Classifying Thinking Traps with Active Learning

Koa Health Challenge

Final Presentation

Elise Anderson, Eszter Pazmandi & Shereena Salas

Barcelona School of Economics

Introduction/Research Question

- **Goal:**
  - Classify unhelpful thoughts into 8 thinking traps as defined by Cognitive Behavioral Therapy (CBT)
- **Why?**
  - In CBT, this exercise helps patients recognize unhelpful thoughts and develop coping strategies
  - By building a model to categorize them automatically, patients can check their work and save themselves and their therapist's time
- **How?**
  - We were given a small dataset of labeled data and a larger unlabeled dataset from Koa Health
  - We create an Active Learning pipeline using NLP (DistilBERT), a tuned Logistic Regression model, and the Ranked Batch query strategy
- **Results**
  - Binary classification between a "thinking trap" and "not a cognitive distortion"
  - However, for multiclass classification more observations per target "thinking trap" category are needed to improve performance

Data

- 2 kinds of data
- Collected via Amazon’s Mechanical Turk
    - Labeled
    - 330 observations
    - Only contains “thinking traps”
        - 8 main + other
- Unlabeled
    - Anonymized
    - 31,126 observations
    - 255 questions
    - Not all observations are “thinking traps”
    - Drop Spanish questions; filter by length: min 4 words & max 5 sentences
    - Final: 18,094 observations

Methodology – first steps

1. Create initial seed data
   a. Mechanical Turk + 700 hand labeled observations - remaining Spanish answers = 1015 observations in the initial seed data
   b. Highly unbalanced
   c. 8 main thinking traps + Other + “not CD”
2. Baseline multi-label text classification Machine Learning (ML) model
   a. Mix of word embedding techniques + ML models
   b. Winner: DistilBERT + Logistic Regression

Number of observations per category

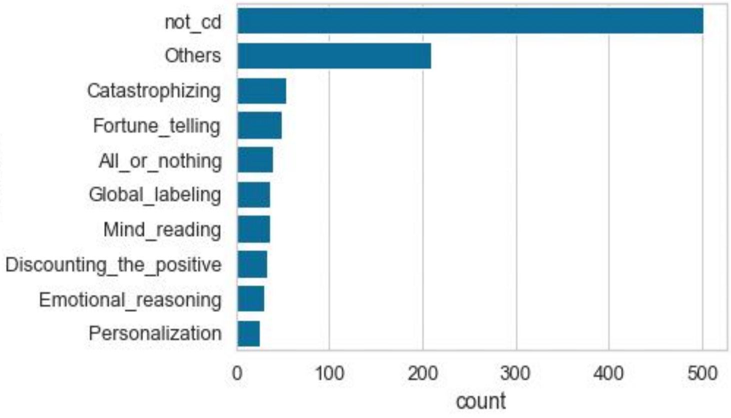
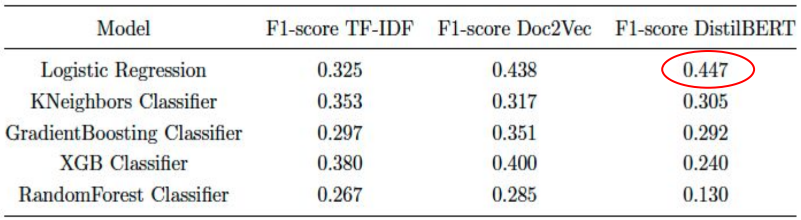


Table 3: Summary of F1-Macro scores by the different word embedding techniques and ML models



Model	F1-score TF-IDF	F1-score Doc2Vec	F1-score DistilBERT
Logistic Regression	0.325	0.438	0.447
KNeighbors Classifier	0.353	0.317	0.305
GradientBoosting Classifier	0.297	0.351	0.292
XGB Classifier	0.380	0.400	0.240
RandomForest Classifier	0.267	0.285	0.130

Methodology - Active Learning

What is Active Learning?

- Active Learning (AL) is a semi-supervised algorithm used to label large amounts of unlabeled data efficiently
- It aims to reduce the amount of data annotated by the human expert by using an iterative process to query unlabeled observations to be labeled.

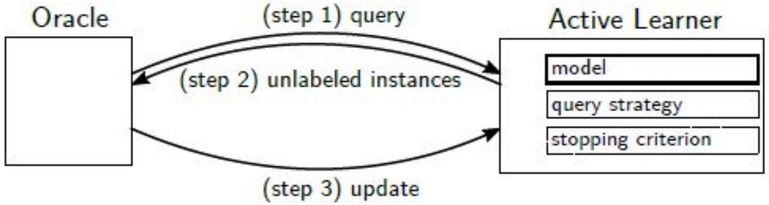
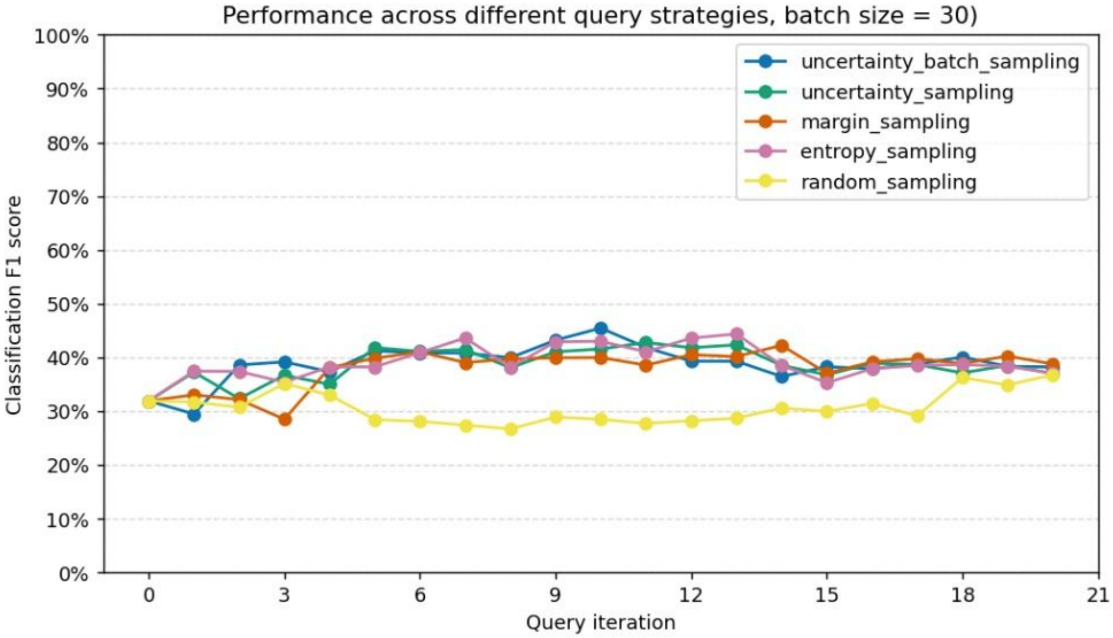


Figure 1: An overview of the AL process: Model, query strategy, and (optionally a) stopping criterion are the key components of an active learner. The main loop is as follows: First the oracle queries the active learner, which returns a fixed amount of unlabeled instances. Then, for all selected unlabeled unstances are assigned labels by the oracle. This process is repeated until the oracle stops, or a predefined stopping criterion is met.

Methodology – Active Learning – Query Strategy

Query Strategy:
Function to decide which observations from the pool should be labeled. The goal is to intelligently select observations to get the highest score (F1) for the least work (labeling).

We tested 5 strategies. They all performed similarly except for random selection, which performed worse.



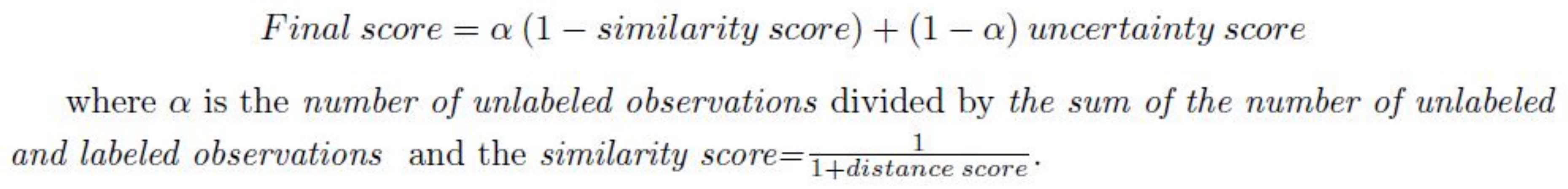
Barcelona School of Economics

Methodology – Active Learning – Query Strategy

Selected: Ranked batch-mode Active Learning (RBMAL).

Combination of uncertainty and similarity: it prioritizes diversity and then shifts the attention to cases where the classifier is most unsure of their classification.

(Cardoso et al. (2017))



$$Final\ score = \alpha\,(1 - similarity\ score) + (1-\alpha)\,uncertainty\ score$$

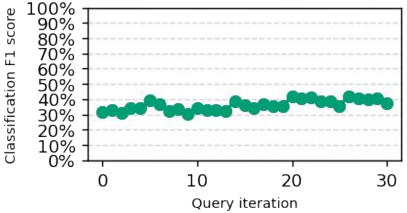
where α is the *number of unlabeled observations* divided by *the sum of the number of unlabeled and labeled observations* and the *similarity score*$=\frac{1}{1+distance\ score}$.

Methodology – Active Learning – Batch Size & Stopping Criteria

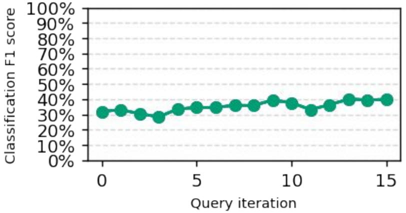
Batch: How many observations labeled per query (100)

Stopping Criteria: Double training data or no improvement of the F1 score

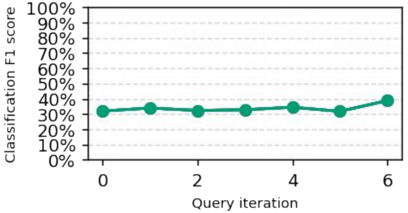
Macro F1-score after 30 queries
(* 20 samples/query = 600 samples): 0.376,
Duration: 7.144 s



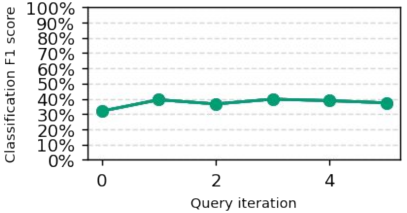
Macro F1-score after 15 queries
(* 40 samples/query = 600 samples): 0.398,
Duration: 3.582 s



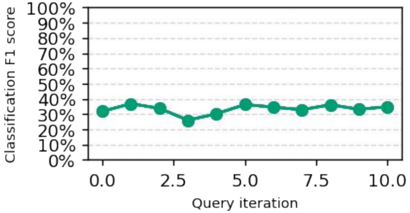
Macro F1-score after 6 queries
(* 100 samples/query = 600 samples): 0.388,
Duration: 1.734 s



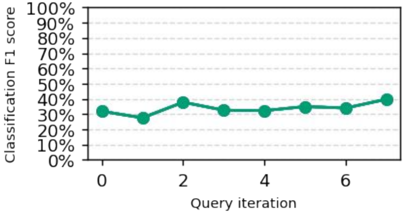
Macro F1-score after 5 queries
(* 120 samples/query = 600 samples): 0.373,
Duration: 1.414 s



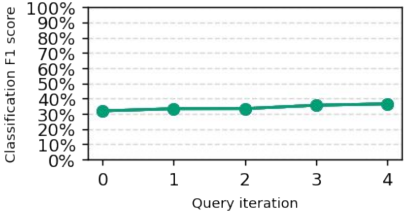
Macro F1-score after 10 queries
(* 60 samples/query = 600 samples): 0.349,
Duration: 2.547 s



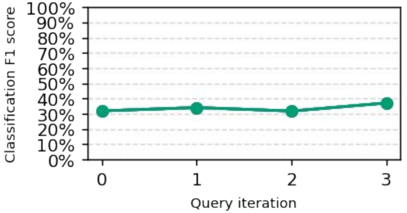
Macro F1-score after 7 queries
(* 80 samples/query = 560 samples): 0.398,
Duration: 1.671 s



Macro F1-score after 4 queries
(* 140 samples/query = 560 samples): 0.366,
Duration: 1.159 s

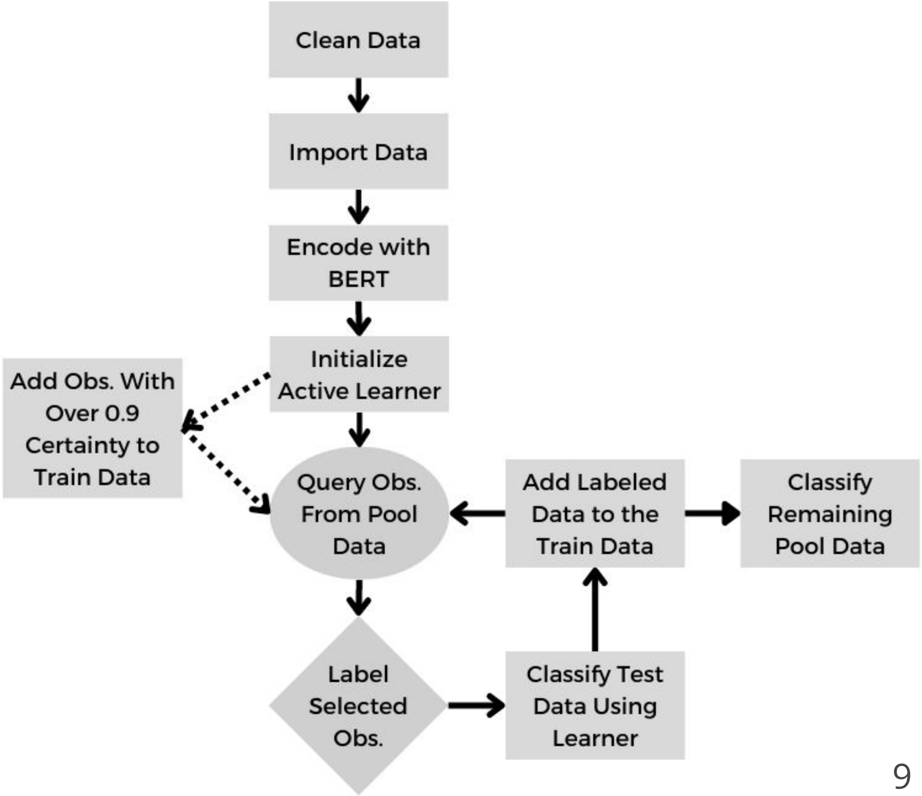


Macro F1-score after 3 queries
(* 160 samples/query = 480 samples): 0.371,
Duration: 0.833 s

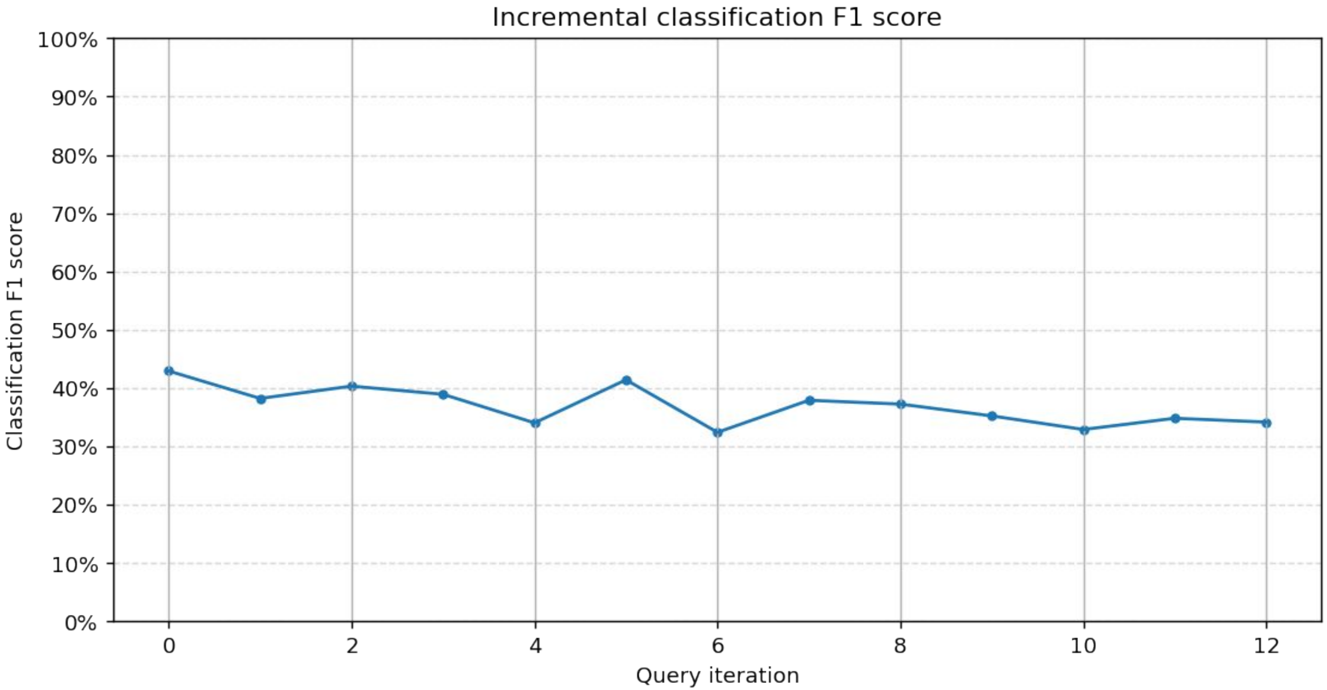


Methodology - Active Learning Pipeline

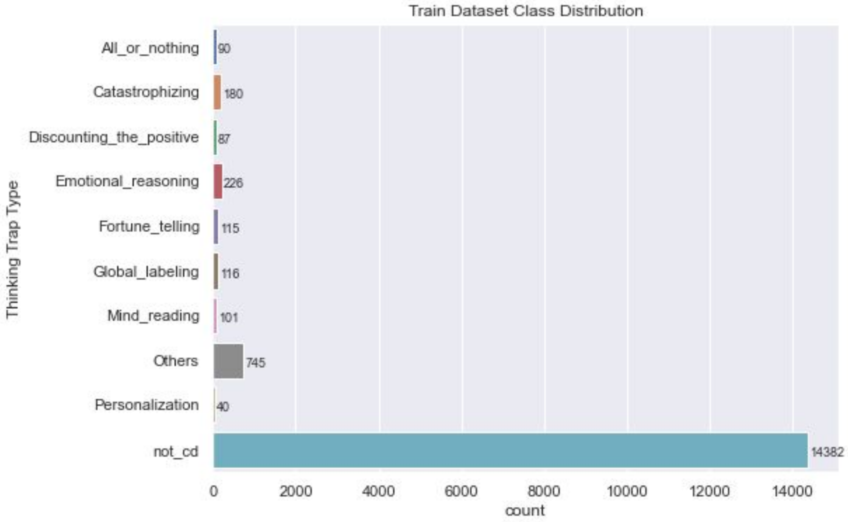
- Import cleaned data (train, test, and pool)
- Encode with DistilBERT model
- Initialize AL with Logistic Regression
  - Add observations categorized with > 90% certainty to train data
- Start loop:
  - Ranked Batch query to return a list of observations
  - Label observations
  - Categorize test data into 8 thinking traps, “other” thinking traps, and not a cognitive distortion
  - Calculate F1 score for test data
  - Add the labeled data to the train data
- End loop after stopping criteria met
- Classify remaining pool data

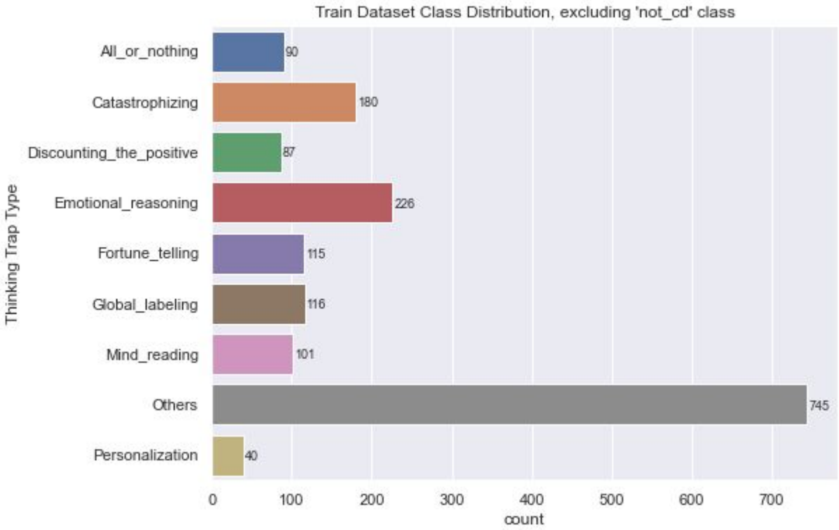

Clean Data
Import Data
Encode with BERT
Initialize Active Learner
Add Obs. With Over 0.9 Certainty to Train Data
Query Obs. From Pool Data
Add Labeled Data to the Train Data
Classify Remaining Pool Data
Label Selected Obs.
Classify Test Data Using Learner

Results – Active Learning



Results – Active Learning



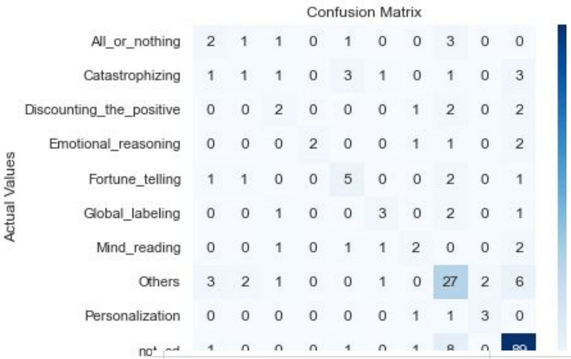


Results – Active Learning

	Seed data training set (out-of-sample performance)	AL training set (out-of-sample performance)	Support (out-of-sample)	AL training set (in-sample performance)
All or nothing	0.25	0.00	8	0.68
Catastrophizing	0.13	0.00	11	0.72
Discounting the positive	0.29	0.18	7	0.67
Emotional reasoning	0.50	0.29	6	0.95
Fortune telling	0.48	0.53	10	0.78
Global labeling	0.46	0.31	7	0.80
Mind reading	0.31	0.40	7	0.86
Others	0.61	0.56	42	0.88
Personalization	0.60	0.22	5	0.89
Not cognitive distortion	0.86	0.81	100	0.99
Macro F1-Score	**0.45**	**0.33**	**203**	**0.82**
Accuracy	0.67	0.62	203	0.97

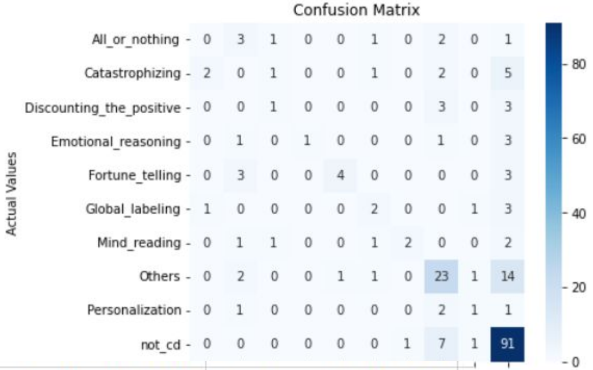
Results – Active Learning

Using only the seed data (initial train set)



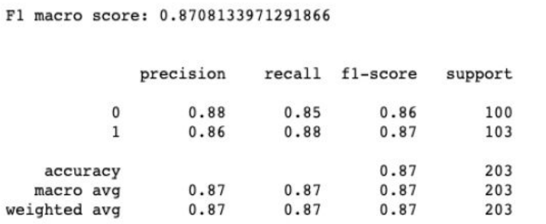
All_or_nothing	2	1	1	0	1	0	0	3	0	0
Catastrophizing	1	1	1	0	3	1	0	1	0	3
Discounting_the_positive	0	0	2	0	0	0	1	2	0	2
Emotional_reasoning	0	0	0	2	0	0	1	1	0	2
Fortune_telling	1	1	0	0	5	0	0	2	0	1
Global_labeling	0	0	1	0	0	3	0	2	0	1
Mind_reading	0	0	1	0	1	1	2	0	0	2
Others	3	2	1	0	0	1	0	27	2	6
Personalization	0	0	0	0	0	0	1	1	3	0
not_cd	1	0	0	0	1	0	1	8	0	89

Using training set from AL



All_or_nothing	0	3	1	0	0	1	0	2	0	1
Catastrophizing	2	0	1	0	0	1	0	2	0	5
Discounting_the_positive	0	0	1	0	0	0	0	3	0	3
Emotional_reasoning	0	1	0	1	0	0	0	1	0	3
Fortune_telling	0	3	0	0	4	0	0	0	0	3
Global_labeling	1	0	0	0	0	2	0	0	1	3
Mind_reading	0	1	1	0	0	1	2	0	0	2
Others	0	2	0	0	1	1	0	23	1	14
Personalization	0	1	0	0	0	0	0	2	1	1
not_cd	0	0	0	0	0	0	1	7	1	91

Actual	Confused with (initial)	Confused with (AL)
All or nothing	Others	Catastrophizing
Catastrophizing	Fortune telling	All or nothing
Discounting the positive	Others, not cd	Others, not cd
Emotional reasoning	not cd	not cd
Fortune telling	Others	Catastrophizing
Global labeling	Others	not cd
Mind reading	not cd	not cd
Others	not cd	not cd
Personalization	Global labeling	Others
not cd	Others	Others



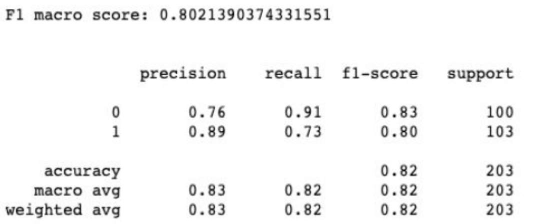
```
F1 macro score: 0.8708133971291866

              precision    recall  f1-score   support

           0       0.88      0.85      0.86       100
           1       0.86      0.88      0.87       103

    accuracy                           0.87       203
   macro avg       0.87      0.87      0.87       203
weighted avg       0.87      0.87      0.87       203
```

(a) Classification report of the initial training set



```
F1 macro score: 0.8021390374331551

              precision    recall  f1-score   support

           0       0.76      0.91      0.83       100
           1       0.89      0.73      0.80       103

    accuracy                           0.82       203
   macro avg       0.83      0.82      0.82       203
weighted avg       0.83      0.82      0.82       203
```

(b) Classification report of the final training set

Possible reasons for the results

- Divergence of test and train distribution
- Very small number of train observations per class
- Using RBMAL instead of building a custom query strategy
  - Prioritizes dissimilarity first before uncertainty. Might be more optimal to use uncertainty first before dissimilarity given the varying characteristics of the instances (not-so-useful observations could be very dissimilar to the current training set)
- Inconsistent labeling
- Multilabel classification
  - no restriction on the number of classes that an instance may be allocated to

Conclusion

What we accomplished:

- We created an Active Learning pipeline that is optimized for Koa Health’s data
- Our model excelled at binary classification
  - “Thinking Trap” vs “Not a Cognitive Distortion”
- However, for classifying “thinking traps” we need more data per category
  - Very few observations per target category in the unlabeled dataset

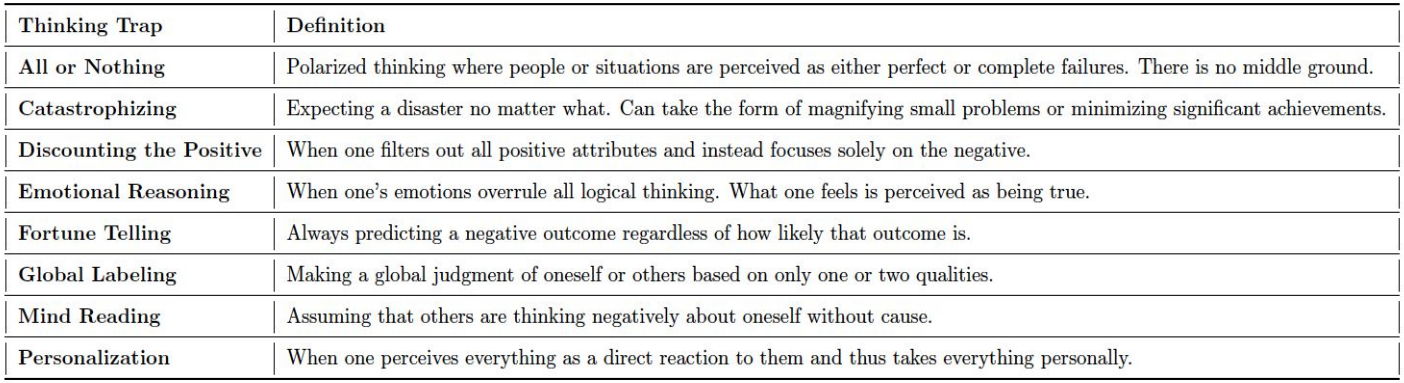
How to apply our work:

- Feed more quality data into the pipeline
  - Filter questions/answers by sentiment
- Customize the query strategy
- To optimize the existing observations, having a trained therapist or a panel of annotators reach consensus would assure quality labeling

Thank you for your attention!

Questions?

Table 1: Definition of the 8 main thinking traps



Thinking Trap	Definition
All or Nothing	Polarized thinking where people or situations are perceived as either perfect or complete failures. There is no middle ground.
Catastrophizing	Expecting a disaster no matter what. Can take the form of magnifying small problems or minimizing significant achievements.
Discounting the Positive	When one filters out all positive attributes and instead focuses solely on the negative.
Emotional Reasoning	When one's emotions overrule all logical thinking. What one feels is perceived as being true.
Fortune Telling	Always predicting a negative outcome regardless of how likely that outcome is.
Global Labeling	Making a global judgment of oneself or others based on only one or two qualities.
Mind Reading	Assuming that others are thinking negatively about oneself without cause.
Personalization	When one perceives everything as a direct reaction to them and thus takes everything personally.

Recommendations for Future Work

- Filter per question, remove questions that would have positive answers
- Customize query strategy based on characteristics of data
- Use of a different question prompt
  - “Add the thought that made you feel bad in that situation”
    - 'Write a balanced version of this thought:',
  - “Most negative take on a problem”
    - re-think and give a more positive angle on the problem
- Multilabel classification
- Quality controls (consensus and time) or expert annotators; text augmentation
- Use DistilBERT/BERT as a classifier not only for word embeddings
- implementing AL incrementally
  - Increase number of sentences

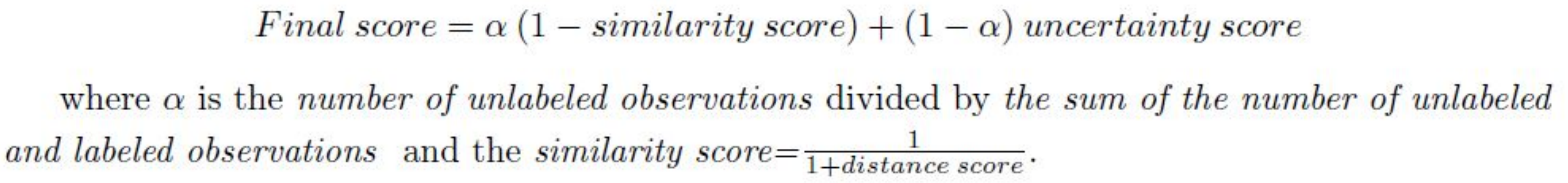
Classify in 2 steps: first binary, then classifying into categories

Literature Review

- Cognitive Behavioral Therapy (CBT) and Thinking Traps
- Studies on Psychology and Machine Learning
- Machine Learning Models for active learning and classification of text data in the psychology domain

Literature Review – Active Learning, scenarios, and key query strategies

- help identify which instances are most informative to be labeled so that the manually labeled observations would be representative of the whole dataset.
- Ranked batch-mode Active Learning (RBMAL) (Cardoso et al. (2017))
- preferred as it decreases the number of algorithm iterations while retaining or even improving the quality of the selected instances.
- Prioritize diversity and then shifting the attention to cases where the classifier is unsure as the amount of labeled information increases.



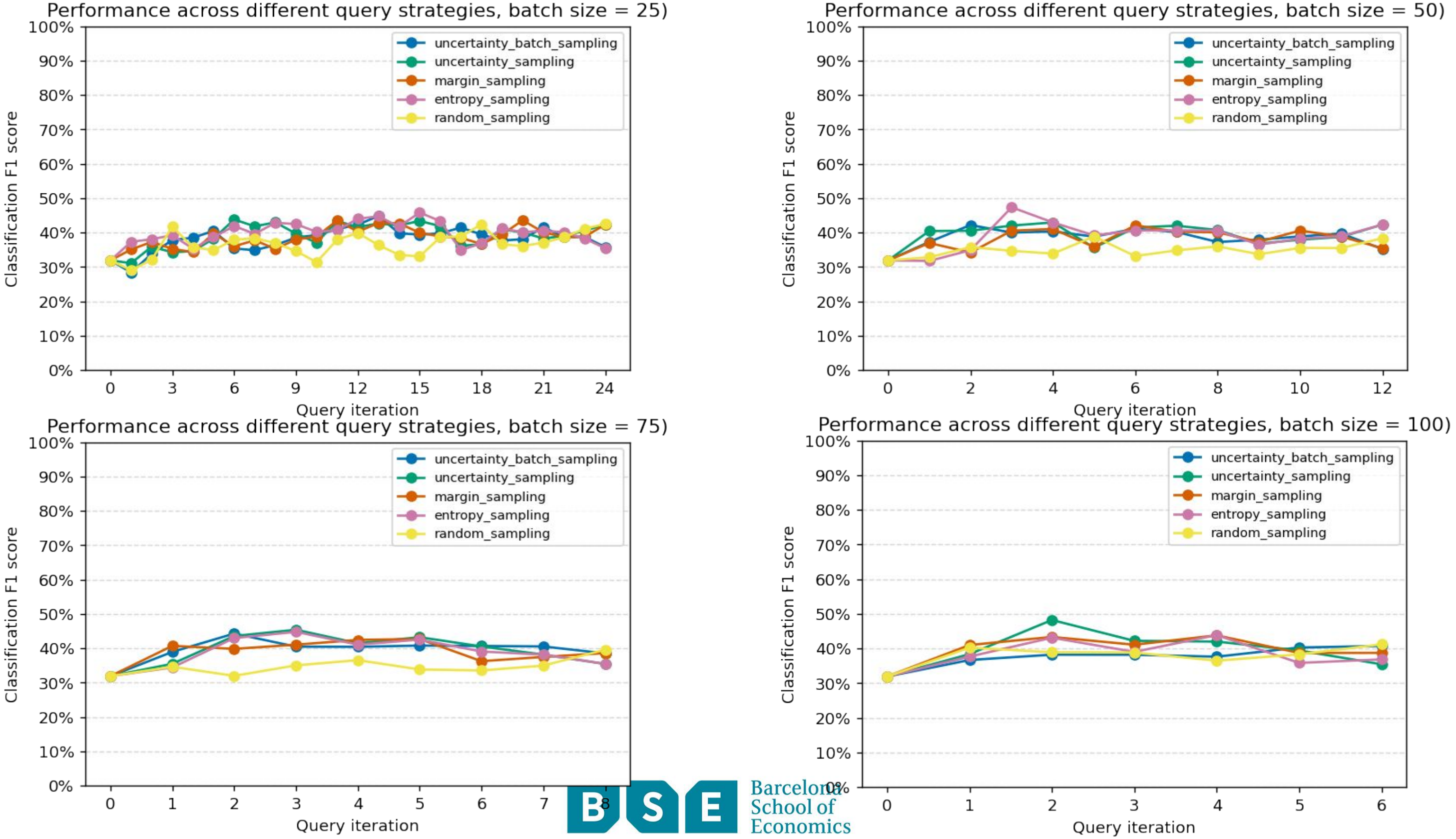
$$Final\ score = \alpha\,(1 - similarity\ score) + (1 - \alpha)\ uncertainty\ score$$

where α is the $number\ of\ unlabeled\ observations$ divided by $the\ sum\ of\ the\ number\ of\ unlabeled\ and\ labeled\ observations$ and the $similarity\ score = \frac{1}{1 + distance\ score}$.

Literature Review - Crowdsourcing/non-experts to label seed data

- Zhao et al. (2020)
  - crowdsourcing is useful to create high-quality annotations with some quality controls in place such as discarding observations that have been annotated too quickly.
  - annotation quality by a 3-worker group and a 5-worker group does not make a difference with regard to annotation reliability

Methodology – Active Learning – Query Strategy


Performance across different query strategies, batch size = 25)
Performance across different query strategies, batch size = 50)
Performance across different query strategies, batch size = 75)
Performance across different query strategies, batch size = 100)
Classification F1 score
Query iteration
uncertainty_batch_sampling
uncertainty_sampling
margin_sampling
entropy_sampling
random_sampling