Twitter, LinkedIn, Yelp, and Github. You will learn to acquire, process, analyze and visualize social media data; to organize your code and workflows with version control systems; and to employ opensource toolkits for data mining and machine learning such as *scikit-learn* or *Gephi*.

Course topics

Taking this course, you should expect to learn about:

- Supervised learning: Crash course on Data Classification.
 - Eager vs. Lazy learning: Decision Tree and k-Nearest Neighbors.
 - Ensemble methods, bagging and boosting: Random Forest and AdaBoost.
 - Classification performance evaluation: Precision/Recall/F1, Accuracy and ROC Curves.
- Unsupervised learning: Crash course on Clustering Data.
 - Distance and similarity measures & K-means clustering.

- Hierarchical Clustering and Dendrograms.
- Density-based clustering.
- Clustering performance evaluation.
- Applications of texts and documents analysis.
 - Natural Language Processing and Part-of-speech tagging.
 - Sentiment Analysis.
 - Topic Modeling.
- Networks.
 - Statistical descriptors of networks: link analysis, centrality, and prestige.
 - Network clustering: modularity and community detection.
 - Dynamics of information and epidemics spreading: threshold and information cascade models.
 - Network visualization algorithms: spring-like layouts, multidimensional scaling, Gephi.
- Collective intelligence.
 - Recommender systems & Collaborative filtering

Prerequisites and course requirements

A basic understanding of programming that will allow you to manipulate data and implement basic algorithms, using any programming language, is recommended but not required. A basic understanding of statistics and algebra will help too.


CAD 4773 only: undergraduate students must have completed Data Structures (COP 4530, minimum grade: C-) and Computer Logic Design (CDA 3201, minimum grade: C-) to enroll in this course.

Programming language

The “official” programming language of the course will be Python. We will use it during hands-on sessions and for the coding assignments. We will use it in conjunction with Jupyter (formerly known as IPython Notebook), an interactive programming environment. You are welcome to use any programming language you like during the hackaton and for the final paper, but please be advised that neither the instructor nor the TA(s) may be able to assist you with language-specific problems if you use a language other than Python.

Textbooks and course materials

Reza Zefarani, Mohammad Ali Abbasi, Huan Liu, “Social Media Mining: An Introduction.” Cambridge University Press. ISBN: 978-1107018853.

 *Required textbook.*

Grading

This is the (tentative) system that will be employed for grading:

| Component | Weight | Description |
|---------------|--------|--|
| Participation | 20% | Class participation, weekly presentation, and engagement. Attendance is mandatory. |
| Assignments | 20% | Five assignments on social Web data analysis and modeling. |
| Midterm exam | 30% | Mid-term Hackathon (grading will be informed by a peer-review system). |
| Final exam | 30% | Final project paper. |

The following misconducts will automatically result in a zero weight for that component of the grade: (1) failing to attend class on the day of your presentation; (2) failing to turn in the assignments by the expected dates; (3) failing to attend meetings of your group's Hackathon and/or final presentation; (4) failing to submit your final paper by the expected date. Extenuating circumstances will normally include only serious emergencies or illnesses documented with a doctor's note.


Course policies

Technology and classroom policies

Class participation and engagement are essential ingredients for success in your academic career. There is a growing body of evidence that laptops and other technological devices are detrimental to the learning experience. Classes of this course, however, will feature two different styles of lectures: traditional-style lectures, and hands-on coding sessions. Therefore, usage of electronic devices is regulated as follows:

Traditional-style lectures Laptops and tablets must be either turned off or in stand-by mode with the lid closed. Exceptions will be granted, for note-taking purposes only, to those who cannot accommodate otherwise; please contact the instructor at the beginning of the semester to obtain a waiver. If you are granted a waiver you must sit in the front rows of the classroom. No email, social media, games, or other distractions will be accepted and will be treated as disruption to academic process (see below).

Hands-on coding sessions During these sessions it is strongly recommended that you use your laptop for coding. If you do not have access to one, please contact the instructor for accommodations. No email, social media, games, or other distractions will be accepted and will be treated as disruption to academic process (see below).

 **At all times** Cellphones and other noise-making devices must be silenced (no vibrate mode). No calls, messaging, email, social media, games, or other distractions will be accepted and will be treated as disruption to academic process (see below).

Student expectations

Mandatory attendance Students will be expected to do all readings and assignments, and to attend all meetings unless excused, in writing, at least 24 hours prior. Please arrive on time for all class meetings.

USF Policies Policies about disability access, religious observances, academic grievances, academic misconduct, and several other topics are governed by a central set of policies that apply to all classes at USF. These may be accessed at: <https://www.usf.edu/provost/faculty-info/core-syllabus-policy-statements.aspx>.


Learning activities

Readings & discussion

There will be readings to do before each class. At the beginning of each lecture (starting week 2), one graduate student will hold a 10 min. presentation on one daily reading and moderate a 5 min. discussion (open to both graduate and undergraduate students) about it. Undergraduate students may fill any empty presentation slots, if any is still available. The list of required readings is available at the end of the syllabus.

Assignments

Throughout the course there will be 5 assignments to be carried out independently by each student. The goal of these assignments is to allow you to track your own progresses and understand whether you are grasping the essential concepts of the course. They will occur tentatively at the end of each of the five parts the course plan (see Schedule). The assignments will consist of part “theory” (including material from the mandatory readings) and part coding tasks. They will be based on topics, problems and questions discussed during class each week.

 *Late assignments:* you have 6 free late days for the entire course. For each extra late day (after the 6th day that is), the score on that assignment will be reduced by 10%.

Mid-term hackathon

The mid-term exam will be a group-based hackathon (<http://en.wikipedia.org/wiki/Hackathon>). The goal of this activity is to develop your intellectual, teamwork, and project management skills. Groups of 3-4 members will be formed (graduate student groups maximum 3) in advance of the mid-term. Each group will work on a different problem approved by the instructor.

A list of problem topic will be made available in advance. Groups may also submit their own project proposals no later than 3 days before the beginning of mid-term week. Project proposals will be subject to approval. Proposal should be 1 page long and should include at least the following information:

Problem definition Clearly state what is the problem being solved.

Motivation Why is this problem relevant?

Approach How the group plans to solve the problem.

References Bibliographic references to at least one relevant related paper.

Proposals that fail to comply with this format may be rejected without review. The rules of the hackathon will be released the week before the mid-term. Each group will receive a 15m slot for presentation of their results, in which each member of the group is expect to discuss at least one critical task of the project. The grading of the projects will be in part based on crowd-sourced ratings attributed by other fellow students and submitted in anonymous form at the end of each presentation day.

Final project and paper

Each student will complete a final project and write a paper report about it. The project may be based on the mid-term hackathon project. Text with other group members cannot be shared, figures/tables can be shared when appropriate, and with proper credit attribution. Grading will be based on soundness (both quality and quantity of original work).

A final paper will be expected. The paper will be at least 1,500 words (*CAD 4773*) and 3,000 words (*CAD 6773*). It should include at least the following sections:

Introduction Clearly state what is the problem being solved. Why is this problem relevant?

Methods How the student plans to solve the problem; clearly discuss the methods being used.

Results Discussion of results, findings, and any limitations; at least one figure or table should be present.

Discussion Related literature and conclusions.

References Bibliographic references.

Code *CAD 6773 only*: Link to GitHub repository with code and data for reproducing the results.

CAD 6773 only: Papers must be formatted using a conference template. The possibility to submit a joint project report will be available for graduate students only (maximum 3 students per group). Please contact the instructor in advance for approval of a joint project.

Schedule


Classes are Mondays and Wednesdays 2:00–3:15 p.m., room BSN 2205. The tentative schedule for the semester is as follows:

| Week | Day | Topics | Readings |
|-----------------------------|------|---|--|
| PART 1: SUPERVISED LEARNING | | | |
| 1 | Mon. | Welcome to the course; introduction to supervised learning | Domingos (2012); Jones (2014), WDM (§ 3.1) |
| | Wed. | Eager vs lazy learning: Decision Trees, k -Nearest Neighbors | Lazer et al. (2014), WDM (§ 3.2, 3.9) |
| 2 | Mon. | Ensemble methods, bagging and boosting, classification performance evaluation | Dhar (2013), WDM (§ 3.3, 3.10) |

| | | | |
|-------------------------------|------|--|---|
| | Wed. | <i>Hands-on session:</i> mining Twitter | Fan and Gordon (2014), MtSW (Chap. 1, pp. 5–26), Twitter API (https://dev.twitter.com/) |
| 3 | Mon. | Labor day; no class | Nuzzo (2014), MtSW (Chap. 1, pp. 26–44) |
| | Wed. | <i>Hands-on session:</i> mining Twitter | Szabo and Huberman (2010), PCI (Chap. 7, pp. 142–165) or alternatively § 1.10 of the scikit-learn User Guide at http://scikit-learn.org/stable/user_guide.html |
| PART 2: UNSUPERVISED LEARNING | | | |
| 4 | Mon. | Introduction to Unsupervised learning. Distance measures, <i>K</i> -means clustering | Vespignani (2009, 2012), WDM (§ 4.1–4.3, pp. 133–147) |
| | Wed. | <i>Hands-on session:</i> mining Twitter | Kosinski et al. (2013) MtSW (Chap. 9, pp. 351–396) |
| 5 | Mon. | Hierarchical clustering, dendrograms | Liben-Nowell and Kleinberg (2008), WDM (§ 4.3–4.5, pp. 147–155) |
| | Wed. | <i>Hands-on session:</i> mining LinkedIn | Schich et al. (2014), MtSW (Chap. 3, pp. 89–132), LinkedIn API (https://developer.linkedin.com/apis) |
| 6 | Mon. | Density-based clustering, clustering performance evaluation | Rodriguez and Laio (2014), WDM (§ 4.6–4.10, pp. 155–165) |
| | Wed. | <i>Hands-on session:</i> mining LinkedIn | Gastner and Newman (2004), PCI (Chap. 3, pp. 29–53) or alternatively § 2.3.1, 2.3.2, and 2.3.6 of the scikit-learn User Guide at: http://scikit-learn.org/stable/user_guide.html |
| PART 3: TEXT & DOCUMENTS | | | |
| 7 | Mon. | Natural language processing, part-of-speech tagging | i Cancho and Solé (2001) WDM (§ 6.5), MtSW (§ 5.3–5.5, pp. 190–222) |
| | Wed. | Sentiment Analysis, <i>hands-on session:</i> mining Yelp | Golder and Macy (2011), MtSW (Chap. 4, pp. 135–180), Yelp Open Dataset (https://www.yelp.com/dataset/documentation/main) |
| 8 | Mon. | Topic modeling | Blei (2012), WDM (§ 6.7) |
| | Wed. | <i>Hands-on session:</i> mining Instagram | Centola (2010, 2011), Instagram API (http://instagram.com/developer/) |
| 9 | Mon. | Mid-term hackaton presentations | |
| | Wed. | Mid-term hackaton presentations | |
| PART 4: NETWORKS | | | |
| 10 | Mon. | Introduction to networks, statistical descriptors of networks | Borgatti et al. (2009); Lazer et al. (2009), NS (Chap. 1–2) |
| | Wed. | <i>Hands-on session:</i> mining Facebook | Cho (2009) WDM (§ 7.1, 7.3–7.4), MtSW Chap. 7 (pp. 279–320) Facebook API (https://developers.facebook.com/) |
| 11 | Mon. | Network clustering | Mucha et al. (2010); Rosvall and Bergstrom (2008), NS (Chap. 9), WDM (§ 7.5) |
| | Wed. | <i>Hands-on session:</i> mining Facebook | Dodds et al. (2003), MtSW (Chap. 2, pp. 45–86) |

| | | | |
|---------------------------------|------|--|--|
| 12 | Mon. | Dynamics of information and epidemics spreading | Metaxas and Mustafaraj (2012) NS (§ 10.1–10.3, pp. 11–29) |
| | Wed. | <i>Hands-on session:</i> tutorial on Gephi | Bond et al. (2012); Kramer et al. (2014), NS (§ 10.4–10.7, pp. 30–58), Gephi Wiki (https://wiki.gephi.org/index.php/Main_Page) |
| 13 | Mon. | Network visualization algorithms | Aral and Walker (2012); Muchnik et al. (2013), PCI (Chap. 12, pp. 300–302) (MDS) or alternatively § 2.2.8 of the scikit-learn User Guide at: http://scikit-learn.org/stable/user_guide.html |
| | Wed. | <i>Hands-on session:</i> tutorial on Gephi | Crandall et al. (2010); Liben-Nowell et al. (2005) |
| PART 5: COLLECTIVE INTELLIGENCE | | | |
| 14 | Mon. | Recommender systems: collaborative filtering algorithm | Koren (2010); Schafer et al. (2007), WDM (§ 12.4) |
| | Wed. | Recommender systems: Non-negative Matrix Factorization algorithm | Lee and Seung (1999), PCI (Chap. 10, pp. 226–249) or alternatively § 2.5.6 of the scikit-learn User Guide at http://scikit-learn.org/stable/user_guide.html |
| 15 | Mon. | Project presentations | |
| | Wed. | Project presentations | |
| 16 | Mon. | Final project reports due on Canvas (11:59 p.m. ET) | |

Reading list

 All papers will be made available on the course website at <http://glciampaglia.com/teaching/cad4773-6773/>

Sinan Aral and Dylan Walker. Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341, 2012.

David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

Robert M Bond, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298, 2012.

Stephen P Borgatti, Ajay Mehra, Daniel J Brass, and Giuseppe Labianca. Network analysis in the social sciences. *Science*, 323(5916):892–895, 2009.

Damon Centola. The spread of behavior in an online social network experiment. *Science*, 329(5996):1194–1197, 2010.

Damon Centola. An experimental study of homophily in the adoption of health behavior. *Science*, 334(6060):1269–1272, 2011.

A Cho. Ourselves and our interactions: the ultimate physics problem? *Science*, 325(5939):406, 2009.

David J Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52):22436–22441, 2010.

- Vasant Dhar. Data science and prediction. *Communications of the ACM*, 56(12):64–73, 2013.
- Peter Sheridan Dodds, Roby Muhamad, and Duncan J Watts. An experimental study of search in global social networks. *Science*, 301(5634):827–829, 2003.
- Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.
- Weiguo Fan and Michael D Gordon. The power of social media analytics. *Communications of the ACM*, 57(6):74–81, 2014.
- Michael T Gastner and Mark EJ Newman. Diffusion-based method for producing density-equalizing maps. *Proceedings of the National Academy of Sciences of the United States of America*, 101(20):7499–7504, 2004.
- Scott A Golder and Michael W Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881, 2011.
- Ramon Ferrer i Cancho and Richard V Solé. The small world of human language. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1482):2261–2265, 2001.
- N Jones. Computer science: The learning machines. *Nature*, 505(7482):146, 2014.
- Yehuda Koren. Collaborative filtering with temporal dynamics. *Communications of the ACM*, 53(4):89–97, 2010.
- Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.
- Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, page 201320040, 2014.
- D Lazer, R Kennedy, G King, and A Vespignani. Big data. the parable of google flu: traps in big data analysis. *Science*, 343(6176):1203, 2014.
- David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Life in the network: the coming age of computational social science. *Science*, 323(5915):721, 2009.
- Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.
- David Liben-Nowell and Jon Kleinberg. Tracing information flow on a global scale using internet chain-letter data. *Proceedings of the National Academy of Sciences*, 105(12):4633–4638, 2008.
- David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628, 2005.
- Panagiotis T Metaxas and Eni Mustafaraj. Social media and the elections. *Science*, 338(6106):472–473, 2012.
- Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878, 2010.

- Lev Muchnik, Sinan Aral, and Sean J Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013.
- R Nuzzo. Scientific method: statistical errors. *Nature*, 506(7487):150–152, 2014.
- Alex Rodriguez and Alessandro Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014.
- Martin Rosvall and Carl T Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.
- J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. In *The adaptive web*, pages 291–324. Springer, 2007.
- Maximilian Schich, Chaoming Song, Yong-Yeol Ahn, Alexander Mirsky, Mauro Martino, Albert-László Barabási, and Dirk Helbing. A network framework of cultural history. *Science*, 345(6196):558–562, 2014.
- Gabor Szabo and Bernardo A Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.
- Alessandro Vespignani. Predicting the behavior of techno-social systems. *Science*, 325(5939):425, 2009.
- Alessandro Vespignani. Modelling dynamical processes in complex socio-technical systems. *Nature Physics*, 8(1):32–39, 2012.