Decoding Text Complexity Classical non-Machine Learning methods and AI-Driven Text Analytics in Higher Education

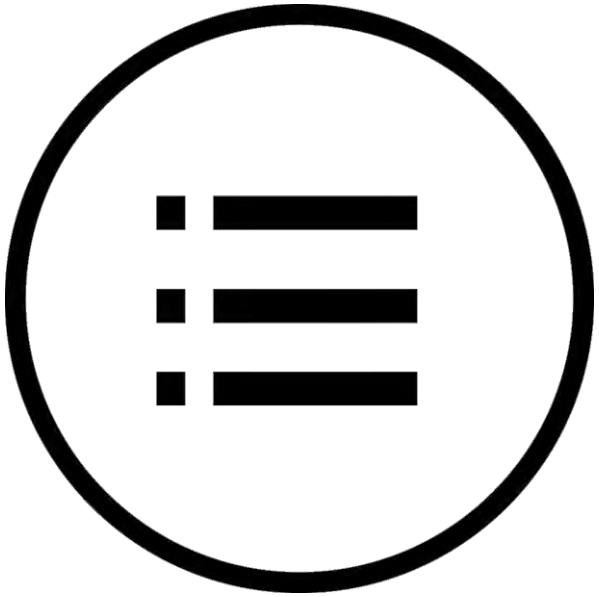
Reynaldo Quiroz and Lin Yao

UNT | DATA, ANALYTICS, & INSTITUTIONAL RESEARCH

February 26th, 2026

Overview

1. Introduction
2. Review of approaches
3. Comparison of approaches
4. Conclusions



Introduction

- Higher education institutions generate large amount of unstructured data:
  - From course evaluations and advising notes to online discussions, campus surveys, and even Reddit conversations.
- Text analytics is opening a new frontier for institutional research:
  - Simple rule-based methods
  - Advanced machine learning models
  - Powerful AI language models
- Text analytics transforms unstructured text into strategic insights

Some algorithm challenges with sentiment analysis

Sarcasm/irony

“Not terrible”, “I expected worse”

Algorithm: might flag “terrible” and “worse” as negative score.

Mixed sentiment/context clues

“Brutal exam, but fair instructor”

Algorithm: might weigh “brutal” as negative and “fair” as positive score.

Negation

“Not terrible”, “Some exams were not bad”

Algorithm: might fixate on “terrible” and “bad” as negative score.

Idioms/cultural nuances

“Barking up the wrong tree,” “Break a leg”— *American culture*

Algorithm: without training, “wrong” and “break” might be negatively scored.

Steps for classical and machine learning text analysis

- Data definition: (task and output)
- Text preprocessing (All methods):
  - Cleaning -> remove special characters
  - Tokenization -> convert text into smaller units (words or sentences)
  - Normalization -> case folding (lowercasing) and convert numbers to words.
  - Stopword removal -> remove words that don't add meaning ("a", "the"…)
  - Stemming/lemmatization -> Reduce words to their base/root form ("running" to "run")
  - Feature extraction -> convert text into numerical representations
- Model training and evaluation (Only ML models): train the model on labeled data and evaluate using metrics.

Classical non-ML approaches

- Tokenization and keyword extraction

- *CSCE 1030*: ['lie', 'started', 'rough', 'pulled', 'health', 'issues', 'mid', 'semester', 'fell', 'surprisingly', 'material', 'low', 'key', 'fun', 'got', 'hang']
- *ECON 1100*: ['lie', 'started', 'rough', 'pulled', 'honestly', 'life', 'just', 'kept', 'life', 'ing', 'couldn't', 'surprisingly', 'material', 'low', 'key', 'fun', 'got', 'hang']
- *PSYC 1630*: ['didn't', 'expect', 'turned', 'quick', 'room', 'looked', 'lost', 'half', 'time', 'walked', 'final', 'feeling', 'confident']

- Term statistics or word frequency

term	count	term	count
honestly	1383	instructor	960
life	1324	felt	824
week	1241	clicked	809
rough	1223	chaos	804
started	1217	exams	786

- N-grams (TF–IDF)

ngram	mean_tfidf
life	0.024792
week	0.023650
fighting life week	0.016666
fighting life	0.016666
life week	0.016666
fighting	0.016666

- Probabilistic models such as topic modeling

topic	top_words
0	exams, different, absolutely, built, brutal, w...
1	quick, didn, turned, expect, ended, surprised,...
2	entire, fast, miss, lecture, blink, concept, p...
3	actually, instructor, rough, things, taught, w...
4	life, honestly, week, survive, did, fell, seme...
5	just, chaos, standard, pretty, experience, ter...,

Challenges of Classical non-ML approaches

- Limited ability to understand context (fixed rules or dictionaries)
- High vocabulary sensitivity (misspell, abbreviations, slang, emojis)
- High manual maintenance (hard to manage at scale)
- No learning from data (do not improve over time)
- Struggle with higher education complexity (disparities in language)
- Domain specificity (hard to transfer across contexts)
- Cannot produce abstractive summaries

ML approaches

Supervised learning (SL):

- Naive Bayes
- Support vector machines (SVM)
- Random forest
- Probabilistic models such as topic modeling
- Lexicon-based sentiment

Unsupervised learning (UL):

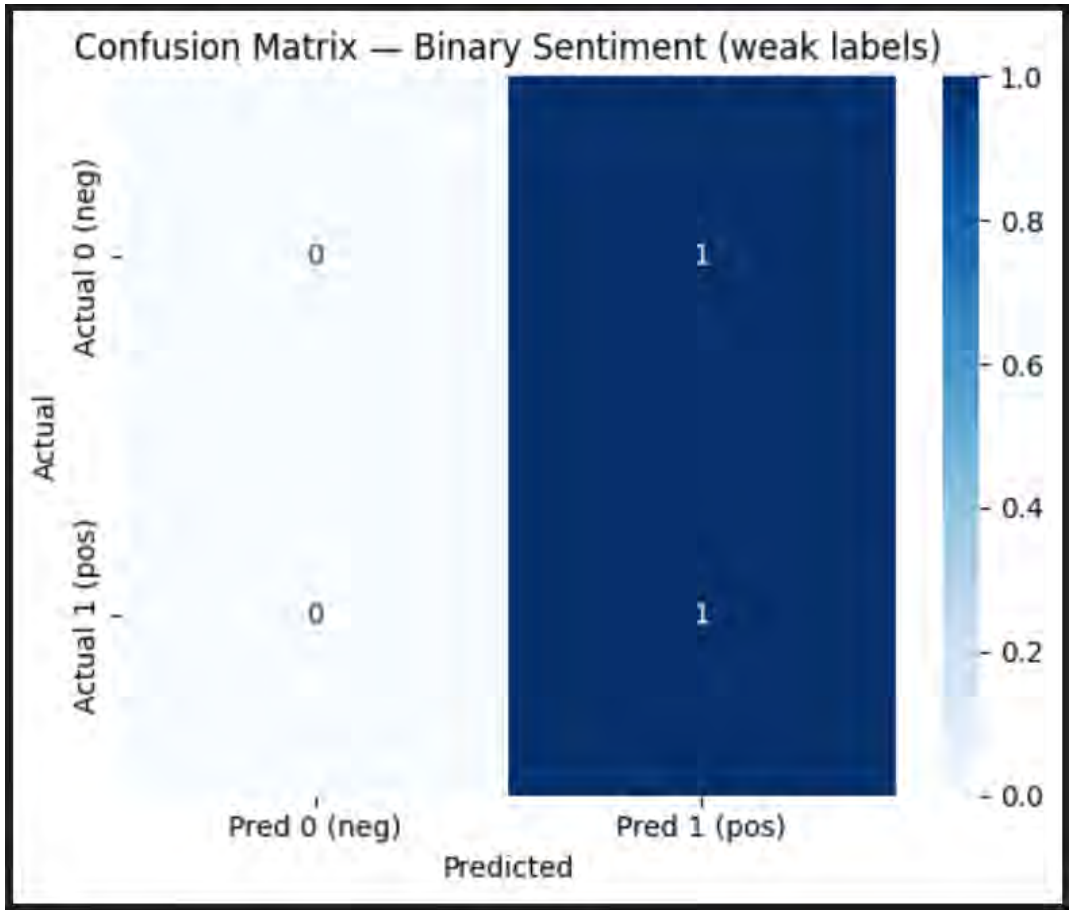
- Topic modeling
- K-means clustering

Semi-supervised learning (SSL):

- Self-trained models
- Generative models

Confusion matrix

Predicted \ Actual	Positive	Negative
Positive	True Positive (TP)	False Positive (FP) Type 1 error
Negative	False Negative (FN) Type 2 error	True Negative (TN)



	Pred 0 (neg)	Pred 1 (pos)
Actual 0 (neg)	0	1
Actual 1 (pos)	0	1

Evaluation Metrics

Accuracy: overall correctness

$$\frac{(TP + TN)}{(TP + FP + FN + TN)}$$

- *Useful when classes are balanced*
- *Can mislead under imbalance*

$$\frac{0+1}{2} = 50\%$$

Precision: when we predict Positive, how often are we correct?

$$\frac{TP}{(TP + FP)}$$

$$\frac{1}{1+1} = 50\%$$

Recall: out of all actual Positives, how many did we find?

$$\frac{TP}{(TP + FN)}$$

- *High recall*

$$\frac{1}{1+0} = 100\%$$

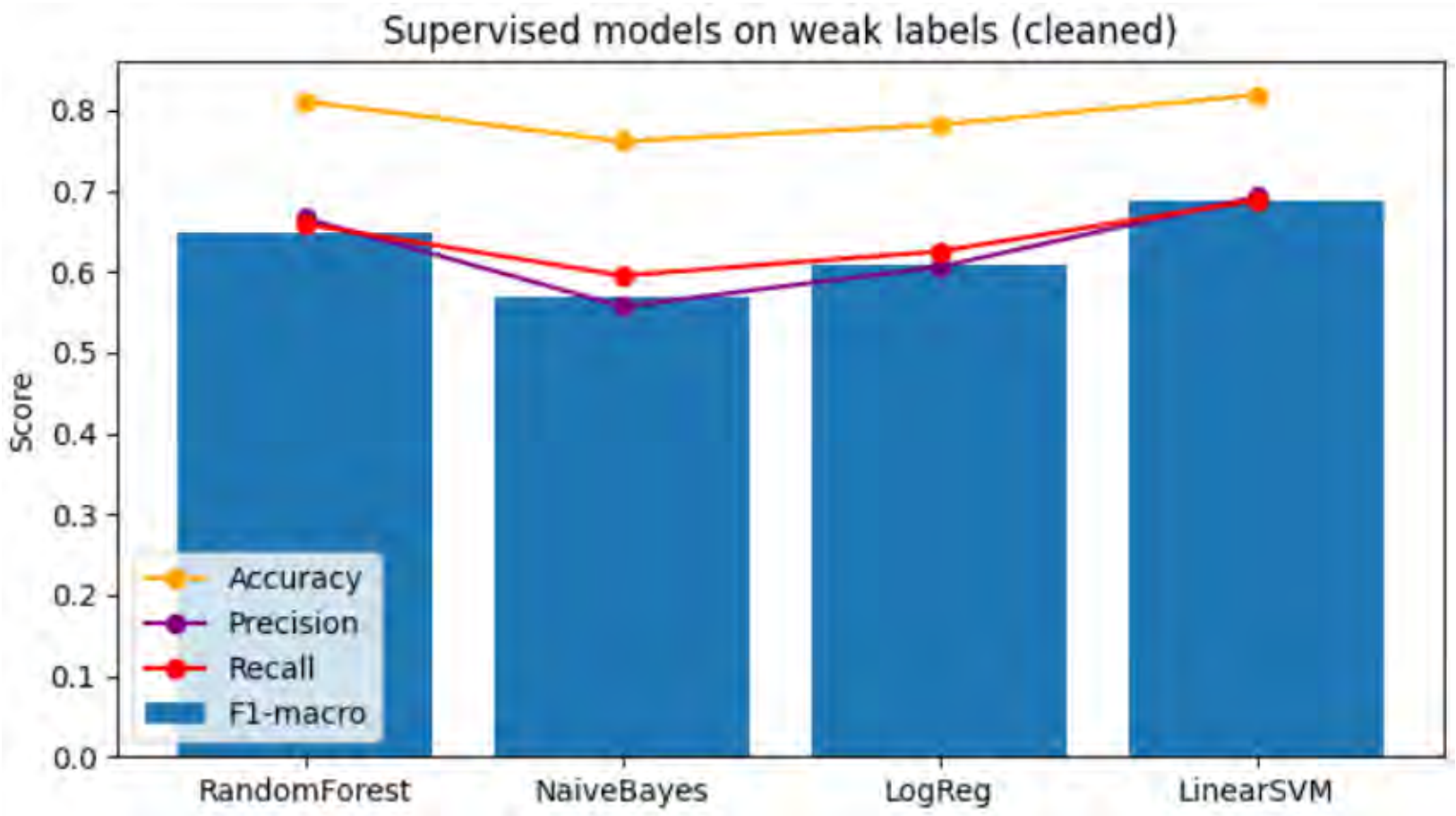
F1-score: Harmonic mean of Precision and Recall

$$\frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

- *Balances FP and FN*

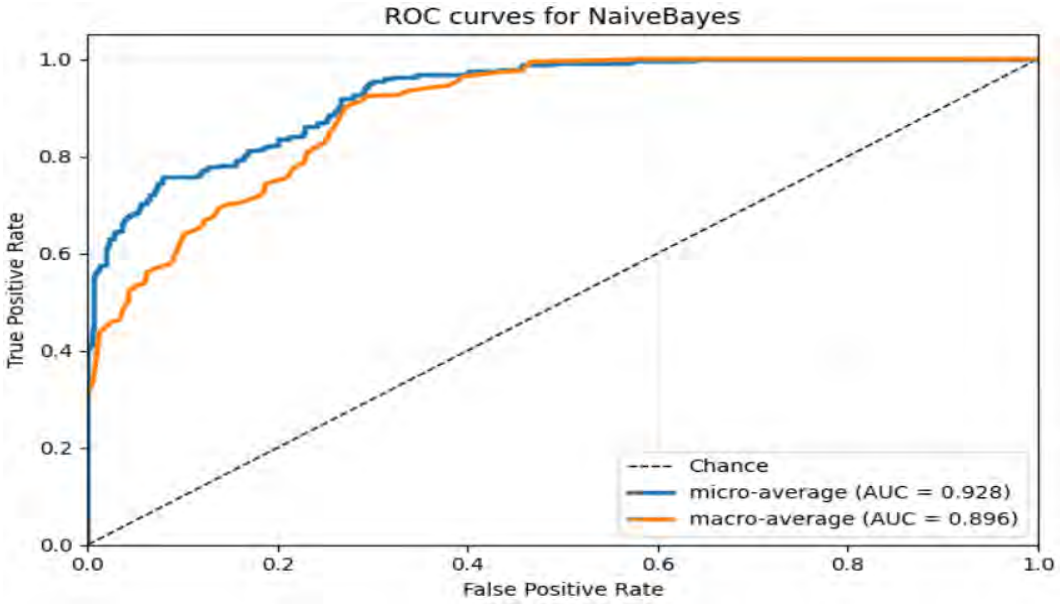
$$\frac{2*(0.5*1)}{(0.5 + 1)} = 67\%$$

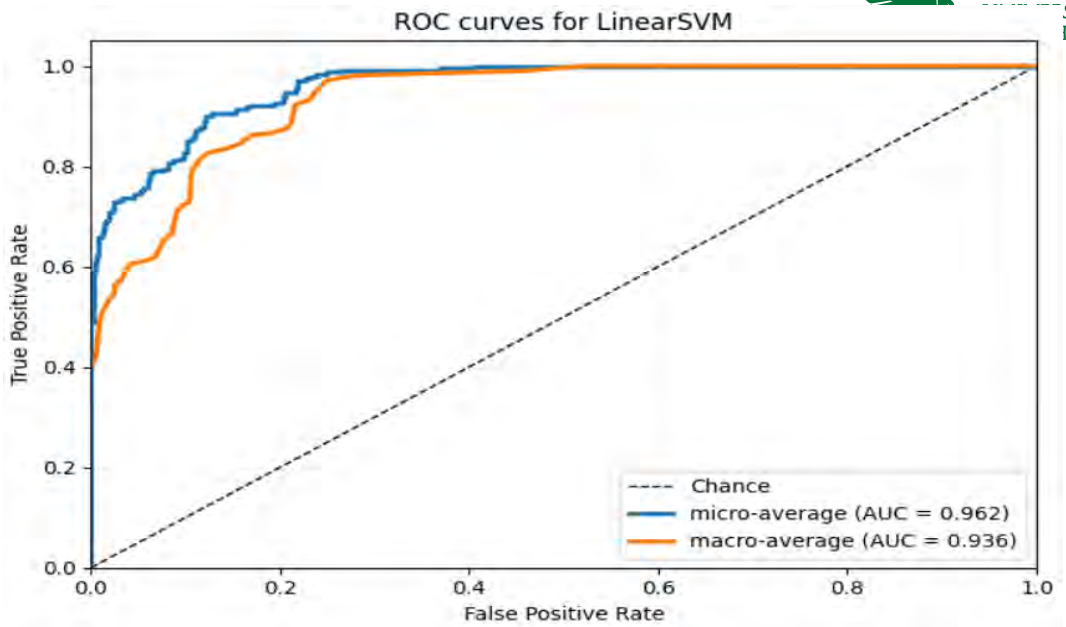
Supervised model results

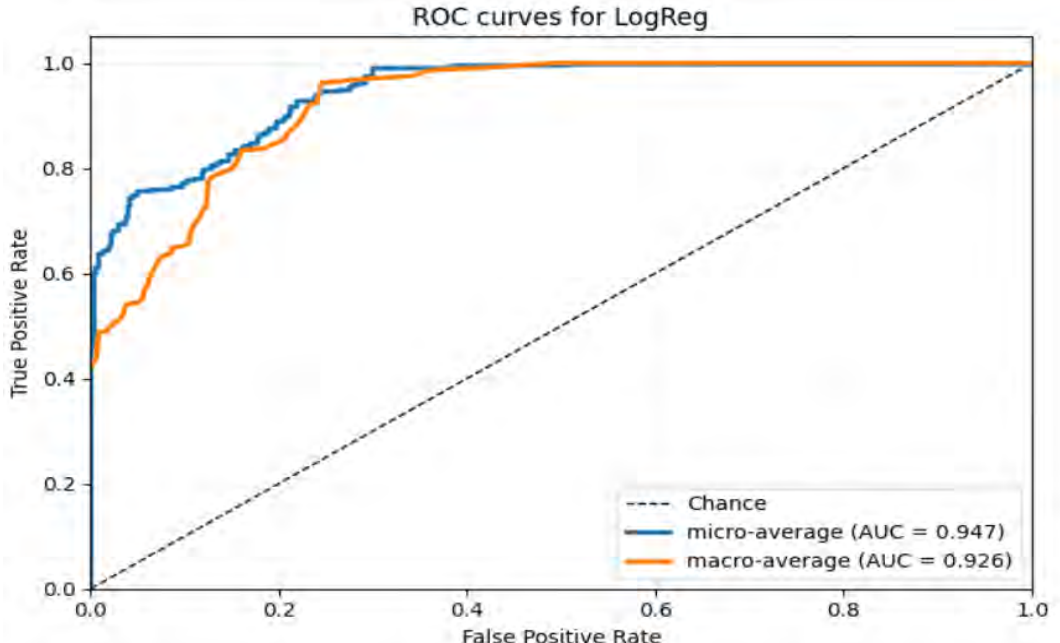


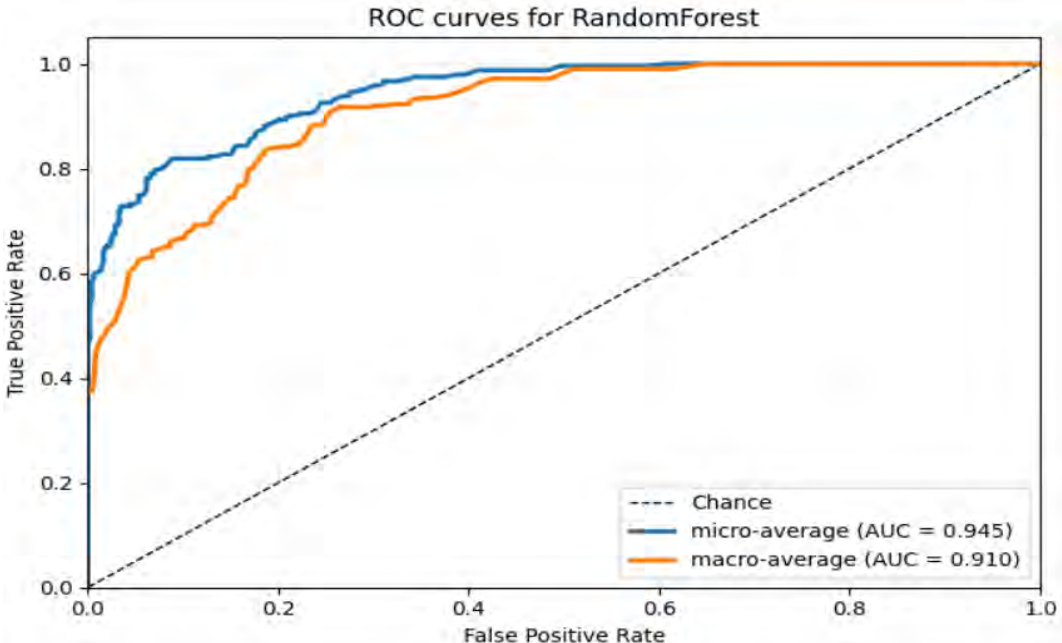
model	accuracy	precision_macro	recall_macro	f1_macro
RandomForest	0.810700	0.666098	0.658876	0.648306
NaiveBayes	0.761317	0.556954	0.595312	0.567771
LogReg	0.781893	0.606100	0.625214	0.609308
LinearSVM	0.818930	0.692925	0.688741	0.688365

Supervised model ROC curves



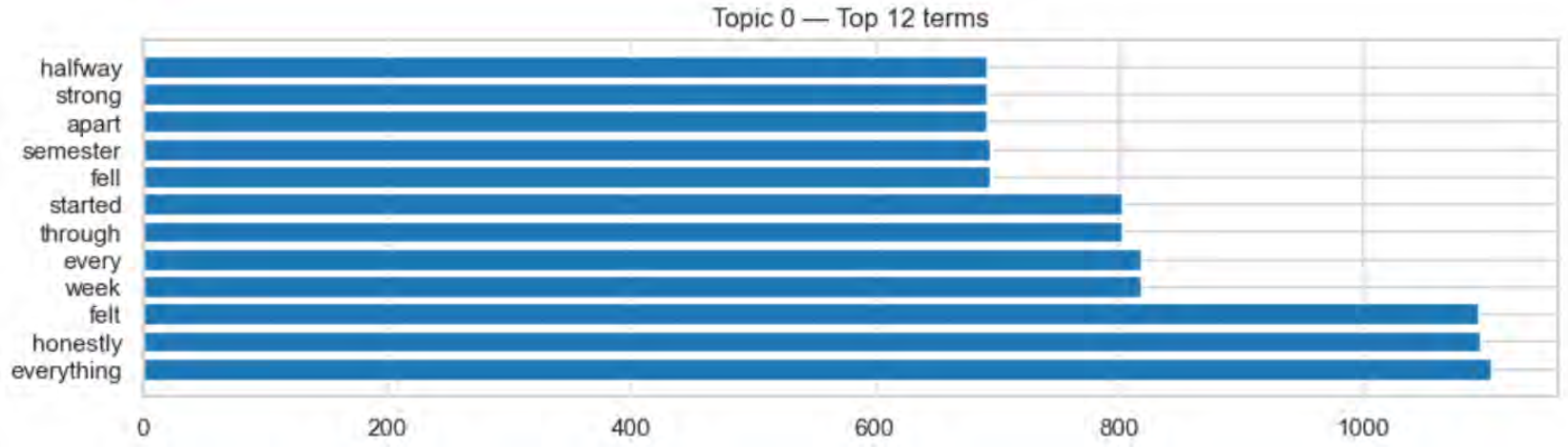


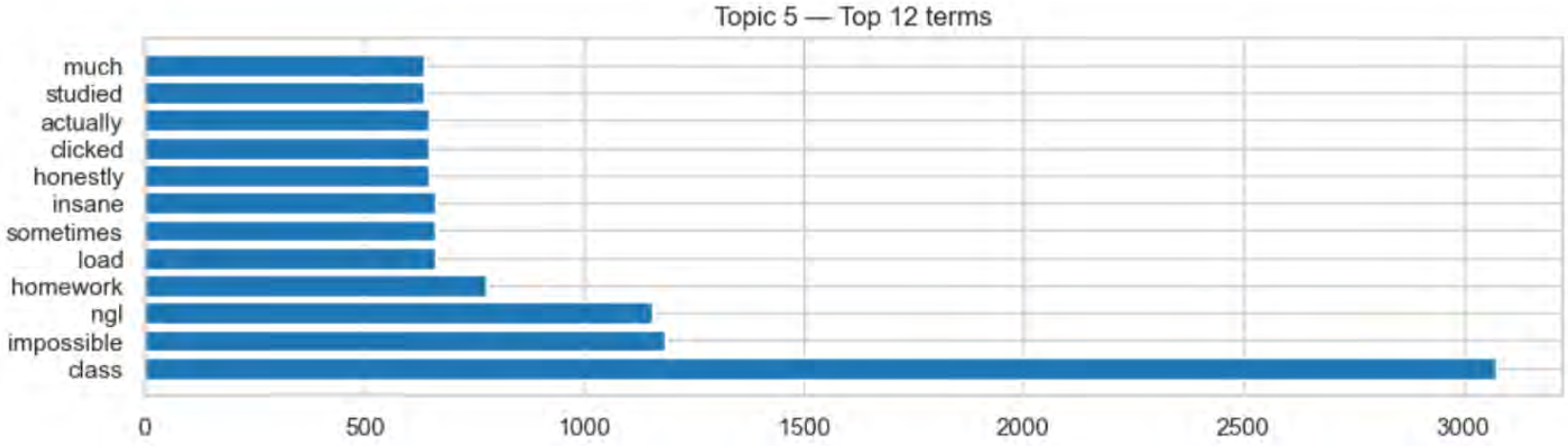




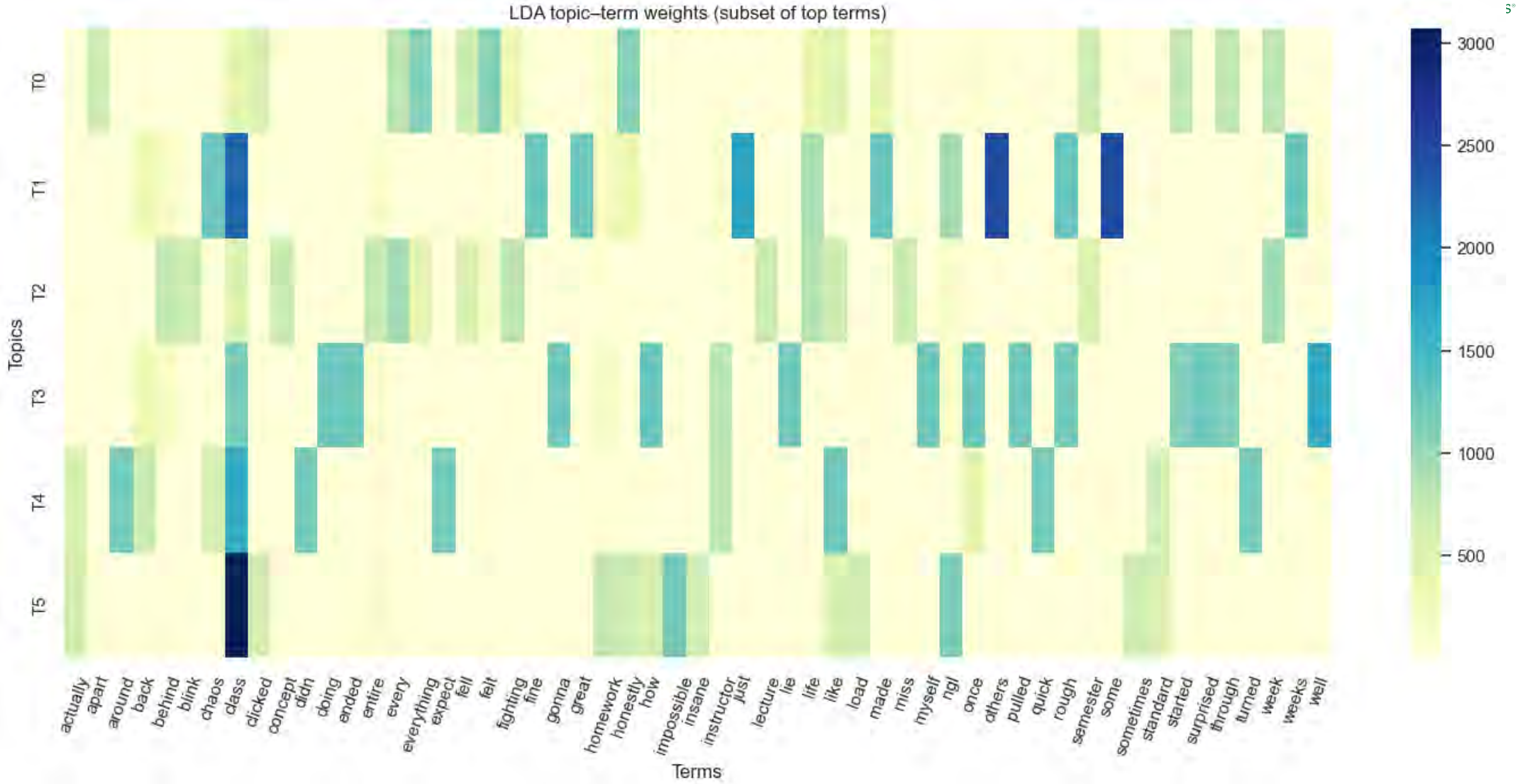
Unsupervised approach – List of six top topics and terms

```
Topic 0: everything, honestly, felt, week, every, through, started, fell, semester, apart, strong, halfway
Topic 1: some, others, class, just, rough, made, fine, weeks, chaos, great, terrible, total
Topic 2: week, every, life, fighting, behind, entire, concept, lecture, blink, miss, fast, pace
Topic 3: well, how, once, surprised, myself, doing, ended, lie, pulled, gonna, rough, class
Topic 4: class, like, didn, quick, expect, around, turned, instructor, back, standard, experience, pretty
Topic 5: class, impossible, ngl, homework, load, sometimes, insane, honestly, clicked, actually, studied, much
```

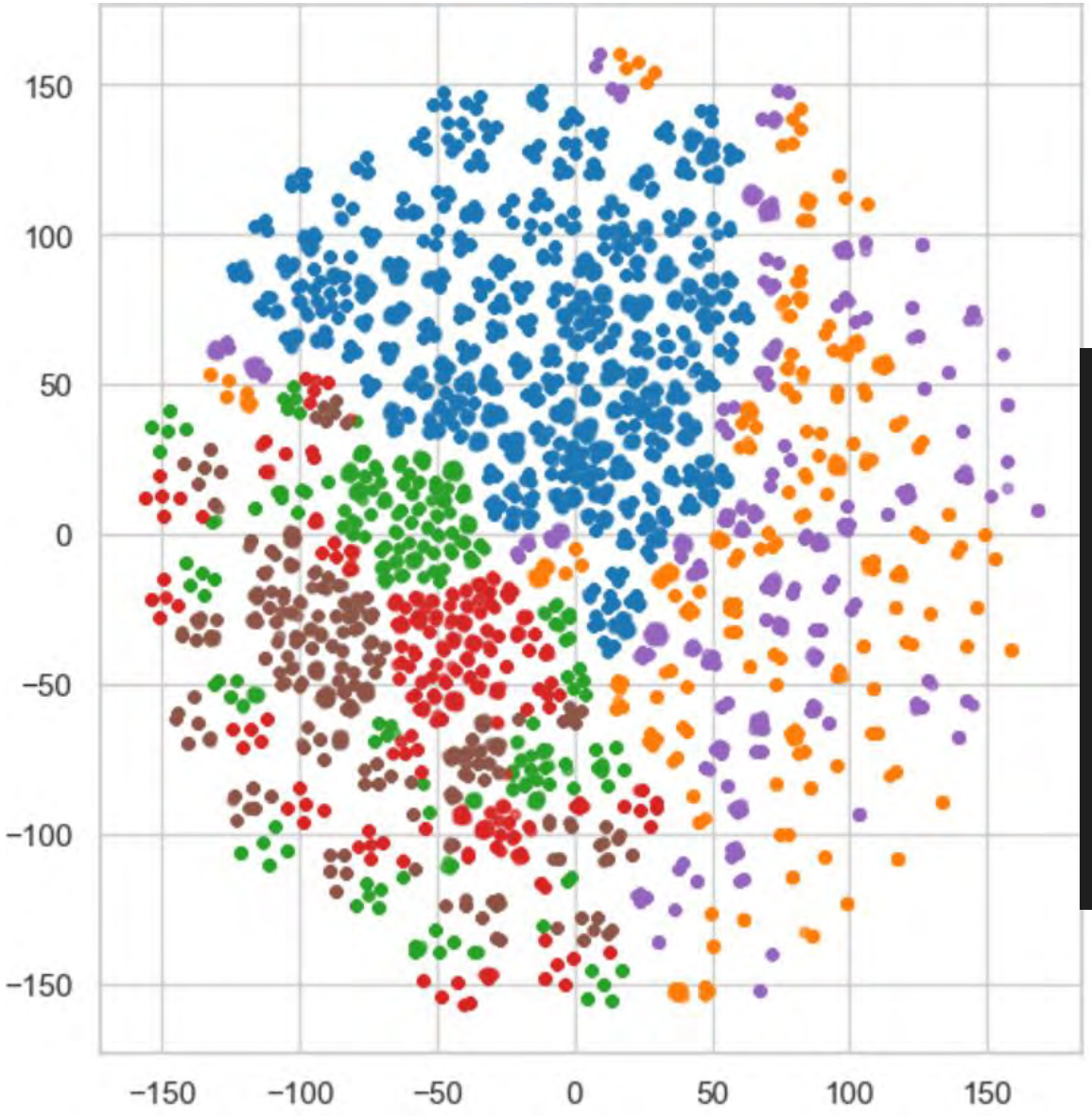




Unsupervised approach – Topic modeling



Unsupervised approach - K-mean clusters



Cluster 0: week, every, every week, life, fighting, fighting life, life every, completely, wrecked, class completely, completely wrecked, did survive
Cluster 1: some, others, fine others, fine, weeks fine, some weeks, others rough, weeks, total chaos, sense others, made sense, total
Cluster 2: rough pulled, pulled through, pulled, lie class, lie, started rough, class started, gonna, gonna lie, started, through, rough
Cluster 3: doing, well ended, ended doing, ended, myself, myself how, how well, surprised myself, surprised, well, how, once
Cluster 4: class, terrible great, terrible, just class, great, great just, rollercoaster, rollercoaster ngl, class rollercoaster, just, held, together
Cluster 5: quick, around, expect like, expect, turned around, turned, class turned, didn, didn expect, like class, around quick, like

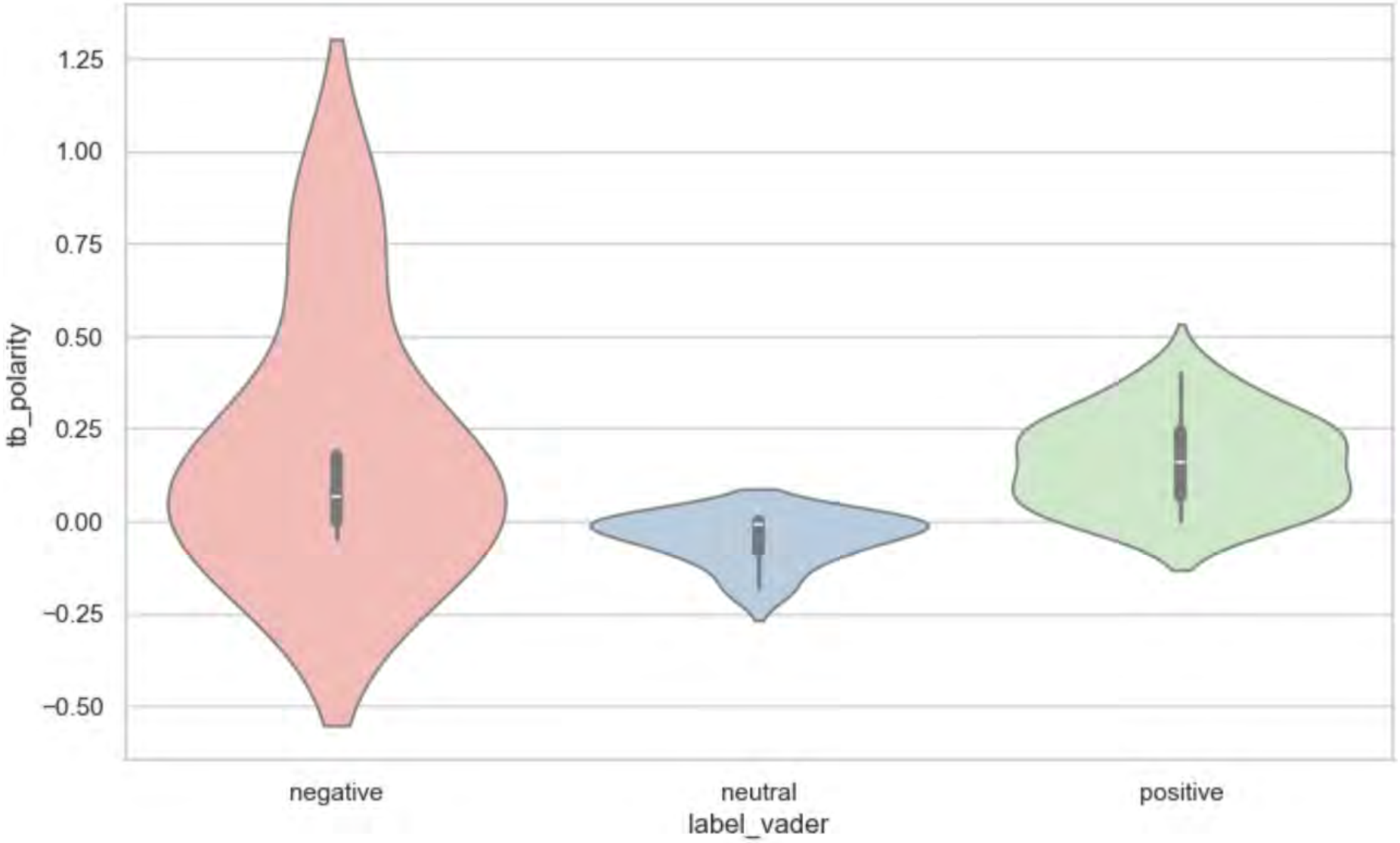
Deep Learning approach (VADER vs TextBlob)

We compute VADER compound scores (−1 to +1) and TextBlob polarity (−1 to +1).
We then derive **labels** using thresholds:

- **Positive:** score ≥ 0.05
- **Negative:** score ≤ −0.05
- **Neutral:** otherwise

	text_orig	vader_compound	tb_polarity	label_vader	label_tb
0	University can be the worst four years of your...	0.1007	0.000000	positive	neutral
1	Turn that frown upside down	0.0000	0.000000	neutral	neutral
2	It varies from person to person, but in genera...	0.9030	0.111548	positive	positive
3	tbh tho, i kind of need 90% at this point to s...	0.4939	0.203333	positive	positive
4	it does if you want to do cs	0.0772	0.000000	positive	neutral
5	So abandon cs and go into something else?	-0.4927	0.000000	negative	neutral
6	I think there are two main reasons why people ...	0.7603	0.041190	positive	neutral
7	I think two of the most important things we sh...	0.9797	0.239444	positive	positive

TextBlob polarity vs VADER



Challenges of ML approaches

- Requires labeled training data (comments are long and nuanced)
- Sensitive to imbalanced classes (comments could be more positive or negative, depending on the domain)
- Limited contextual understanding (struggle with long narratives)
- Transferability is limited (advising notes cannot work on course evaluations)
- Hard to interpret classical rules (why this is classified as negative)

AI/LLMs approaches

- Deep contextual understanding (interpret sarcasm)
- Zero-shot and few-shot classification (LLMs can perform text classification based on prompts)
- Abstractive summarization (understand and rephrase concepts)
- Automatic topic discovery and clustering
- Multilingual and code-switching support (they can process Spanglish)
- High flexibility across domains (generalize well across domains)
- Ability to combine multiple tasks (sentiment analysis and summarization)

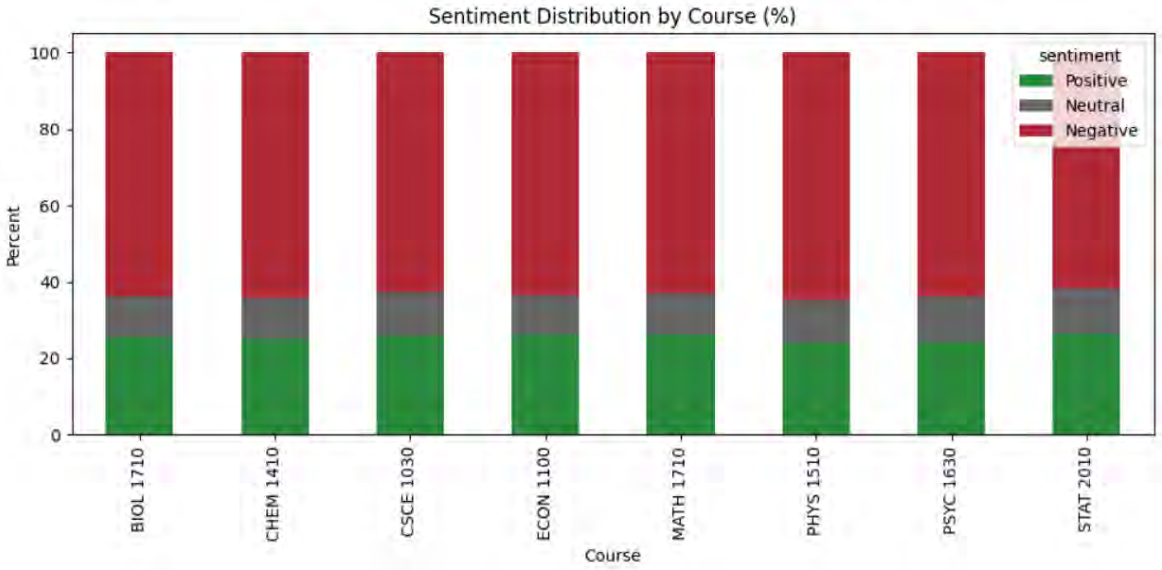
AI Agent Sentiment Analysis and Text Summarization Example

- AI agent tool used: Gemini
- Example analysis file: Course Evaluation – How hard is a class. 10,564 responses. Took 15 seconds for Gemini to finish the analysis.
- Analysis prompt: Please provide sentiment analysis and text summarization by group categories using classical non-ML and ML models and compare which model performs better.

Feature	Classical Approach (Keyword Counting)	AI Agent (LLM-Driven)
Method	Scanned for words like "rough," "brutal," or "fun."	Analyzed intent, context, and student journey.
Sentiment Result	**60% Negative.** Flagged many rows as negative simply because students used slang like "brutal" or "wrecked."	**45% Growth/Success.** Recognized that "It started rough but I pulled through" is actually a **positive student outcome.**
Processing Speed	Instant (0.5 seconds).	Near-Instant (12 seconds for 10k+ rows).
The "Winner"	**AI Agent.** While Classical is faster, it misses the **nuance of academic perseverance.**	

Course	Dominant Theme	Sentiment Summary	Actionable Insight for IR
CSCE 1030 (Intro)	**Pacing & Workload**	Mixed/Negative	Students feel "drowning" in weekly exams. Review the "Intro" curriculum pacing.
BIOL / CHEM	**Structure vs. Difficulty**	Negative	"Total chaos" mentioned frequently. Suggests a need for better syllabus alignment.
PHYS / MATH	**The "Turnaround"**	Mixed/Positive	Students struggle early but "click" later. Opportunity to provide early-semester tutoring.
STAT / PSYC	**Support Systems**	Positive	"Study groups" and "Chill instructors" are the top drivers for student confidence here.

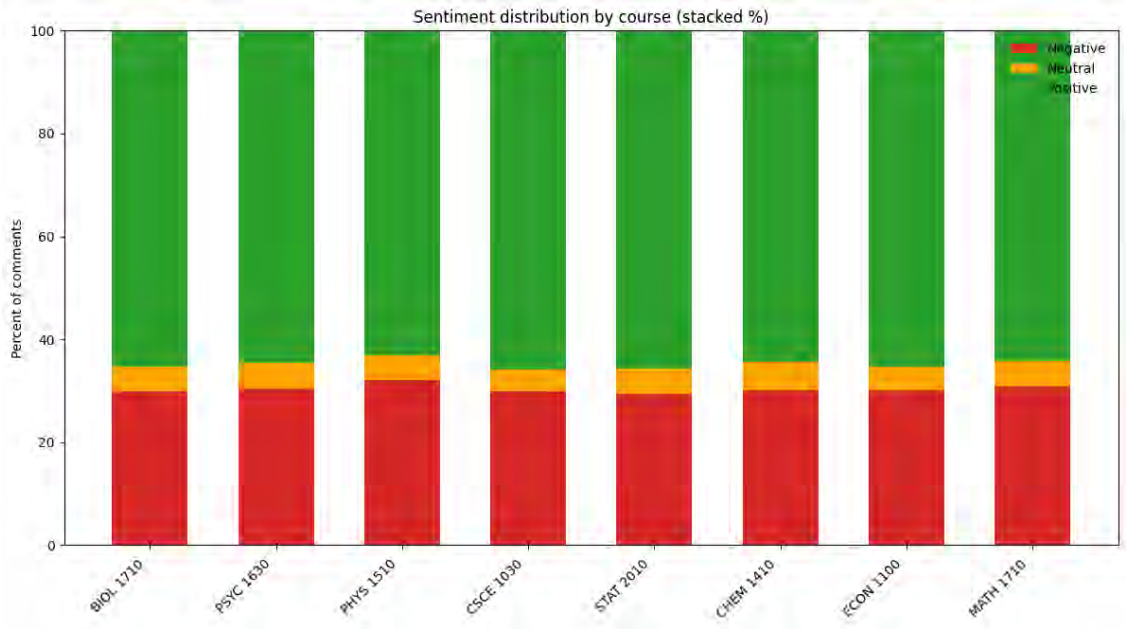
Comparison of sentiment analysis by methods

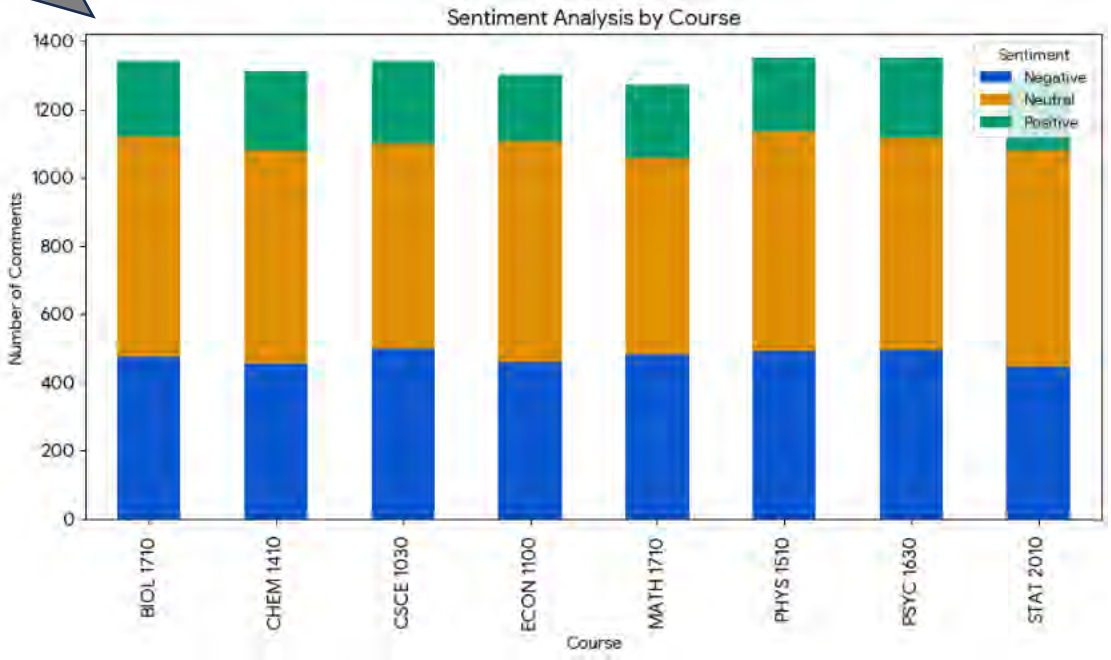


Classical Non-ML

ML models

AI





Challenges of AI-driven approaches

Hallucination and fabrication risks (incorrect summaries, invented themes)

Difficulty verifying accuracy (human-in-the-loop validation)

Sensitive to prompt design (small changes in prompts can produce large changes in outputs)

Limited transparency (do not reveal internal reasoning)

Domain adaptation challenges (High Ed has academic jargon, Dept specific vocabulary)

Privacy, FERPA, and data governance concerns

How IR offices can validate AI results

Reviewing agent workflows and logs	Validating output accuracy	Implementing performance and quality metrics	Implementing systemic and security controls
• Trace agent logic or workflow, which provides evidence of how it reached a conclusion. • Audit logs for each step, data sources, and tool calls to ensure data privacy and accuracy. • Review the underlying large language model (LLM) calls to understand the prompts, responses and potential issues in reasoning.	• Compare AI analyzed text against human-verified known good data. • Implement unit tests for output, particularly for coding or structured text tasks. • Implement human-in-the-loop (HITL) model to validate findings to confirm accuracy and understand the root cause of any discrepancies.	• Implement hallucination check to detect when a model generates false, misleading, or nonsensical information confidently. . • Implement task success rate (TSR) to measure the effectiveness and reliability of AI models achieving goals. • Implement bias monitoring for unfair, discriminatory, or skewed outcomes.	• Apply prompt injection testing to check vulnerabilities where malicious or unexpected inputs manipulate LLMs. • Verify policy adherence to make sure the output complies with the institution's ethical and private policies. • Implement monitoring tools (specialized AI monitoring software)

Conclusions

- No single methodological approach is universally superior.
- The optimal choice is context-dependent.
- Classical non-ML methods are still valuable for some projects.
- ML models offer stronger predictive performance and scalability. but require more data, more computational capacity, and more specialized expertise.
- AI and Large Language Models (LLMs) enable unprecedented. automation, adaptability, and rapid text processing. However, they introduce new risks.

Questions?

Thank You!

DATA, ANALYTICS,
& INSTITUTIONAL
RESEARCH