Syllabus for “Text as Data”

Instructor

Harm H. Schütt
Professor of Financial Accounting

Email

harm.schuett@whu.edu

Course dates

April 7th, 9th, 10th, 11th, 2025

Course location

7th: K-001
9th: C-101
10th: D-101
11th: E-103

Course hours

09:45 – 17:45

(4 blocks of 90min each per day, We might not need all blocks)

Course Overview

The course is aimed at doctoral students and teaches the current state of the art in drawing inferences from data—specifically, text data.

A fundamental problem in empirical research is measurement. Complex and difficult-to-define concepts are often components of important theories we want to test. Think of investor sentiment in behavioral trading models or economic uncertainty and capital constraints in theories of firm investment cycles.

The lack of appropriate measures stymied sufficient advances in theory testing for many years. Then, around 2008, a series of papers in Economics and Finance pioneered the idea that texts can inform about such important concepts. For example, Tetlock (2008) measured investor sentiment as the number of negative-tone words in a daily Wall Street Journal market commentary column. These papers sparked a textual analysis revolution that finally enabled significant developments in core theories. Today, textual analysis techniques are invaluable in every applied empiricist’s toolbelt.

This course aims to equip you with such a tool belt. It introduces you to a modern treatment of drawing inferences from text data. It introduces a framework for data analysis in general and one for measuring concepts from text specifically.

The course is roughly divided into three parts. The first, shorter, part introduces a modern framework for drawing inferences from data. This part introduces basic concepts. It also shows how to use graphs (DAGs) to derive research designs.

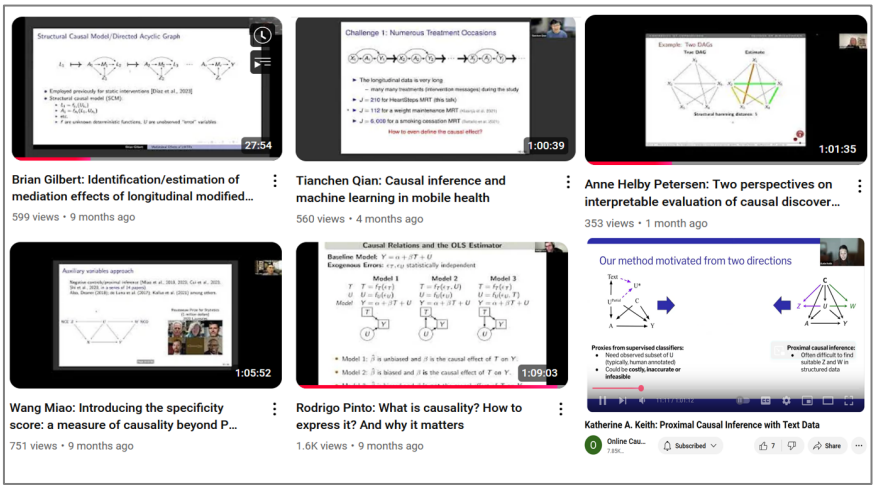

Figure 1: Stanford’s online causal inference seminar talks: DAGs describe assumed variable relations

The second part introduces textual analysis using a framework that divides textual analysis (or any measure generation) into two connected steps: quantification and mapping

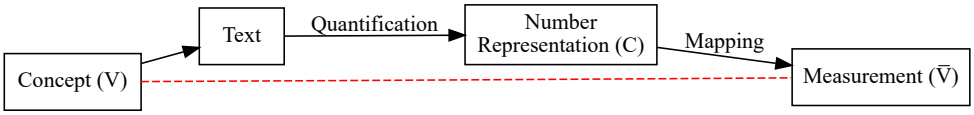

Figure 2: Text as data framework (Gentzkow, Kelly, Taddy, 2019)

Quantification concerns quantifying a text into machine-readable form, such as the bag-of-words representation. Mapping encompasses methods, such as word lists and supervised or unsupervised methods, that turn numerical representations into the measure of interest.

You will be introduced to commonly applied approaches for both steps**.** You will learn which approaches are advisable given a specific text and the concept to be measured. We will see multiple examples of how the concept to be measured influences certain texts and suggests specific quantification and mapping steps.

The third part introduces GenAI as a tool to create measures of important concepts. The advent of generative AI applications has generated new possibilities for deriving concept representations. Yet, it also poses several challenges. We will discuss and test-run the use of GenAI on several examples.

Learning goals

1. **Develop a Comprehensive Understanding of Inference and Measurement in Empirical Research.** Students will learn to critically evaluate the theoretical underpinnings of causal inference and prediction, utilizing Directed Acyclic Graphs (DAGs) to visualize and reason through complex relationships among variables, confounders, moderators, and mediators. Students will demonstrate proficiency in diagnosing measurement errors and their impacts on empirical identification and inference.
2. **Master Advanced Textual Analysis Techniques for Empirical Measurement.** Students will acquire the skills necessary to quantify textual data into numerical representations and effectively map these into meaningful empirical constructs. They will proficiently apply and critically evaluate methodologies such as dictionary-based approaches, document similarity measures, supervised and unsupervised machine learning techniques, and pre-trained language models like BERT to empirically measure complex concepts such as sentiment, economic uncertainty, and competition.
3. **Critically Assess and Implement LLM Methods in Textual Analysis.** Students will gain expertise in leveraging LLMs as an advanced methodological tool for textual analysis. They will analyze and articulate the methodological strengths, limitations, and potential biases inherent in LLM-based measurement techniques, demonstrating the capability to apply LMM responsibly and effectively to quantify complex concepts in novel research contexts.

Preparation before class starts

Do three things before class. First, browse the two papers cited in the references at the end of the syllabus. You do not have to analyze every statement made there. I want you to have heard the terms and be familiar with some ideas before class. Second, sign up on GitHub and install the software (see below). Third, are you a complete Python beginner? If so, see whether you can still take a quick Datacamp beginner course. It is not required but can help you a lot.

Teaching Mode

My goal is to make this course as fun and useful to you as I can. For every method we will discuss, I will show you several examples of important and interesting concepts that were measured using this method. My hope is that this will help you think about the concepts important in your own research and ideas on how to measure them. The course also involves quite a bit of in-class exercises. The course is roughly divided into two-thirds lectures and one-third practicing the methods under my guidance. As a result, the number of seats is limited. You can do the exercises alone or in groups. The exercises are not graded; they are simply there to foster learning. Playing around with real code and experiencing how sensitive certain measures are to design choices is an invaluable part of the learning experience.

During the sessions, I will provide Python code, and you will be asked to apply the methods and coding patterns you learned in selected exercise sessions and answer certain questions. To make this work, you must bring your laptop with the required software installed and ready to go. I have collected the installation instructions, all data, exercises, and code in a private GitHub repository. Send me an email with your GitHub username a few days before class starts so that you get access to the material and know what to install. *Important:* I will not have the time to troubleshoot installation problems during class. If there are errors, it is a good exercise to try to solve the problem using GenAI, stackoverflow.com, or the package documentation. If you cannot get the software to run after honest effort, please get in touch with me before the course starts.

We all want the sessions to be interesting and enjoyable, with lots of discussion. Therefore, feel free always to ask clarifying questions. This is a methods course, so the onus is to ensure you understand and digest the ideas and can apply the methods. We will look at many use cases, and I will ask you to critique methods and critically discuss the limitations of inference and possible extensions.

Required Level of Python Skills

Given the time we have, I cannot give a full introduction to everything Python can do. But I will introduce everything you need to know to apply textual analysis methods. Learning and internalizing these methods takes practice! The course is designed to give you code and pointers to help you practice and apply the learned material to your own problems.

If you are new to coding, I highly encourage you to take one of the many excellent free introduction tutorials found online (see, for instance, the Datacamp course catalog). You will likely have a much smoother ride if you invest three hours into such an online tutorial beforehand.

Course Policies

Grading Policy

This is a pass/fail course. Passing depends on your participation and the deliverable. I require you to be present on all days to pass the course successfully. I also require a well-meant effort to engage with the material and class. The deliverable is a short 3-page research proposal. The proposal consists of three parts: the first contains a research question, motivation, and expected contribution to the prior literature part (1-3 paragraphs). The second outlines the theory to be tested with accompanying DAG. The third part describes a research design derived from the DAG and a description of how core concepts are measured using some of the methods we discussed in class.

E-mail Policy

You can always write me an e-mail; I am sometimes a bit slow to respond, but I will respond. Generally, if you have an administrative question, please look at the syllabus first. Emails with questions that should be clear from just looking at the syllabus will be a very low priority.

Tentative Course Schedule

	Day	Subject	Exercises
A modern framework for drawing conclusions from data	Day1	Statistical Inference in 2025 - Analyzing is generalizing and comparing - Causal inference versus prediction - Trade-offs: Expressive models versus black box models	
		Research design principles - Reasoning about identification: From theory to DAGs to empirical design - The building blocks of DAGS	Simulation exercises: Confounders, Moderators, Mediators, Colliders, and fixed effects
		The importance of measurement for identification - Types of measurement error - Signal-to-noise - Example: Identifying unobserved confounders from text	Simulation exercises: Correlated measurement error
		Quantifying uncertainty - Prediction errors versus standard errors - Why and when to cluster standard errors - Example: LLM-generated variables and standard errors	i.t.
Classical Textual Analysis	Day 2	Measuring concepts with text data - Sentiment, economic uncertainty, and competition – The textual analysis revolution in Economics, Finance, and Accounting - The Gentzkow, Kelly, Taddy (2019) Framework	
		Parsing text data - Introduction to parsing using Python - Parsing using GenAI tools	Preparing and structuring financial documents Extracting number references in conference calls
		Turning text into numbers - Pre-processing text - Bag-of-words representation - Introduction to Spacy NLP tools - Introduction to the SKlearn ML library - Word-embeddings representation - Doc2Vec - Finding informative features - best-practices	

	Day	Subject	Exercises
	Day 3	Measuring concepts by word classification – Dictionary approaches - Mapping word counts to concepts - When does it work well? - Designing word lists	Computing sentiment scores Exploring token occurrences
		Measuring concepts by document similarity – Cosine similarity - Similarity as a powerful analogy concept - Cosine similarity from bag-of-words	Measuring business model similarity
		Measuring concepts by document classification – Supervised approaches - Prediction approaches based on training data - Naïve Bayes, penalized regressions, and gradient boosting - Multilabel and Multiclass problems	Classifying hate speech Named-entity recognition
		Measuring concepts with pre-trained models – BERT, FINBERT, all the BERTs - BERT - Example: Sautner et. al (2023) "Firm-Level Climate Change Exposure"	i.t.
		Measuring concepts by document classification – Unsupervised approaches - Clustering approaches: One topic per document - Model-based approaches and Latent Dirichlet Allocation: multiple topics per document - Introduction to the genism topic modeling library	Explore trends in scientific articles
GenAI Textual Analysis	Day 4	Measuring concepts using GenAI - GenAI is an n-gram prediction model - GenAI sentiment - Scoring and confidence scores - Text parsing - Reasoning: GenAI and FSA - Numerical reasoning and lookahead bias – Current issues with GenAI applications	Measuring litigation risk, Sentiment, and disagreement Causes and determinants of cultural values

References

Gentzkow, Matthew, Bryan T. Kelly, and Matt Taddy. (2019) "Text as Data." *Journal of Economic Literature* 57(3).

Dikolli, Shane S., Thomas Keusch, William J. Mayew, and Thomas D. Steffen (2020) "CEO behavioral integrity, auditor responses, and firm outcomes." *The Accounting Review* 95(2): 61-88.