MEDINFO23

8 – 12 JULY 2023 | SYDNEY, AUSTRALIA

KING'S College LONDON

Identifying Mentions of Pain in Mental Health Records Text: A Natural Language Processing Approach

project website

Jaya Chaturvedi

PhD student

King's College London, UK

@JayaChatur

@PainandMH

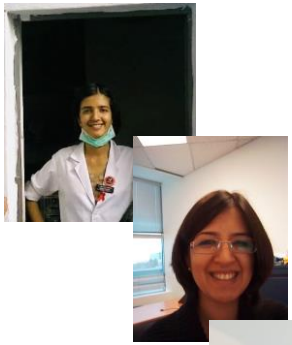
AIDH IMIA

Layout of Presentation

- About me
- Background
- Methods
- Results
- Conclusion
- Acknowledgements
- Q&A

About me

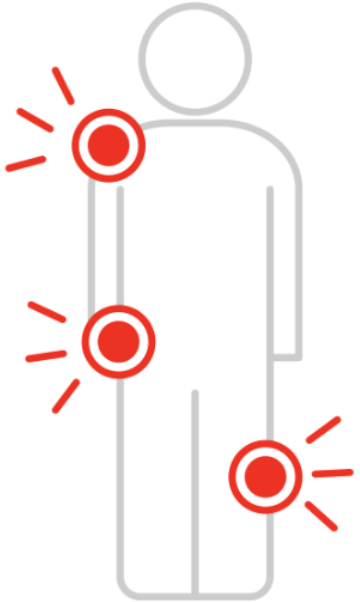
- Dentist
- Love for Data
- Passionate about mental health and pain research
- Final year of my PhD (submitting in 1.5 months!)



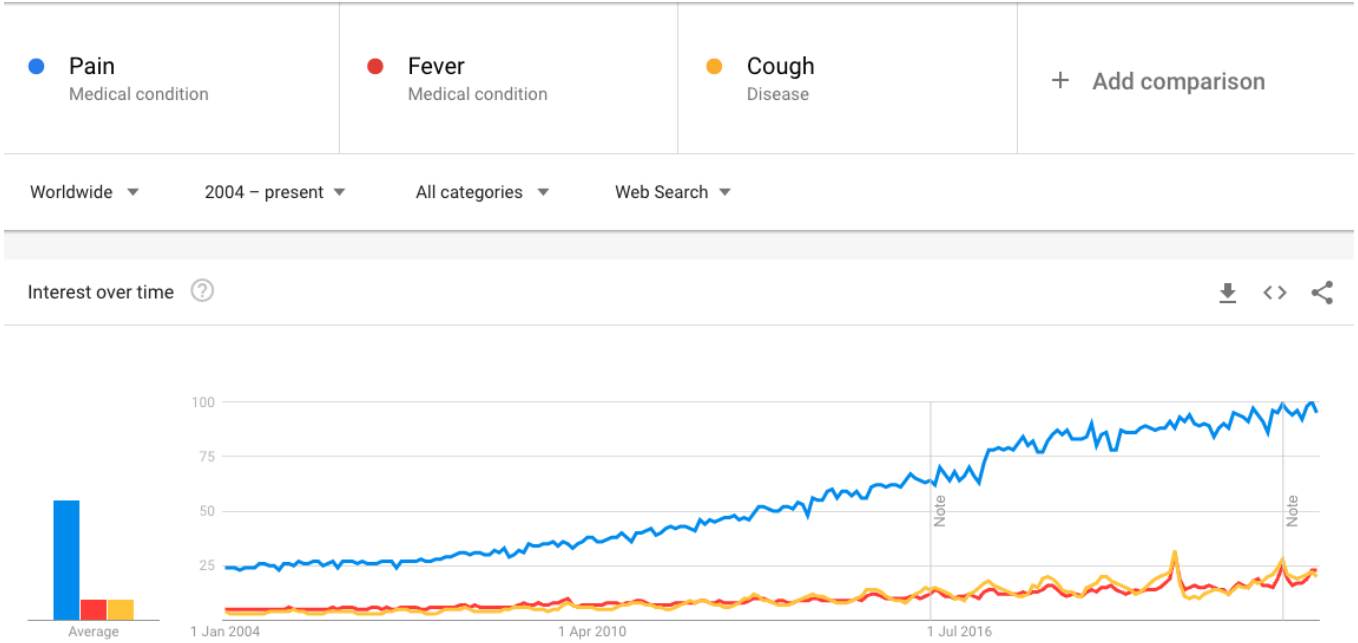
Background

Pain

- an unpleasant sensory and emotional experience
- very subjective in nature
- high co-occurrence of pain and mental health disorders
- common reason for people to access healthcare facilities



Pain

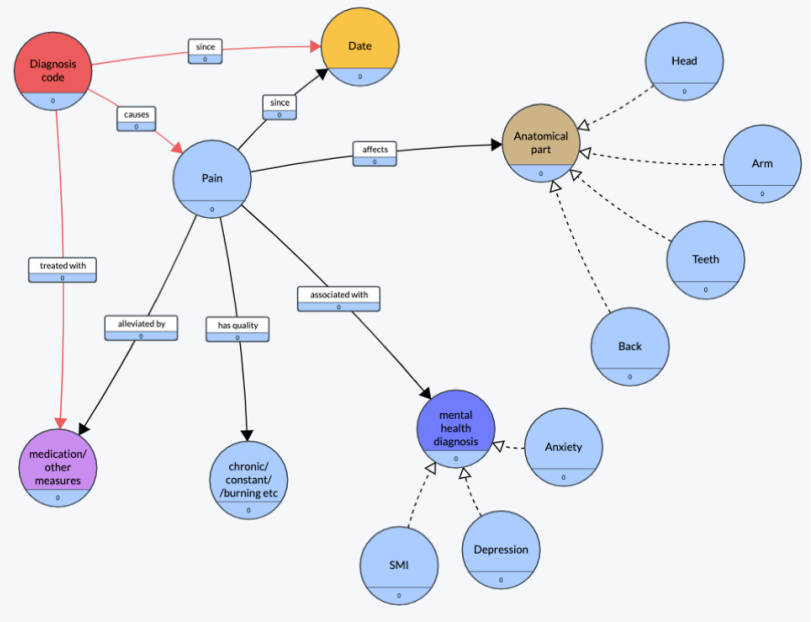

Pain
Medical condition
Fever
Medical condition
Cough
Disease
Add comparison
Worldwide
2004 – present
All categories
Web Search
Interest over time
100
75
50
25
Note
Note
Average
1 Jan 2004
1 Apr 2010
1 Jul 2016

Conceptual Diagram of Pain

Developed after exploration of 4 text sources:

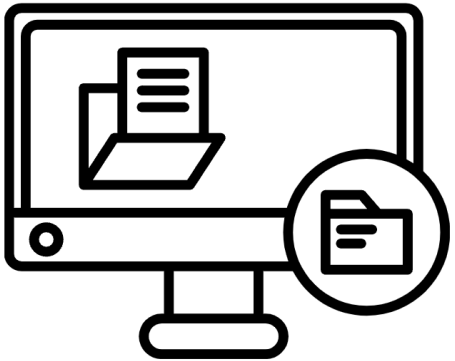
2 EHR databases (MIMIC-III and CRIS)

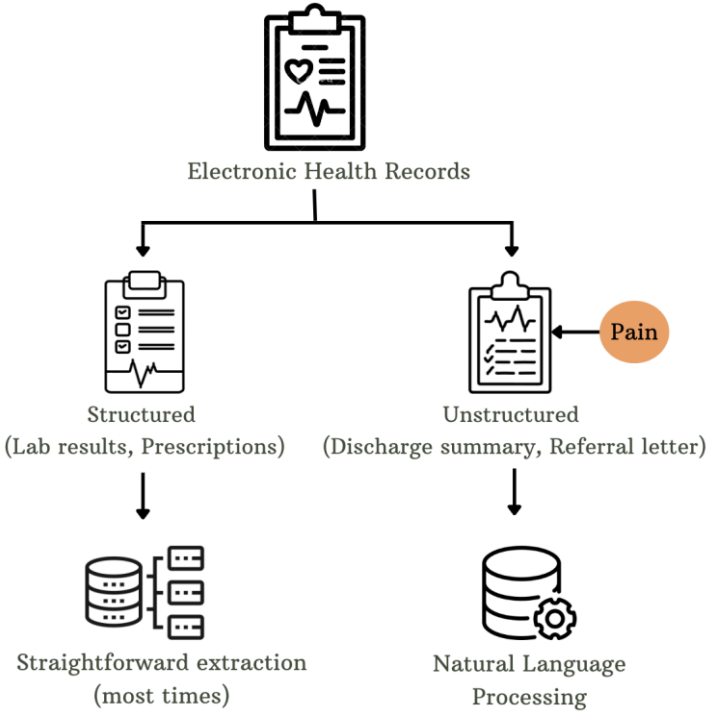
2 social media platforms (Twitter and Reddit)

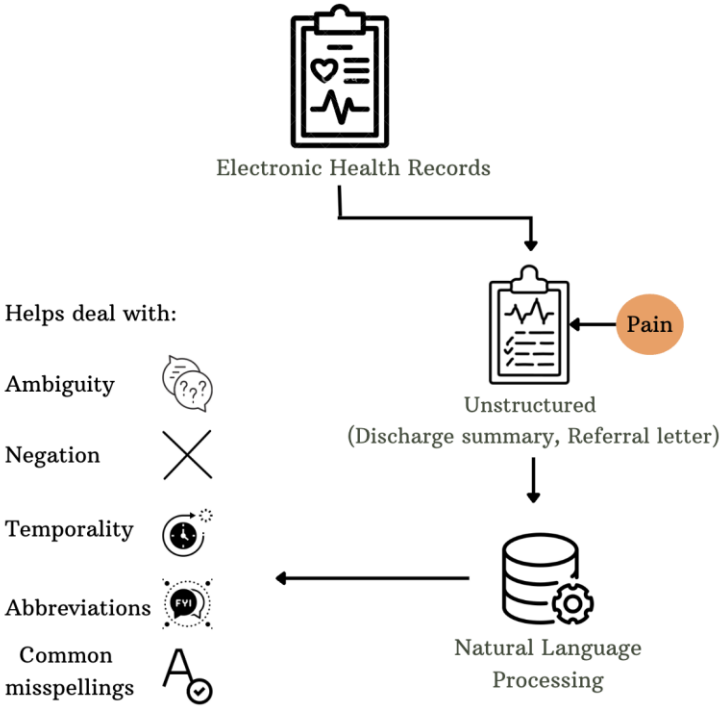


Electronic Health Records

- longitudinal compilations of electronic data pertaining to a person's medical history or healthcare
- increasingly used in research as they provide the opportunity to explore patient symptoms and findings
- unstructured narratives within clinical notes help provide context to the patient's pain

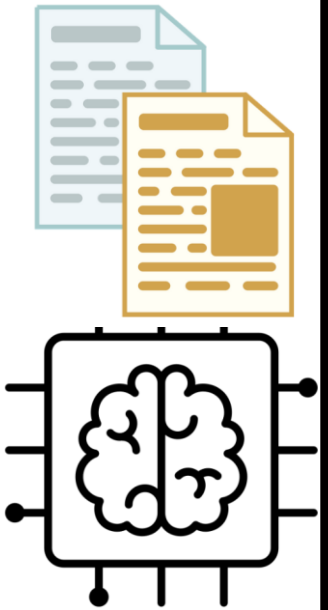



Electronic Health Records
Structured
(Lab results, Prescriptions)
Unstructured
(Discharge summary, Referral letter)
Pain
Straightforward extraction
(most times)
Natural Language
Processing


Electronic Health Records
Pain
Unstructured
(Discharge summary, Referral letter)
Natural Language
Processing
Helps deal with:
Ambiguity
Negation
Temporality
Abbreviations
Common
misspellings

Natural Language Processing

- a subfield of artificial intelligence used to leverage rich textual sources for information extraction and retrieval
- due to the ambiguous and sometimes metaphorical nature of how pain is used in communication, recent advances in NLP which adopt contextual and metaphorically informed methods could contribute to a deeper comprehension of how pain affects health and the utilization of healthcare resources in the treatment of pain

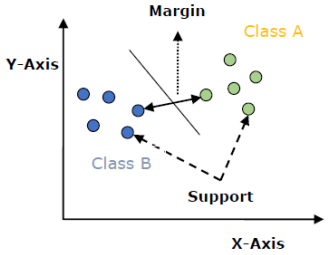


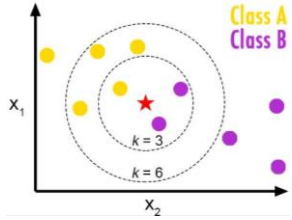
Natural Language Processing

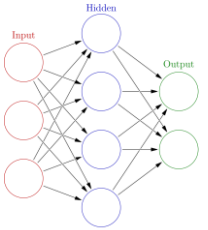
- a commonly used machine learning based NLP approach is text classification, in which labels are assigned to units of text (sentences/paragraphs/documents)
- within the healthcare domain, this can be used to classify presence or absence of features such as symptoms/diagnosis/smoking status and so on

Natural Language Processing

- commonly used classification algorithms include Support Vector Machines and K-Nearest Neighbours
- recent state of the art approaches use embedding models and transformer-based neural network architectures, such as BERT (unsupervised language models from large general corpora – transfer learning by fine tuning on smaller datasets)







Natural Language Processing

- BERT base model is trained on 3.3 billion words from the general domain (Wikipedia and BookCorpus)
- many healthcare domain related models are emerging such as PubMedBERT, BioBERT, ClinicalBERT, UmlsBERT and SAPBERT
- developed after recognition of the need for specialized models due to linguistic differences between general and biomedical text

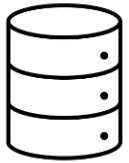
Aim of this work

- develop an NLP application for a sentence-level classification of mentions of pain within clinical text
- two BERT models were trained - bert_base and SAPBERT - and compared to two conventional models - support vector machines (SVM) and K-Nearest Neighbours (KNN)
- the best performing application will be used to extract relevant pain information from large EHR databases for use in further epidemiological studies and pain related research

Methods

Data Source

- an anonymised version of EHR data from The South London and Maudsley NHS Foundation Trust (SLaM), one of the largest mental healthcare organizations in Europe, is stored in the Clinical Record Interactive Search (CRIS) database
- CRIS contains over 30 million documents, averaging 90 documents per patient



Ethics and Data Access

- ethics approval for CRIS has been granted by Oxford C Research Ethics Committee (reference 18/SC/0372)
- research projects that use the CRIS database are reviewed and approved by a patient-led oversight committee
- an opt-out model is in place for service users, and is advertised in all publicity material and initiatives

Data Extraction

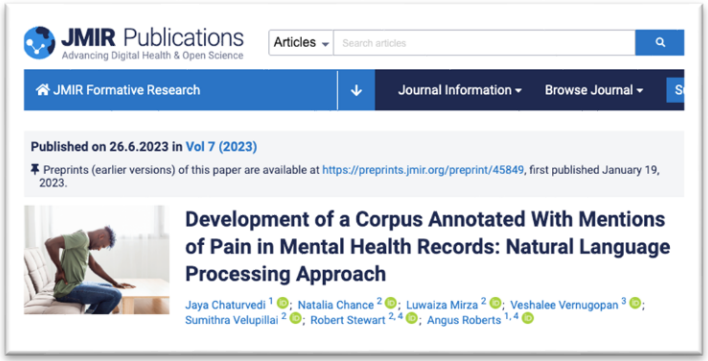
- pain can be described in numerous ways, using a variety of terms
- a lexicon of pain terms was developed from literature, biomedical ontologies, and word embedding models
- lexicon was used to help identify which documents in CRIS might be discussing pain

Data Extraction

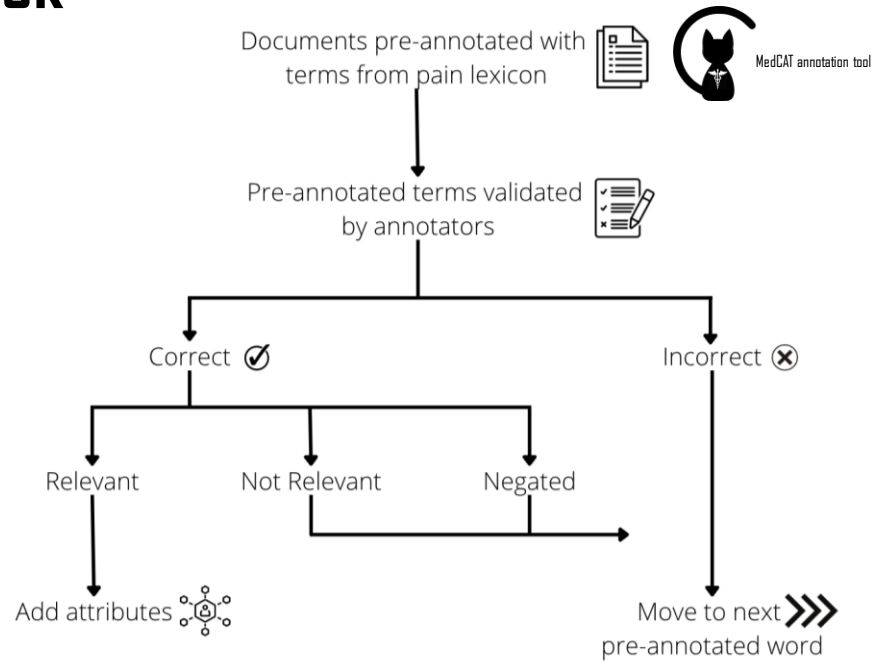
- documents containing pain terms were extracted using SQL
- no time or diagnosis filter was applied to the extraction

Annotation Task

- extracted documents were used to create a corpus of text discussing patient pain by labelling/annotating spans of text as being about pain or not
- 3 medical student annotators



Annotation Task



Annotation Task

SENTENCE	KEYWORD	CORRECT	RELEVANT	PAIN CHARACTER	ANATOMICAL PART	PAIN MANAGEMENT
She is in constant pain	pain	yes	yes	other	NA	NA
He used to suffer from severe headaches	ache	yes	yes	other	mentioned	NA
He burnt down his house	burn	No	NA	NA	NA	NA
She has back pain due to injury.	pain	yes	yes	NA	mentioned	NA
It is painful to think about the past	pain	yes	No	NA	NA	NA
She is taking pain medication because of a pulled muscle	pain	yes	yes	NA	mentioned	medication

NLP Application

- spans labelled by the annotators were used as gold standard training data for development of the NLP application (span of 200 characters before and after a pain mention)
- annotations were split into train/test/validation sets at a proportion of 80/10/10 respectively
- pre-processing: lowercasing, removal of stop words, and tokenisation

NLP Application

- 4 models were trained

Model	Tokenizer	Pre-processing	Other Parameters
1. Support Vector Machine 2. K-Nearest Neighbour	NLTK	Lowercase, stopword, white space and punctuation removal, lemmatize and tokenize	Tf-Idf vectorizer Default parameters from sklearn
3. BERT	bert_base_uncased	Tokenize Prepend sentence with special token [CLS] and append with special token [SEP] Pad and truncate sentence to max length (default is 511)	Epochs: 3 Batch size: 16 Optimizer: AdamW, learning rate 3e-5
4. SAPBERT	cambridgeltl/SapBERT-from-PubMedBERT-fulltext		Epochs: 4 Batch size: 16 Optimizer: AdamW, learning rate 2e-5

Results


Traffic Sources Overview
Direct Traffic
Search Engines
Referring Sites
Map Overlay
Visitors Overview
Visitors
2,958

Data Extraction

- 1,985 randomly selected documents from 723 patients were extracted that contained pain related keywords from the lexicon
- most common diagnosis codes for these extracted patients were Mood disorders (ICD10 chapters F30-39) (33% of patients)
- average of 8 annotations per patient

Annotations

- inter-annotator agreement (IAA) of 90% (Cohen's kappa 0.88) was achieved after four rounds (each round containing 200 documents) of triple annotations
- adjudication was carried out on documents that had disagreements
- annotators were given separate set of documents after good IAA was achieved

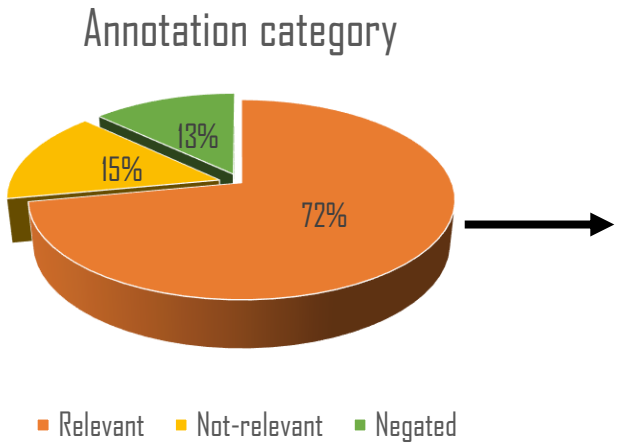
Annotations

- 5,644 gold standard annotations were obtained
- Final classes:

class 1: Relevant (72%)

class 0: Not-relevant (28%)

(Negated was combined with not-relevant)



Attribute within Relevant	Mentions
Anatomy	45%
Pain character	11%
Pain management	10%

Evaluation of the NLP applications

- a single GPU (Tesla T4) was used for training the models
- K-fold cross validation was carried out for evaluation of the models, and 95% confidence intervals were calculated

Model	Precision	Recall	F1-score (average from 10-fold cross validation)
Support Vector Machine	0.86 (0.83-0.88)	0.98 (0.97-0.99)	0.91 (0.90-0.93)
K-Nearest Neighbour	0.84 (0.81-0.87)	0.91 (0.89-0.93)	0.87 (0.85-0.89)
BERT	0.96 (0.94-0.97)	0.98 (0.97-0.99)	0.97 (0.96-0.98)
SAPBERT	**0.98 (0.97-0.99)**	**0.99 (0.98-0.99)**	**0.98 (0.98-0.99)**

Error Analysis

- common disagreements during annotation process were when an instance could be interpreted as physical or metaphorical, such as "..causing him pain",
- and hypothetical mentions such as

"..she feared the pain" and "?migraine"

Error Analysis

- false positives with BERT models were instances such as

"..wishing to project his pain on others",

"..headaches were a common adverse effect reported by the trial"

Error Analysis

- false negatives with BERT models were instances such as

"denying symptoms other than stomach ache",

"..if pain increases",

"bruised arm is painful, no other worrying findings"

Conclusion

- pain is very subjective and ambiguous in its description, making it hard for clinicians to code pain within structured fields of EHRs
- free-text fields within the EHR provides clinicians the flexibility to describe the pain in the patient's own words or based on their interpretations

Conclusion

- ambiguous nature of pain was highlighted during this project, especially during the annotation process where it took multiple rounds for three clinically trained annotators to agree on the meanings and interpretations of the pain mentions

Conclusion

- this is a novel approach towards extracting information about pain from mental health records, leveraging the unstructured clinical notes to identify patients with relevant mentions of pain
- such cohorts of patients can further be used in epidemiological and other pain related research with more confidence in the actual occurrence of pain when mentioned in the text

Acknowledgements

- Supervisors
  - Dr Angus Roberts
  - Prof Robert Stewart
  - Dr Sumithra Velupillai
- Medical student annotators
  - Natalia Chance
  - Veshalee Vernugopan
  - Luwaiza Mirza

Thank you!

Questions?