Sentiment Analysis for Biochem Class Reviews

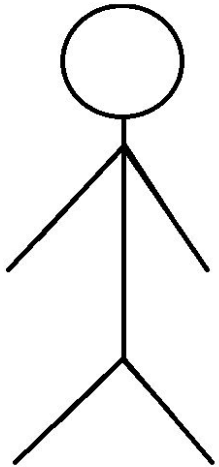
Maddie Bonanno

The nature of STEM science courses

- Usually lecture style
- Exams/tests usually make up large percent of final grade
- Often emphasize memorization over understanding

Is there another way?

Breaking the mold



There exists a professor who runs their courses in a different manner:

- Focus on classroom environment
  - Accessibility
  - Climate
  - “Human-centeredness”

Consider biochemistry with:

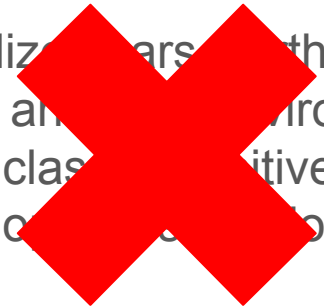
No timed exams, little to no lecturing, no memorization required

This approach is [understandably] incredibly divisive: some students love it, others hate it

Big question: how does a focus on classroom environment impact student course reviews?

Data

Goal: Utiliz ars th of survey data [pre a ironment focus] to clas itive vs negative c ochemistry class



Due to time constraints + factors outside of my control, I was NOT able to obtain most of the professor's survey data

Was aiming for 300+ responses → was given about 50

ACTUAL Data

- Approximately 50 survey responses [end of year review from students enrolled in biochemistry]
- Supplemented this with reviews from rate my professor
    - Limited to reviews of target professor

Total vocabulary = 56 words :/

Processing

- Started with basic “bag of words” model → logistic regression
- Moved to word2vec word embedding → logistic regression

	precision	recall	f1-score	support
0	0.20	0.20	0.20	5
1	0.20	0.20	0.20	5
accuracy			0.20	10
macro avg	0.20	0.20	0.20	10
weighted avg	0.20	0.20	0.20	10

	precision	recall	f1-score	support
0	0.40	0.40	0.40	5
1	0.40	0.40	0.40	5
accuracy			0.40	10
macro avg	0.40	0.40	0.40	10
weighted avg	0.40	0.40	0.40	10

Supplementing the dataset

Small # of documents = small vocabulary = less for model to work with

I expanded the dataset by using rate my professor reviews of other biochemistry/associated courses

50 → 170 reviews

Much better performance on models

Bag of Words → Logistic

	precision	recall	f1-score	support
0	0.75	0.60	0.67	10
1	0.85	0.92	0.88	25
accuracy			0.83	35
macro avg	0.80	0.76	0.78	35
weighted avg	0.82	0.83	0.82	35

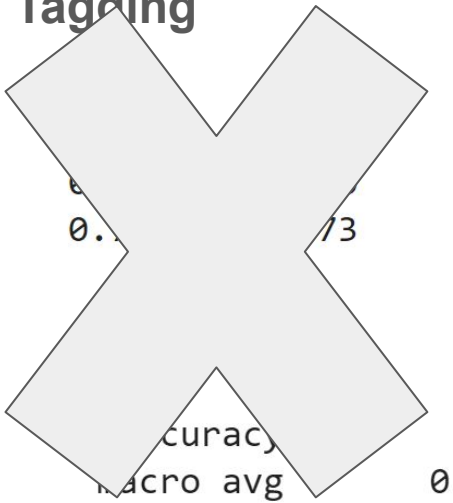
Word2Vec → Logistic

	precision	recall	f1-score	support
0	0.80	0.40	0.53	10
1	0.80	0.96	0.87	25
accuracy			0.80	35
macro avg	0.80	0.68	0.70	35
weighted avg	0.80	0.80	0.78	35

More advanced modeling

Moved to a recurrent neural network for sentiment analysis

First Idea: **Predicting POS Tagging**



accuracy				173
macro avg	0.53			173
weighted avg	0.77	0.	73	173

0.7303773376080472

← *Small dataset*

Large dataset →

			0.72	422
macro avg	0.55	0.60	0.55	422
weighted avg	0.76	0.72	0.73	422

0.7272680695472887

Actually predicting sentiment

Built a recurrent neural network for sentiment analysis

First: **One Hot Encoding**

- Trained on 126 sentences, validated on 21 sentences, tested on 21 sentences

Large dataset →

	precision	recall	f1-score	support
0	0.40	0.67	0.50	6
1	0.82	0.60	0.69	15
accuracy			0.62	21
macro avg	0.61	0.63	0.60	21
weighted avg	0.70	0.62	0.64	21

accuracy = 0.6190476190476191

What is happening right now?

Gridsearch = currently running

- Trying to improve performance of model

I expanded past one-hot-encoding and have been trying: **Glove word embeddings**

- Technically have not gotten these embeddings integrated into the model yet, but it is a work in progress

Challenges and Limitations

- I labeled and collected the data myself
  - Bias in labeling and sampling
- I did not gain access to the data I wanted to use
  - Had to change project direction and alter goals many times
  - Smaller set of data
- Definitely have had challenges doing hyperparameter tuning → still finalizing these values

Thank you