Ontology-driven Self-supervision for Adverse Childhood Experiences Identification Using Social Media Datasets

Jinge Wu[1,2], Rowena Smith[2] and Honghan Wu[1]
[1]Institute of Health Informatics, University College London, London, United Kingdom
[2]Usher Institute, University of Edinburgh, Edinburgh, United Kingdom

23th July 2022

By Jinge Wu

Jinge.wu.20@ucl.ac.uk

UCL

THE UNIVERSITY of EDINBURGH

HDRUK Health Data Research UK

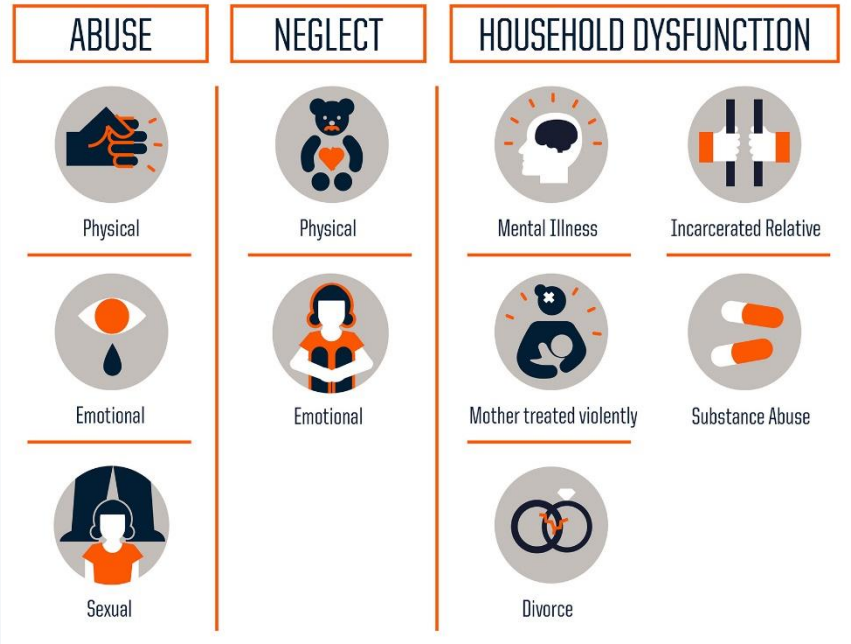
1. Research Background & Motivations

2. Methods

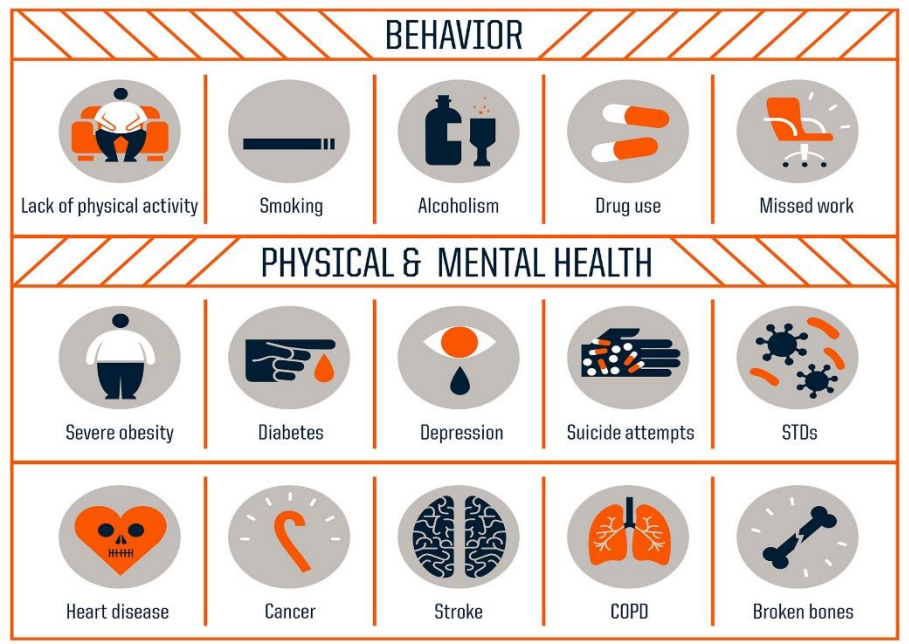
3. Experiments & Results

4. Conclusion & Discussion

Research Background



Adverse Childhood Experiences (ACEs)



Source: Centers for Disease Control and Prevention

Research Motivations

It is difficult to find clues from clinical notes

The ACE causal route is complicated and most of such information is not coded in structured format, requiring the ability to interrogate multimodal data including structured and unstructured Electronic Health Records (EHR) data

Current **ACE ontology** not directly applicable for **Natural Language Processing**

Insufficient relevant, **open-accessible** EHR data for training models

Manual annotation is time-consuming

ACE ontology

Adverse Childhood Experiences Ontology

Last uploaded: January 25, 2019

Summary | Classes | Properties | Notes | Mappings

Jump to:

- ACES_Scores
- Activity
- Adverse_Childhood_Experiences
- Clinical finding
- Clinical findings
- Conceptual Entity
- Disease, Disorder or Finding
  - Disease
    - Developmental disorder
    - Drug-related disorder
    - mental disease
      - Mental disorder due to drug
        - Psychoactive substance use disorder
          - Psychoactive substance dependence
            - Alcoholism
              - **Alcohol abuse**
              - Alcohol Abuse

Source: BioPortal

Details | Visualization | Notes (0) | Class Mappings (41)

Preferred Name	Alcohol abuse
ID	http://purl.bioontology.org/ontology/SNOMEDCT/15167005
label	Alcohol abuse
prefLabel	Alcohol abuse
subClassOf	Psychoactive substance abuse Substance abuse Alcoholism Substance Abuse

Alcohol abuse

UMLS CUI: C0085762

Add to List

Semantic Types: Mental or Behavioral Dysfunction

Definitions (2)

The use of alcoholic beverages to excess, either on individual occasions ("binge drinking") or as a regular practice. ... **(NCI)**

Atoms (134)

Filter by Vocabulary

Reset Filters [x]

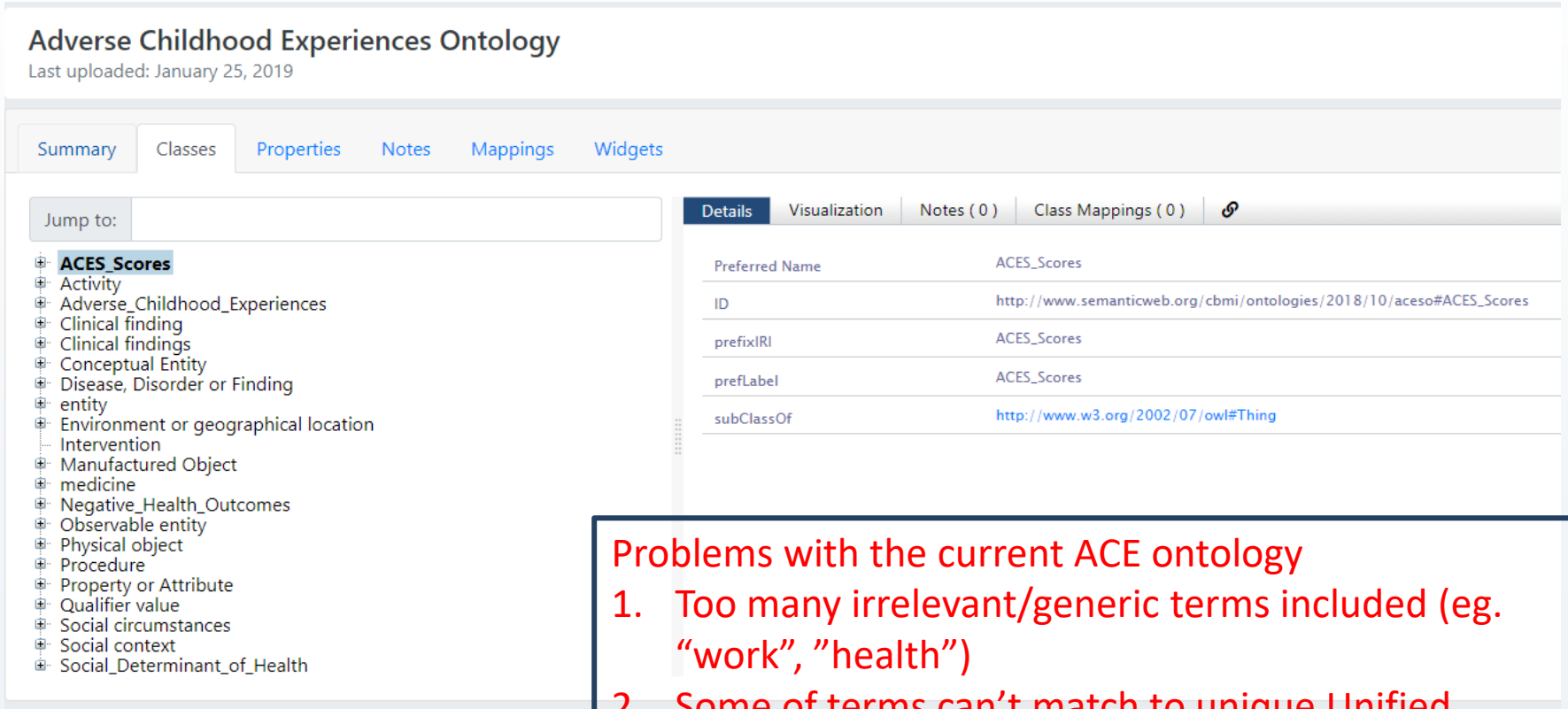
Search Atoms

Name	AUI	Vocabulary	Term Type	Code
Alcohol abuse	A0388935	MTH	PN	NOCODE
Alcohol Abuse	A7755846	MSH	PEP	D000437
Ethanol Abuse	A32284317	MSH	ET	D000437
Abuse, Alcohol	A0387703	MSH	PM	D000437
Abuse, Ethanol	A32284806	MSH	PM	D000437
Alcohol abuse	A2922613	SNOMEDCT_US	PT	15167005
Alcohol abuse (disorder)	A3536058	SNOMEDCT_US	FN	15167005
AA - Alcohol abuse	A3235734	SNOMEDCT_US	SY	15167005
Ethanol abuse	A3238860	SNOMEDCT_US	SY	15167005
Alcohol abuse	A13075252	MDR	PT	10001584
Alcohol abuse	A25683378	MDR	LLT	10001584

Source: UMLS

Research Motivations

Current ACE ontology not applicable for Natural Language Processing



Source: BioPortal

Problems with the current ACE ontology

1. Too many irrelevant/generic terms included (eg. “work”, ”health”)
2. Some of terms can’t match to unique Unified Medical Language System (UMLS) codes, some even don’t have codes

1. Research Background & Motivations

2. Methods

3. Experiments & Results

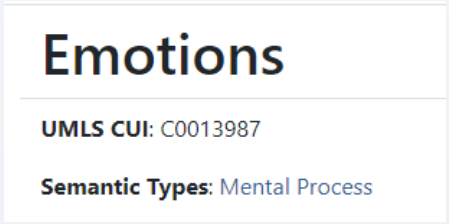
4. Conclusion & Discussion

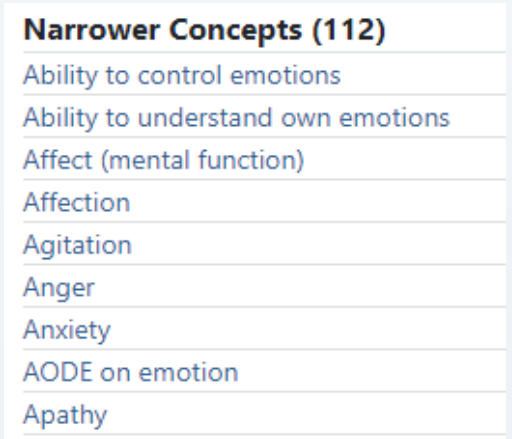
ACE concepts

1. Leaf nodes extract from ACESO, i.e. narrow concepts

2. Manual check and add additional terms

3. Find all descendants and map to UMLS concepts





Named entity recognition (NER)

SemEHR -open source toolkit that integrates text mining and semantic computing for identifying mentions of UMLS concepts from texts. This allows us to identify as many variants of disease mentions as possible and contributes to disambiguation

Why I wish i had a mental disorder. Oh yes passive agressive title. I did what people suggested me to. My family. My ex. Strangers online. "please seek out professional help, OP". Well, they won't help me. I have such an intense fear of abandonment. And uncontrollable mood swings. And unstable self image. The psychiatrist tested me for borderline personality disorder. I fit a lot of the symptoms, but not enough to qualify for a diagnosis. I know that NOT having BPD is good, because BPD is bad. But if I would've had it, they would have helped me and given me treatment.

I self harm. I hate myself. I engage in certain risk taking/self destructive relations with men (seeking out men who tell me to hurt myself is one example). I'm sad a lot. I have no friends whatsoever. I'm diagnosed with social phobia. But the psychiatrist isn't offering a treatment plan for these things. She is talking about discharging me. She knows I struggle with all of these problems.

I actually wish I had a serious mental disorder so that psychiatrists would do.... something.

False labels (not ACEs)

True labels (ACEs)

Example of NER task and annotation results from Reddit

Concept embeddings

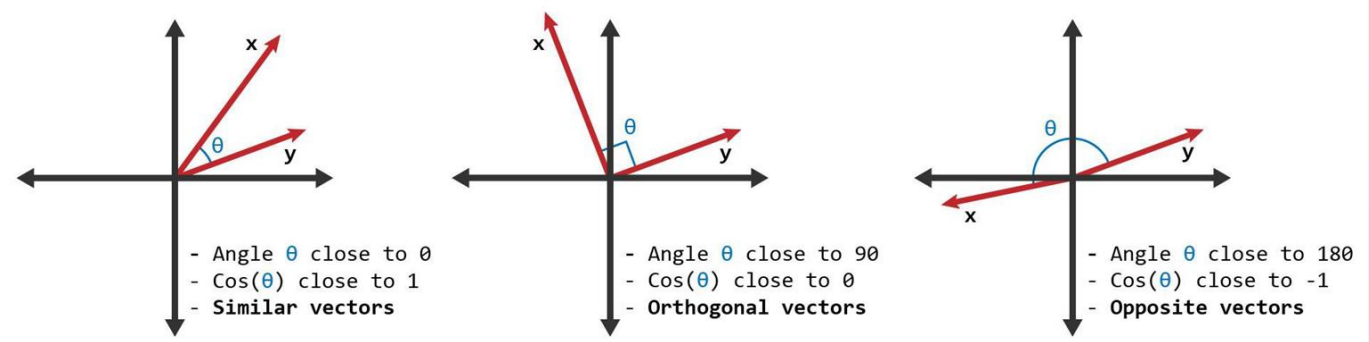
Take the contextual information into consideration

1. Document-level matrix: co-occurrence of individual concepts in each document (n x m)
   m: number of unique concepts; n: number of documents/samples
2. Concept-level matrix: co-occurrence of individual concepts (m x m)

Cosine similarity

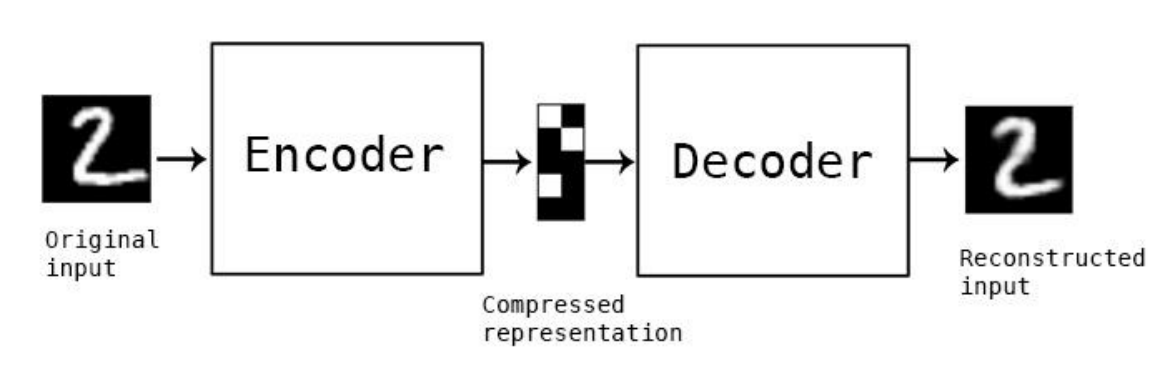
$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

If similarity > threshold:
positive value (i.e ACE)
Else:
negative value (i.e. not ACE)

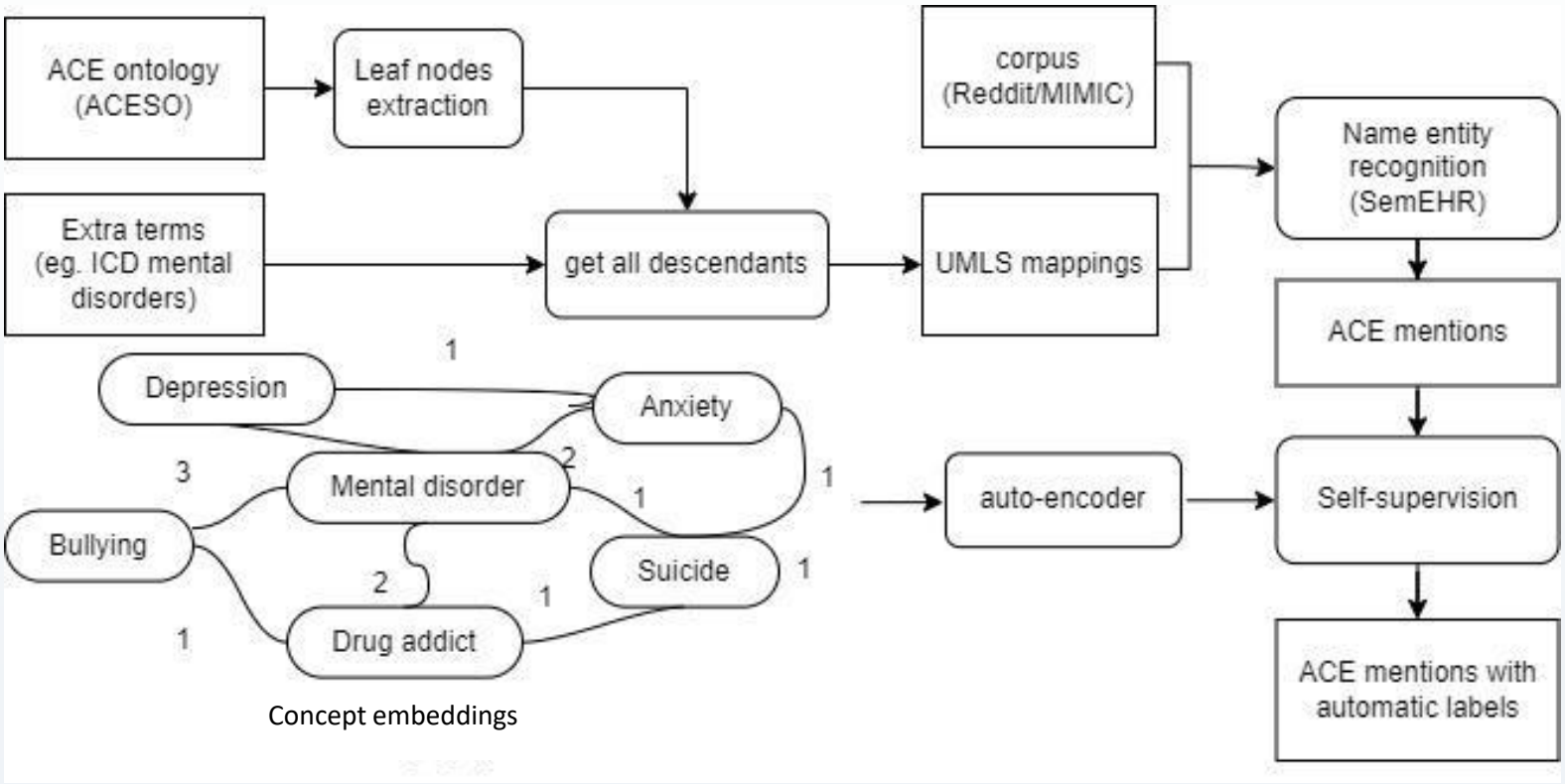


Dimensionality reduction: Auto-encoder

1. To deal with scarcity of concept matrix
2. Compress it into the lower dimensions and then decode the data to reconstruct the original input



Research Methodology & Structure



1. Research Background & Motivations

2. Methods

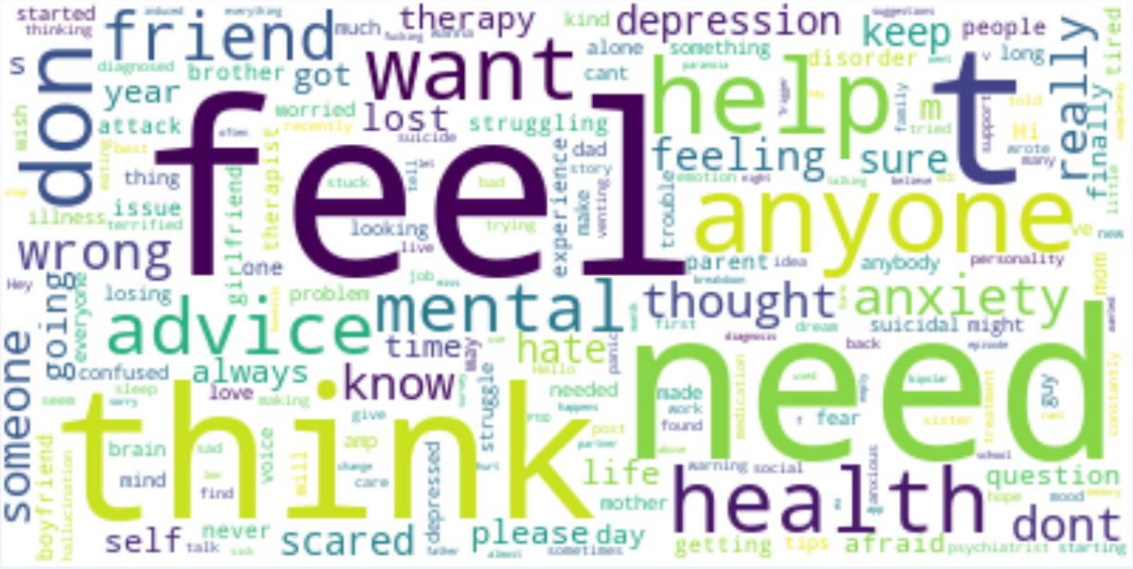
3. Experiments & Results

4. Conclusion & Discussion

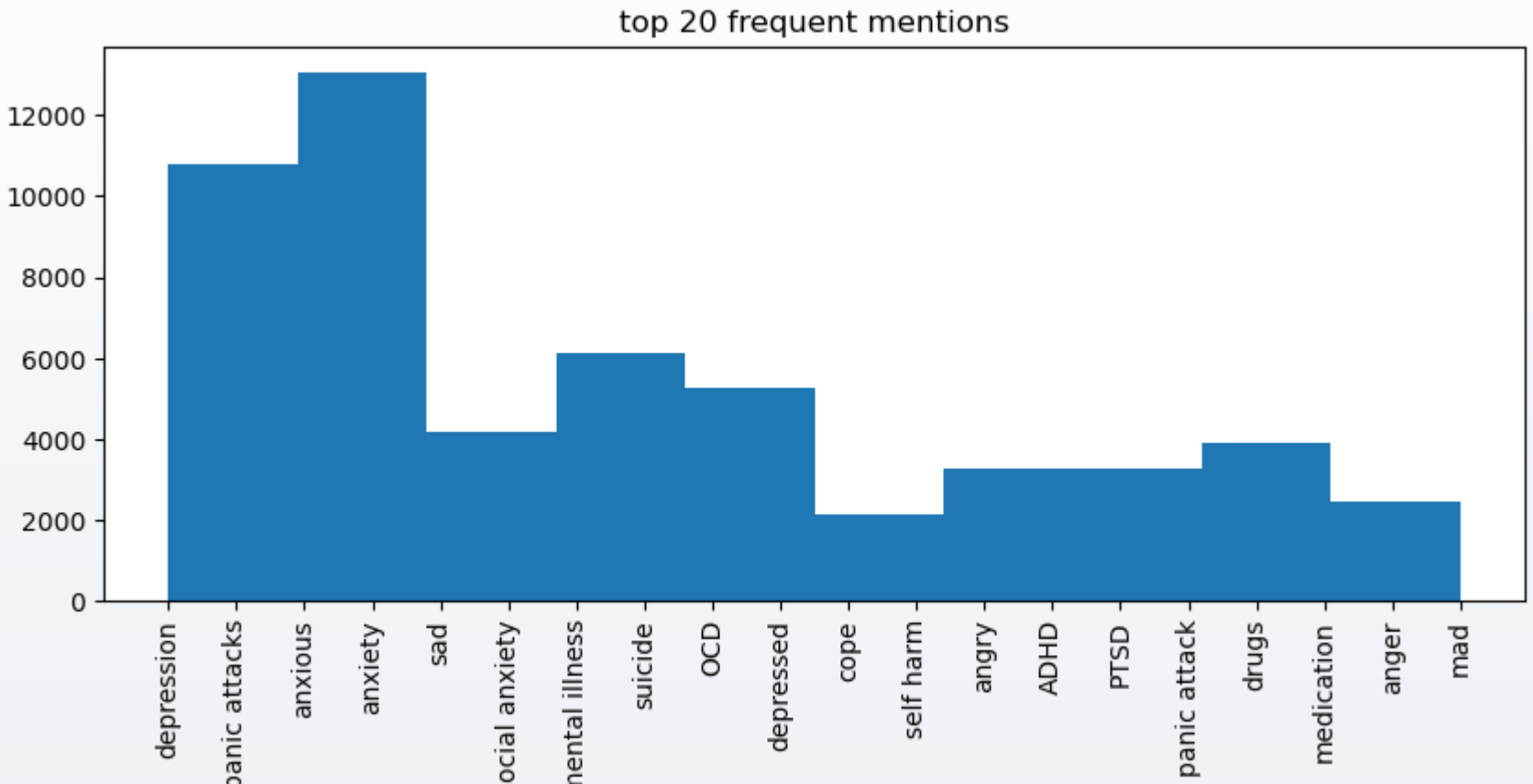
Data collection

Social media corpus – **Reddit**

- With the topic of mental health
- 32439 posts between 2018-2020

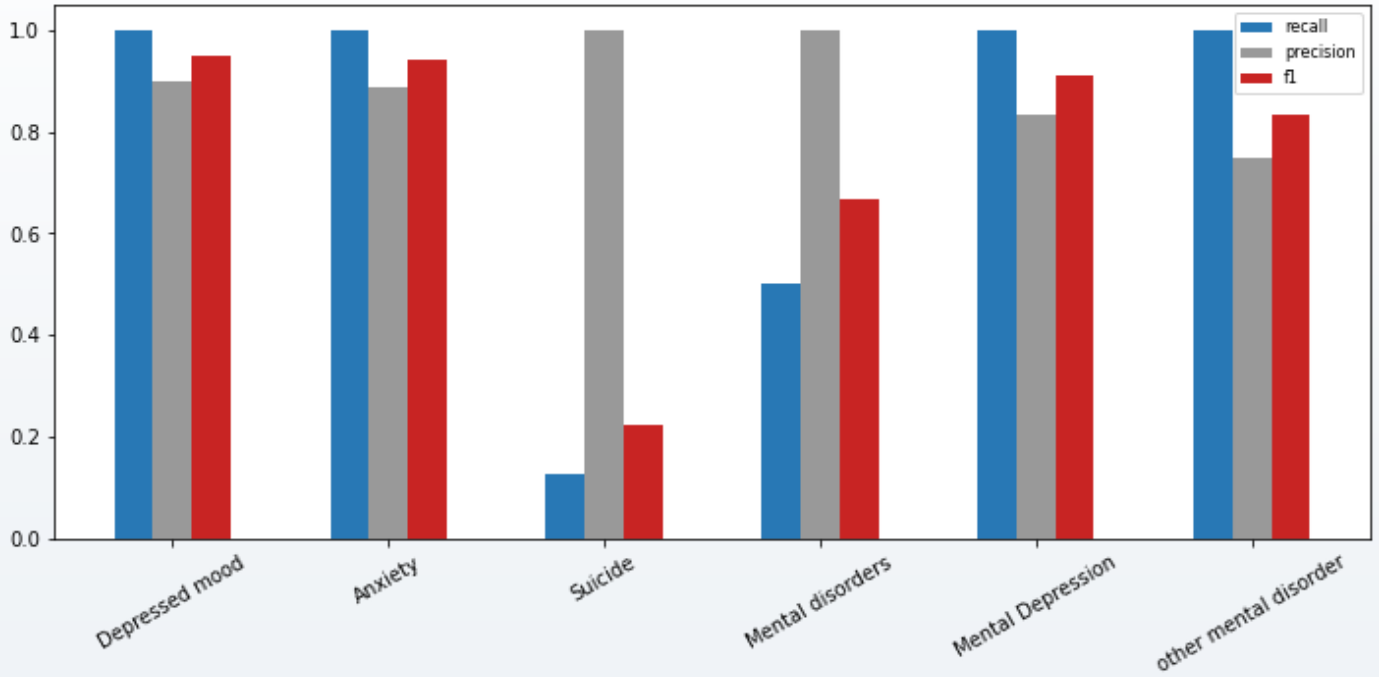


Experiments and Results



Experiments and Results

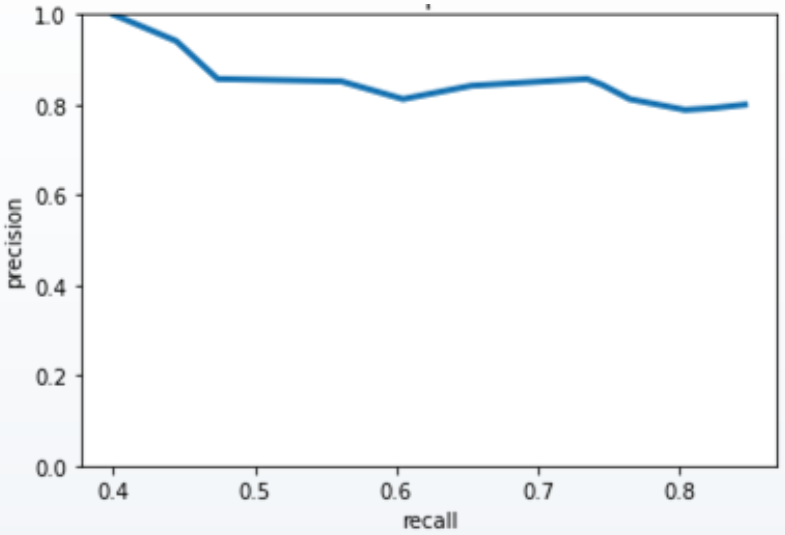
Accuracy at individual concept level for Reddit data



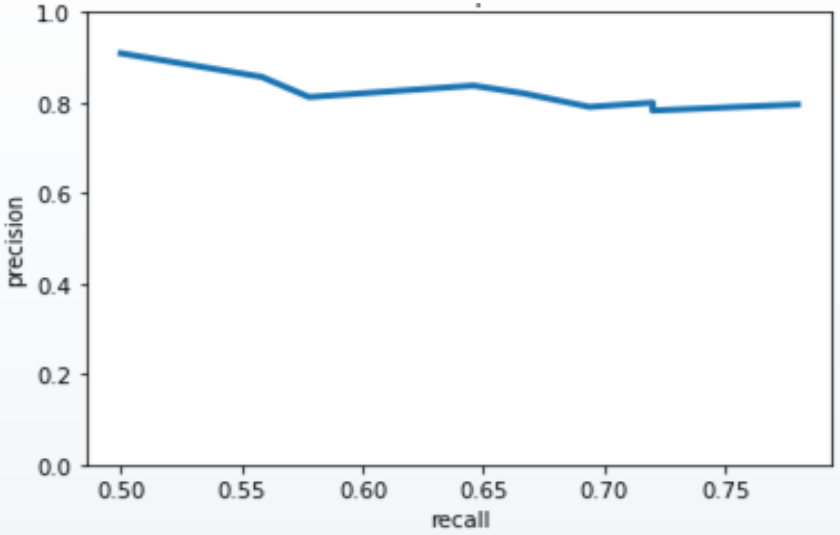
Low recall for “suicide”: there are many terms to express suicide,
For example, "end my life", however, the model failed to identify this

Experiments and Results

Precision-recall area under curve (AUC) without auto-encoder

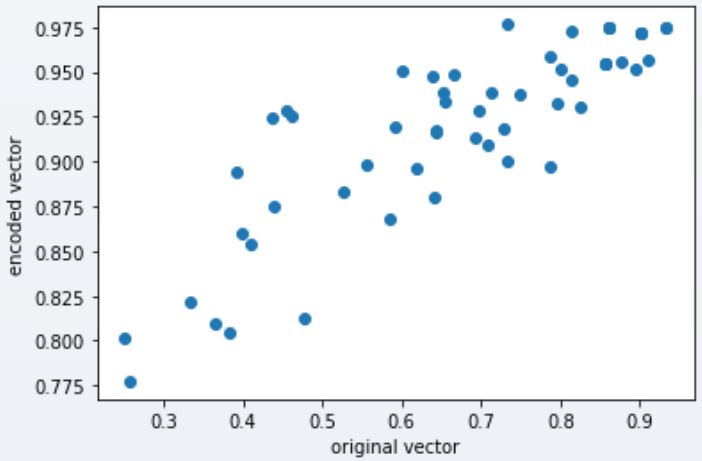


Precision-recall area under curve (AUC) with auto-encoder



This shows our pretrained auto-encoder retains useful information from the original vectors

Scatter plot of original vector vs encoded vector



1. Research Background & Motivations

2. Methods

3. Experiments & Results

4. Conclusion & Discussion

Conclusion

Automated ACE identification from free-text data particularly in low-resource environments

Identified extra ontological terms from UMLS and developed comprehensive mappings to its narrower concepts

Proposed practical approaches of self-supervision by utilising the concept co-occurence graph – beneficial for automatic labelling

Created reusable concept embeddings

Discussion

1. Future plan:
    i. Experiments on EHR data
    ii. Enhancement for ACE identification model
        - Informal language (Eg. kill myself -> suicide)
        - Words or phrases (Eg. “can't go back to sleep”)
    iii. Create a public accessible benchmark for ACE identification

Thank you