RUTGERS
UNIVERSITY | CAMDEN
School of Business

Generative AI and Textual Analysis in Finance

Instructor: Prof. Tengfei Zhang
Office: BSB 258
Email: tengfei.zhang@rutgers.edu
Course Time: Tuesday at BSB 336 and Online (Synchronous), 6:00 pm – 8:50 pm
Office hours: 3:00 pm – 11:00 pm Tuesday (Office or Zoom)

Course Description and Objectives

Objective: This course aims to explore the application of textual content for understanding financial phenomena. The course offers a comprehensive and hands-on introduction to unstructured data mining in finance. Students will start with an introduction to R/Python. Next, students will learn how to extract financial information from various online sources, such as stock prices in Yahoo Finance, financial filings, earnings conference calls, media news, Tweets, corporate websites, and more, using existing APIs and web scraping tools. Then, using Generative Pre-training Transformer (GPT) as a copilot, students will learn state-of-the-art Natural Language Processing (NLP) and Large Language Model (LLM) methods, including bag-of-words, sentiment analysis, word embedding, topic modeling, and sentence embedding (BERT), and OpenAI's text embeddings and fine-tuning. The course will delve into multiple cutting-edge topics, such as how manager tone, media news, and analyst reports affect the financial market (e.g., stock prices and portfolio construction) and corporate finance (e.g., ESG and corporate culture). Students will gain the in-depth skills to prepare, process and interpret textual data and apply the programming techniques to solve real-world problems.

Learning outcome: Upon completing the course, students will obtain comprehensive knowledge of textual mining and NLP methods and grasp practical skills to collect, process, analyze, visualize, and interpret textual data in finance. By the end of the course, students will be equipped to apply programming techniques to address real-world financial challenges and leverage textual data to gain valuable insights for financial decision-making.

It is my intention to provide a collaborative and supportive learning environment where students will learn from one another both in and out of the classroom. To that end, **modifications to this syllabus might be warranted** as determined by the instructor as I assess the learning needs of this class.

Required Course Materials

Course GPT: https://chatgpt.com/g/g-Uh9x65TAw-finance-textual-analyst (all class materials are embedded in this GPT. Please feel free to ask any questions related to this course, seeing it as a teaching assistant or helper. You need to have ChatGPT Plus subscription to access this.)

No required textbook. Recommendations:
Analytics for Finance and Accounting: Data Structures and Applied AI, by Sean Cao, Wei

Jiang, and Lijun Lei.

Tidy Finance with R, by Christoph Scheuch, Stefan Voigt, and Patrick Weiss. Free available at https://www.tidy-finance.org/r/ and https://www.tidy-finance.org/python/

Text Mining with R: A Tidy Approach, by Julia Silge and David Robinson. Free available at https://www.tidytextmining.com/

R for Data Science (2nd edition), by Hadley Wickham, Mine Çetinkaya-Rundel, and Garrett Grolemund. Free available at
https://r4ds.hadley.nz/

The Future of Finance with ChatGPT and Power BI: Transform your trading, investing, and financial reporting with ChatGPT and Power BI, by James Bryant and Aloke Mukherjee (highly recommended if you want to use ChatGPT for trading and financial app/GPT development)

The Predictive Edge: Using Generative AI and ChatGPT in Financial Forecasting, Wiley, by Alejandro Lopez-Lira (general reading material)

Prerequisites

- Finance: You must have a basic understanding of the fundamental concepts in finance: the time-value of money; the relation between risk and return; the basic features and valuation of stocks and bonds. These topics are covered in 52:390:301 (a prerequisite course). I expect you to review the material from 52:390:301 as necessary.
- Computers/Software: Many of the examples in lectures and problem sets may require R/Python coding and ChatGPT (or a similar product). You are required to have a ChatGPT PLUS plan. You will need access to a computer and familiarity with Excel data analysis. I will assume that you know how to use spreadsheets to perform some basic analysis.

Communication

- Canvas: Reading assignments, homework, quizzes, and exams will be done through Canvas. Additional materials will be the syllabus, resources (articles and examples), Power Point slides, announcements, guides, etc. To access this system, go to http://canvas.rutgers.edu to log in, and click on the course on the dashboard.
- Rutgers email: USE YOUR RUTGERS EMAIL ADDRESS. All communications to students will be done using the Rutgers email address provided to you. Please forward your Rutgers email to your personal email if necessary. Not checking your Rutgers email is not an excuse for missing any communications.
- You must have a laptop/desktop available during our class time. We will have a lot of hand-on experiments.

Grading

The percentage that each component contributes to your final grade is:

Class participation: 15%

Class paper presentation and Q&A: 10% & 5%
AI frontier presentation and Q&A: 10% & 5%
Creating your own AI agent: 10%

Student project (writing): 10%
Project presentation: 20%
Project presentation Q&A: 15%

Bonus points (up to 10%): class participation and behavior expectations (up to 4%); midterm/final course evaluation (3%); attend course-related seminars and speaker series (2%); Bloomberg training certificate (1%); Generative AI online course certificates (2%).

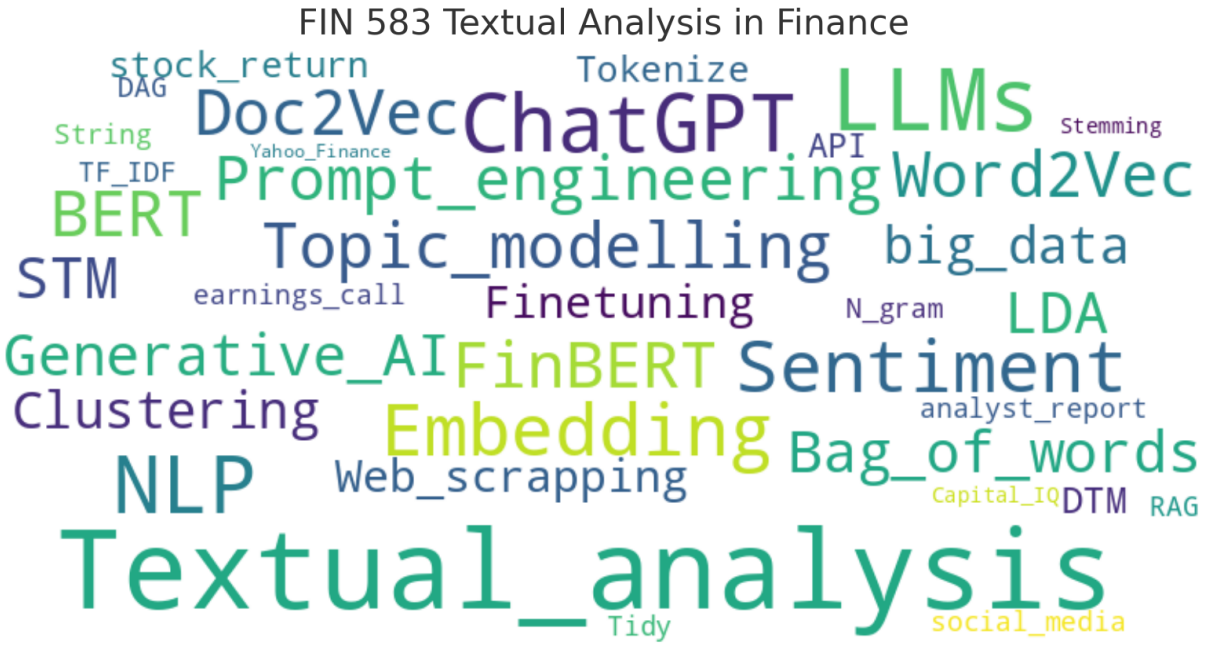
Grading Scale

89.5%-100%	A
84.5%-89.4%	B+
79.5%-84.4%	B
74.5%-79.4%	C+
69.5%-74.4%	C
60%-69.4%	D
Below 60%	F

Course outline and assigned readings*:

Week	Date	Day	Topic
1	9/02	Tu	Syllabus, class outline, course GPT, basic R coding
2	9/09	Tu	Generative AI introduction and prompt engineering
3	9/16	Tu	Generative AI applications in finance and GPT API coding and GPTStudio
4	9/23	Tu	AI agent
5	9/30	Tu	Stock return analysis with Tidy Finance; Data collection and web scrapping; Preprocessing, Bag-of-words, TF-IDF
6	10/07	Tu	Sentiment analysis with 10-K filings and annual reports
7	10/14	Tu	Word Embedding with earnings conference calls
8	10/21	Tu	Topic modelling with news and calls
9	10/28	Tu	Sentence embedding (FinBERT) with analyst reports
10	11/04	Tu	Document embedding using GPT text embedding with Glassdoor reviews; Retrieval-augmented generation (RAG)
11	11/11	Tu	GPT fine-tuning
12	11/18	Tu	GPT new topics (AI agent, multimodal learning, and AI-based hypothesis generation)
13	11/25	Tu	No Class (Thanksgiving break)
14	12/02	Tu	Student presentations
15	12/09	Tu	Student presentations
16	12/16	Tu	**FINAL EXAM** (student project due)

*This course schedule is tentative and subject to modifications under instructor's discretion on students' needs and progress.



Student Project Guidelines

Goal: Assume you are a financial analyst tasked with applying textual analysis in finance to predict stock returns or corporate performance. Students are free to choose any textual data, conduct coding, generate results, analyze these results, present their findings, and answer questions in class. Additionally, you will need to submit a written report.

Research Idea:
Feel free to develop your own research ideas. The only restriction is that your idea must be related to finance. If you don't have a specific idea, you may choose from the following list:

1. Using Glassdoor review data to measure firms' toxic culture (e.g., overtime work culture in investment banks) and then studying how toxic culture affects firm performance (e.g., profitability, employee turnover, financing, and stock valuation).
2. Using news (headlines and content) data to measure firms' culture changes (e.g., due to work-from-home policies) and then studying how these cultural changes affect firm performance.
3. Using Glassdoor review data to measure firms' ESG scores.
4. Using Reddit crypto subreddit data to predict daily/weekly crypto returns.
5. Using Reddit subreddit data to predict meme stock returns.
6. Using earnings conference call data to measure a firm's risk (e.g., AI risk, ESG risk, regulatory risk) and then predicting stock returns.
7. Using earnings conference call data to analyze topic and sentiment differences between executives and analysts, and then studying how these differences predict stock returns.
8. Using 10-K filings to extract the most important risk factors and then predicting stock returns.

Data: The data you use should include textual data for at least S&P 500 firms, assets, or stocks. The instructor will provide earnings conference call data, Glassdoor, RavePack news, Reddit, Compustat, and CRSP data. If you choose to use other datasets, you should be able to find them online (check GitHub, Hugging Face, and Kaggle) or web scrape them yourself.

Model: There are no restrictions on the textual analysis models you choose, as long as you can demonstrate that your selected model meets your research goals. Example models include bag of words, sentiment analysis, word embedding, sentence embedding, ChatGPT classifications, GPT embeddings, and fine-tuning.

Outputs:

1. Presentation and Q&A: A 15-minute presentation followed by a 5-minute Q&A session. Please pick one of the two dates as your presentation date in Canvas-People-Group.
2. Written Report: Include an introduction, literature review, hypothesis generation, data, model, results, discussion, conclusion, reference list, figures, tables, and appendix. The report should be no less than 1,000 words. You are required to upload your report to Canvas-Assignment. You are encouraged to use R Markdown to showcase your work.

Discussion: To keep track of your progress, please create a discussion thread on Canvas for your project. Feel free to ask questions there so that the instructor and other students can assist you.

1. Idea Generation Stage: Describe your idea, how to develop it, and conduct a literature

review.
2. Data Collection Stage: Describe the data you plan to use, provide details about the data, and explain why it is suitable for addressing your idea.
3. Model Selection Stage: Discuss a list of potential models you would like to try or use, and evaluate their advantages and weaknesses.
4. Coding Stage: Describe your coding work and any unresolved issues. Feel free to post coding questions or your R Markdown.
5. Result Analysis Stage: Briefly discuss your results, analyze whether they support your idea, and explore potential reasons if they do not.
6. Presentation and Q&A: You can request brainstorming questions here before your presentation.
7. Writing Report: Feel free to post your writing report so that other students can comment on and learn from your work.

Behavior Expectations

I expect you to treat the instructor and other students with respect and courtesy. I expect students to arrive on time and stay for the entire class.

It will be helpful if you read the assigned chapters and attempt to work homework problems before we discuss them in class. If you are unprepared your learning efficiency will be much lower and ultimately you will have to work harder to get thesame grade!

Students are required to turn off their phones (or turn to silence) while they are in class. Please inform me BEFORE CLASS if you are expecting an emergency call and must leave your phone turned on.

General/Administrative

Pronouns: This course affirms people of all gender expressions and gender identities. Feel free to correct me on your preferred gender pronoun. If you have any questions or concerns, please do not hesitate to contact me.

Chosen Name (Preferred Name): If you have a chosen name or preferred name other than what is listed on the roster, kindly let me know. If you would like to have your name changed within the rosters officially, go to:
https://deanofstudents.camden.rutgers.edu/chosen-name-application

Disability Services/Accommodations

The University is committed to supporting the learning of all students and faculty will provide accommodations as indicated in a Letter of Accommodation issued by the Office of Disability Services (ODS). If you have already registered with ODS and have your letter of accommodations, please share this with me early in the course. If you have or think you have a disability (learning, sensory, physical, chronic health, mental health or attentional), please contact https://success.camden.rutgers.edu/disability-services.

Accommodations will be provided only for students with a letter of accommodation from ODS. Their services are free and confidential. Letters only provide information about the accommodation, not about the disability or diagnosis.

Academic Integrity

The Academic Integrity policy can be found at http://studentconduct.rutgers.edu/student-conduct-processes/academic-integrity/

Students are responsible for understanding the principles of academic integrity and abiding by them in all aspects of their work at the University. Students are also encouraged to help educate fellow students about academic integrity and to bring all alleged violations of academic integrity they encounter to the attention of the appropriate authorities.

Academic Integrity means that you (the student) must:

- properly acknowledge and cite all use of the ideas, results, or words of others,
- properly acknowledge all contributors to a given piece of work,
- make sure that all work submitted as your own in a course activity is your own and not from someone else
- obtain all data or results by ethical means and report them accurately
- treat all other students fairly with no encouragement of academic dishonesty

Adherence to these principles is necessary in order to ensure that:

- everyone is given proper credit for his or her ideas, words, results, and other scholarly accomplishments
- all student work is fairly evaluated and no student has an inappropriate advantage over others
- the academic and ethical development of all students is fostered
- the reputation of the University for integrity is maintained and enhanced.

Failure to uphold these principles of academic integrity threatens both the reputation of the University and the value of the degrees awarded to its students. Every member of the University community therefore bears a responsibility for ensuring that the highest standards of academic integrity are upheld. Violations are taken seriously and will be handled according to University policy.

Code of Student Conduct

Rutgers University seeks a community that is free from violence, threats, and intimidation; is respectful of the rights, opportunities, and welfare of students, faculty, staff, and guests of the University; and does not threaten the physical or mental health or safety of members of the University community, including in classroom space.

As a student at the University you are expected adhere to the Code of Student Conduct.
To review the code, go to the Office of Community Standards:
https://deanofstudents.camden.rutgers.edu/student-conduct