Fall 2025

CS 415/515

Social Media Data Science Pipelines

Last updated: August 22, 2025

Instructor: Kai-Cheng Yang
Email: yangkc@binghamton.edu
Office: G06A, Engineering Building
Lecture: Tuesday & Thursday 9:45am - 11:15am
Office hours: Tuesday 12:30pm - 2:30pm and by appointment

Teaching Assistant: TBD

COURSE DESCRIPTION

The focus of this course is on applying data science techniques to large-scale social media. The topics covered include large-scale data collection and management, exploratory analysis and measurement techniques, data visualization, hypothesis testing and statistical modeling, and predictive, real-time analytics. Students will build an end-to-end analysis pipeline and use it to answer questions about online events as they occur. The goal of the class is to provide students with a methodological toolbox, the technical skills to make use of these tools, and the experience of using them on real world data.

USEFUL LINKS

The course website is https://yangkclab.github.io/social-media-ds-course, which contains the schedule and other course resources. Please check the website regularly for updates.

The up-to-date syllabus can be downloaded from https://github.com/yangkclab/social-media-ds-course/blob/main/syllabus/syllabus.pdf.

CREDIT HOURS

This course is cross-listed as CS 415 and CS 515.

Both sections are 3 credit hours, which means that in addition to the scheduled lectures/discussions, students are expected to do at least 6.5 hours of course-related work each week during the semester. This includes things like: completing assigned readings, participating in lab sessions, studying for tests and examinations, preparing written assignments, and other tasks that must be completed to earn credit in the course.

COURSE OBJECTIVES

This course is designed to provide a solid foundation and background in performing data science on social media. In particular, upon successful completion of this course, you will be able to:

- Build a continuous data system for social media.
- Manage collected data.
- Design and execute various measurements on social media.
- Model and analyze online behavior via social media.
- Create visualizations that help understand social media phenomena.

PREREQUISITES AND CO-REQUISITES

- CS 350 Operating Systems
- CS 375 Design & Analysis Algorithm
- MATH 327 Probability with Stat Methods or equivalent
- Know at least one programming language well

RELATIONSHIP WITH ABET

- Student Outcome 5 (Function effectively as a member or leader of a team engaged in activities appropriate to the program's discipline): All programming projects are required team projects of 2-3 students.
- Exposure to information management: This course is a designated course for this requirement.

TEXTBOOK AND REFERENCE BOOKS

Material in this class is delivered via lecture and reading research papers; there is no textbook.

COURSE FORMAT AND TOPICS

This class combines lectures with research paper reading and discussions.

The lectures will cover the fundamentals of data science on social media. The following is a non-exhaustive list of topics that will be covered in the lectures:

- What is Data Science and what does social media have to do with it?
- Data collection
- Social media data formats
- Social media data management with RDBMS/NoSQL
- Applications of probability and statistics, with an emphasis on hypothesis testing
- Applications of Machine Learning
- Visualization

The reading materials will be recent research papers that are related to the topics covered in the lectures. The main topics that will be covered in the reading materials will be:

- Dataset and data collection
- Algorithmic bias
- Inauthentic behaviors
- Ethics and data access
- Generative AI and social media

LECTURE NOTES AND SUPPLEMENTAL MATERIALS

- Lecture notes will be provided via PDFs or other formats delivered in class.
- All paper reading assignments will be made available via Brightspace.

ASSIGNMENTS

- Paper readings. The best way to start understanding what you can do with data science is to explore the state-of-the art. The best way to do that is to read research papers and that is what we will do in this class. There will be regular paper readings. For each paper, there will potentially be an in-class quiz. It is expected that all students come prepared (i.e., read the paper) and participate in the discussion to the best of their abilities.
- There will be three projects. Each project has three parts: a proposal, an implementation, and a report. Projects are to be completed in groups of 2-3 students.

Important notes about assignments:

- Late assignments may sometimes be accepted with penalty, which will typically be 5% per day late (including weekends and holidays). We will not accept assignments more than 5 days after the due date unless there is a very compelling reason.
- For project submissions, please ensure your programs are clear, well-structured, and easy to evaluate. All programming assignments should have:
  - An adequate explanation of the design of your program.
    - You should be prepared to answer good-faith, technical questions asked about your design and implementation during 1:1 sessions with the instructor.
  - Documented code:
    - Ideally, you use whatever documentation tools are available in the language you decide to implement in, but at minimum, all modules, classes, and functions should

have a documentation header that explains what the code does.

- Grading disputes, regrading and missing grades.
  - Should you dispute any grading, please be aware that we will not re-grade any single issue you have. Instead, your work will be re-evaluated from scratch. The new grade may be higher, lower, or stay the same. This new grade will not be changed.
  - No regrading can be requested two weeks after the date when graded work is returned to students.
  - The scores of your assignments will be made available to you after the assignments are graded.

METHOD OF ASSESSMENT

The following percentage weights will be used to assess student work:

- Paper reading quizzes are worth 15%.
- Three programming assignments (projects) 85% split evenly across all three projects.

GRADING DETERMINATION

Your final grade for this course is largely based on your performance relative to the performance of other students in the class. In other words, if your work is consistently better than average, you are likely to receive an A. The specific break down of grades is:

- A: 100–90
- B: 89–80
- C: 79–70
- D: 69–60
- F: 59–0

There are no +/- grades.

ACADEMIC HONESTY EXPECTATIONS AND VIOLATION PENALTY

- Cheating on quizzes of any kind, including, but not limited to, the use of electronic devices, “cheat sheets,” or looking at another student’s quiz are considered instances of cheating and will be reported as Category 1 academic dishonesty violation. You will also receive a one letter grade deduction (e.g., from A to B). More than one incident of cheating of any kind will result in an F for the entire course.
- The School of Computing at Binghamton wrote a letter to all computer science students about the importance of academic honesty. The letter is available at https://www.binghamton.edu/watson/about/academic-honesty.html.
- Please review the academic honesty document and make sure that you understand it!
- Each assignment must include the following statement, verbatim, followed by your group members’ names in a file called "HONESTY.md":

> "We have done this assignment completely on our own. We have not copied it, nor have we given our solution to anyone else. We understand that if we are involved in plagiarism or cheating, we will have to sign an official form that we have cheated and that this form will be stored in our official university records. We also understand that we will receive a grade of 0 for the involved assignment and our grade will be reduced by at least one level (e.g., from A to B) for our offense, and that we will receive a grade of "F" for the course for any additional offense of any kind."

Failure to submit your HONESTY.md file with the above text, verbatim, will result in your project not being graded and you receiving a 0 for the submission.

- For this course, programming assignments (projects) are all team projects. Certain open-source tools/software are permitted to be used (see the description of each project for details). Used open-source tools/software must be clearly acknowledged in the submitted project report.
- Additionally, each project submission must include a statement of contribution, which describes which group members did what part of the assignment in a file called "CREDITS.md". Failure to submit your CREDITS.md file will result in your submission not being graded and receiving a 0 for the submission.
- The use of generative AI tools is allowed for this course given that students follow the guidelines detailed below. Violation of the guidelines and inappropriate use of generative AI tools will be considered cheating and will be reported as Category 1 academic dishonesty violation. More than one incident of cheating of any kind will result in an F for the entire course.

GENERATIVE AI POLICY

Since this is not a generative AI course, we take a moderate stance on the use of generative AI tools: they are allowed but not encouraged. Students are encouraged to complete coursework independently to maximize learning outcomes. However, we recognize the potential benefits of generative AI tools in the learning process. Below are guidelines for students who choose to use these tools.

Generative AI tools include but are not limited to:

- LLMs, such as ChatGPT, Claude, and Gemini
- Coding assistants, such as Cursor, GitHub Copilot, and Claude Code.

Students are allowed to use generative AI tools to:

- Enhance their understanding of course material, such as gathering information and explaining concepts.
- Clarify ideas and polish their writing.
- Assist with project implementation.

Students should be aware that generative AI outputs can be erroneous, and they are fully responsible for verifying accuracy. Additionally, AI outputs may not meet assignment requirements. Since students are ultimately accountable for their submitted work, the quality will be reflected in their grade.

Students who choose to use generative AI tools must include an AI usage statement in their submitted assignments. The statement should include the following information:

- The generative AI tool used.
- The prompt used to generate the output.
- How the AI-generated content was integrated into the assignment.

The following are not allowed:

- Using generative AI tools to generate entire assignments without modification. Submitted work should reflect the student's own understanding, ideas, and effort.
- Any use of generative AI tools without properly acknowledging it in the AI usage statement.

MANAGING STRESS

If you are having any issues with personal or academic stress at any time during the semester, I encourage you to seek support. I do care about your wellbeing, and if my class becomes a pain point for you, you should feel free to reach out; I'm available to talk. Additionally, a wide range of campus resources are available to provide help, including:

- Dean of Students Office: 607-777-2804
- University Counseling Center: 607-777-2772
- Interpersonal Violence Prevention: 607-777-3062
- Office of International Student & Scholar Services: 607-777-2510

CLASS ATTENDANCE REQUIREMENT

Attendance is required and attendance will be checked regularly. If you are not present when attendance is checked, it will be counted as missing the class. Showing up late is considered missing the class.

COMMUNICATION

Students must use their Binghamton email address for all course communication. Emails from non-binghamton.edu domains will not receive a response.

All emails must include "[CS415]" or "[CS515]" in the subject line for proper identification. Emails without this subject line identifier may be ignored.

I will make every effort to respond to student emails within two business days. Please plan accordingly—emails sent shortly before deadlines may not receive timely responses.