Identifying causal relationships in diabetes related tweets

Adrian Ahne, PhD Student

Datacraft, 25/11/2021

Inserm
epiconcept
smart health
université PARIS-SACLAY

Context

Research focus - World Diabetes Distress Study (WDDS)

WDDS =

Social media data

+

Clinical data

Smartphone applications

Connected objects

WDDS objective: Better understand the burden of diabetes and diabetes distress using real-world data

Diabetes Distress →

Emotional burden	**Stress**	**Associated with day-to-day disease**
Fatigue	**Anxiety**	Diet, Physical activity, Fear of hypoglycemia, Fear of complications, Work, Family support,...

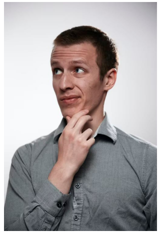
Digital epidemiology

- > 30 M diabetes-related tweets (e.g. insulin, hypoglycemia, blood sugar, #T1D) since 2017

- Data driven research (instead of hypothesis driven)

- First publication:
  - Identification of diabetes (distress) patterns, concerns and emotions in tweets
  - Machine pipeline to clean data
  - => *FastText* embeddings + Kmeans

- What else to do with all those tweets?



epiconcept smart health

- First idea: Identify risk factors for Hypoglycemia (low blood sugar) based on Tweets
  - cause: risk factors
  - effect: hypoglycemia

- But too less tweets referring to Hypoglycemia

Identification of cause and associated effect relationships

in general in our diabetes related tweets

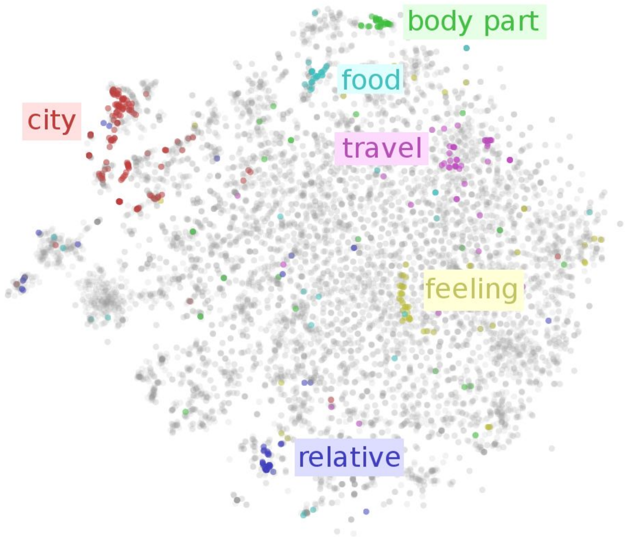
Warning!

This has nothing to do with causal inference (in epidemiology)

Transformers / BERT

Key breakthroughs driving the boom in NLP

- Ability to generate meaningful fixed-size vector representations: *word embeddings / dense vectors*
  - Ex.: word2vec, FastText, Glove



- Ability to generate context-aware word sentence representations using the *transformer* architecture
  - Ex.: BERT based models use the *transformer*

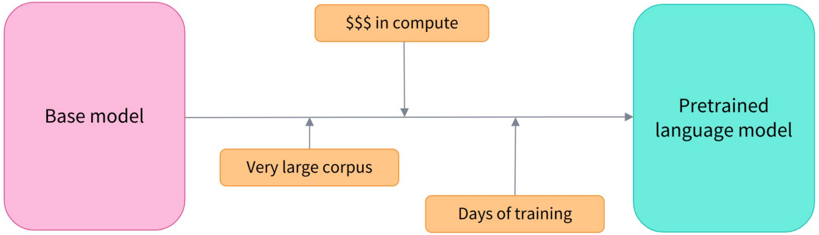
Ex.: “The money on my bank account”, “At the river bank”

Transformers calculates two different embeddings for bank

Transfer learning

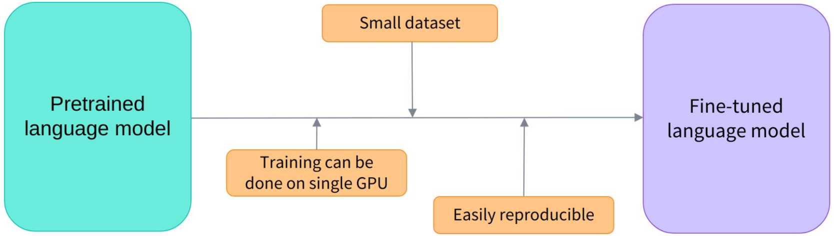
- Neural networks require massive training data

- Labeled data is scarce or expensive in many real-world applications



⇨ **Transfer learning**

- leverages knowledge in similar domains
- reduces training time



Pictures taken from Huggingface

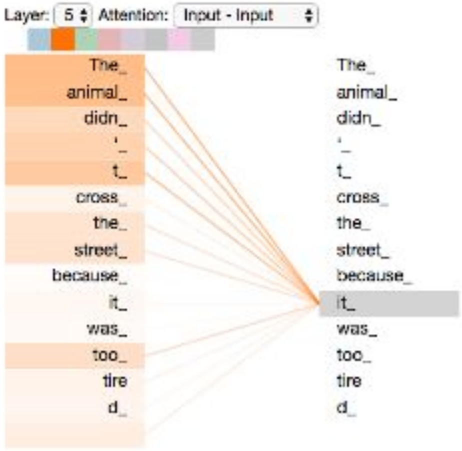
Transformers

- **Previously**: Recurrent neural networks (e.g. LSTMs) dominated sequence-to-sequence models

- **Problem**:
  - Weak in modeling long sequences
  - Difficult to parallelize

- **Solution:** *Attention* mechanism
  - weights different input words
  - allows the model to focus on the relevant parts of the input

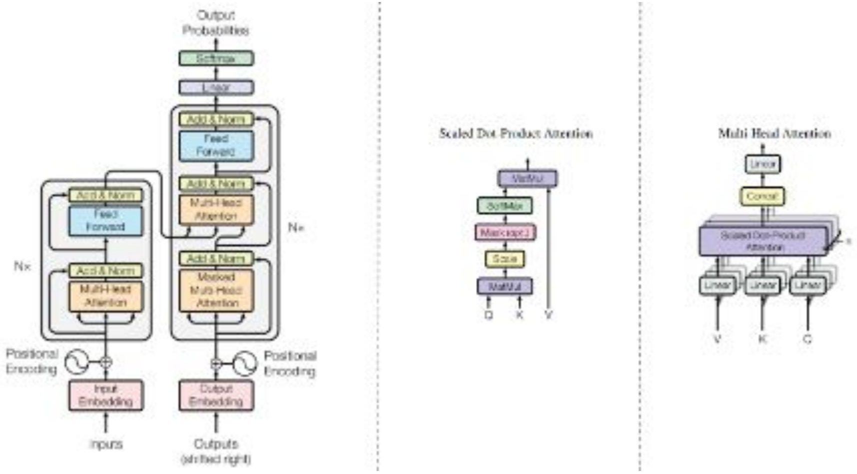
$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$



  - Great visualisation:
    - https://jalammar.github.io/illustrated-transformer/

Transformers

- encoder-decoder mechanism
- scaled dot-product attention
- multi-head attention
    - “looks at a text sequence from different angles”
- Sequences are processed in parallel
    - words are handled simultaneously rather than word-by-word => parallelisation

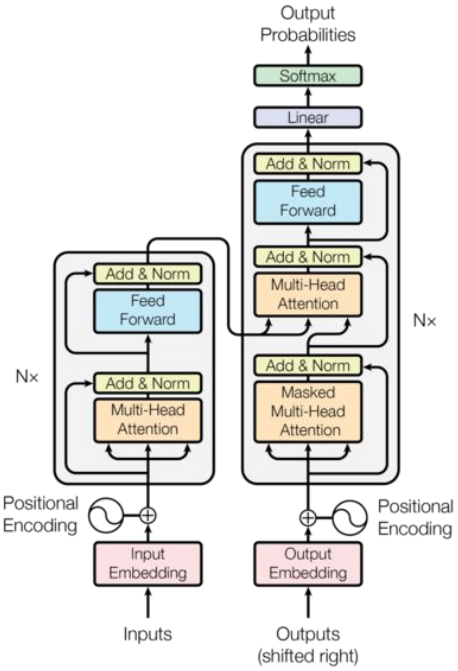


epiconcept
smart health

Transformers

- Encoder (left): Receives input and builds features (vector)
- Decoder (right): Uses features to generate target sequence

- Encoder-only models:
  - Good for tasks that require understanding of the input
  - e.g. sentence classification, NER
  - ALBERT, BERT, DistilBERT
- Decoder-only models:
  - Good for generative tasks, such as text generation
  - GPT, GPT-2
- Encoder-decoder models:
  - Good for generative models that require an input
  - e.g. translation or summarisation
  - BART, T5



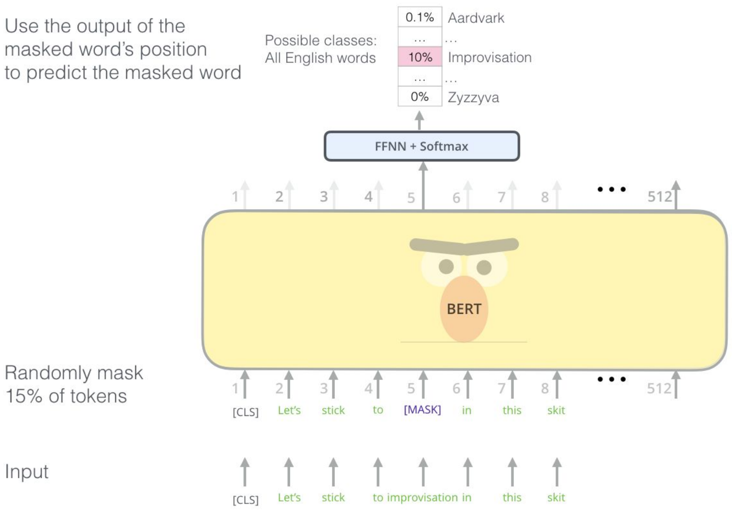
Bidirectional Encoder Representations from Transformers (BERT)

- 2 versions:
  - BERT_base : 12 transformer encoder layers; 110 million parameters; dim 768
  - BERT_large: 24 transformer encoder layers; 340 million parameters; dim 1024

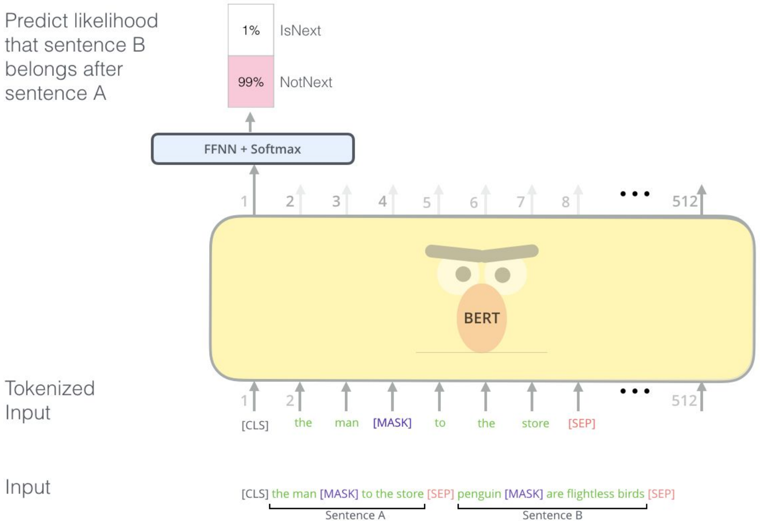
- Typically, transfer learning:
  - use pretrained BERT model
  - fine-tune on task-specific data

- Training objectives:
  - Mask language model
  - Next sentence predictions

Masked language modeling



Next sentence prediction



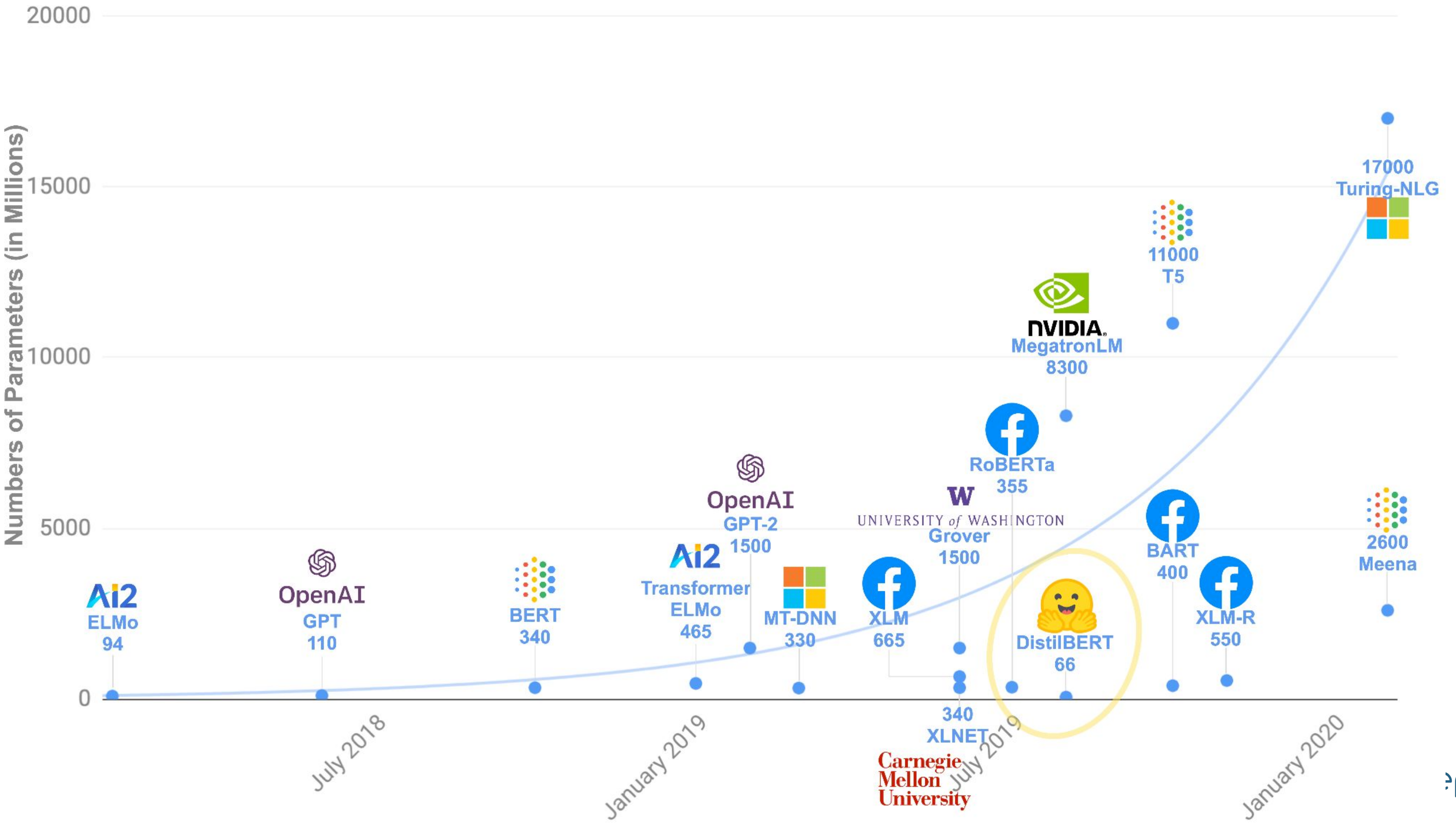
*https://jalammar.github.io/illustrated-bert/

Transformers

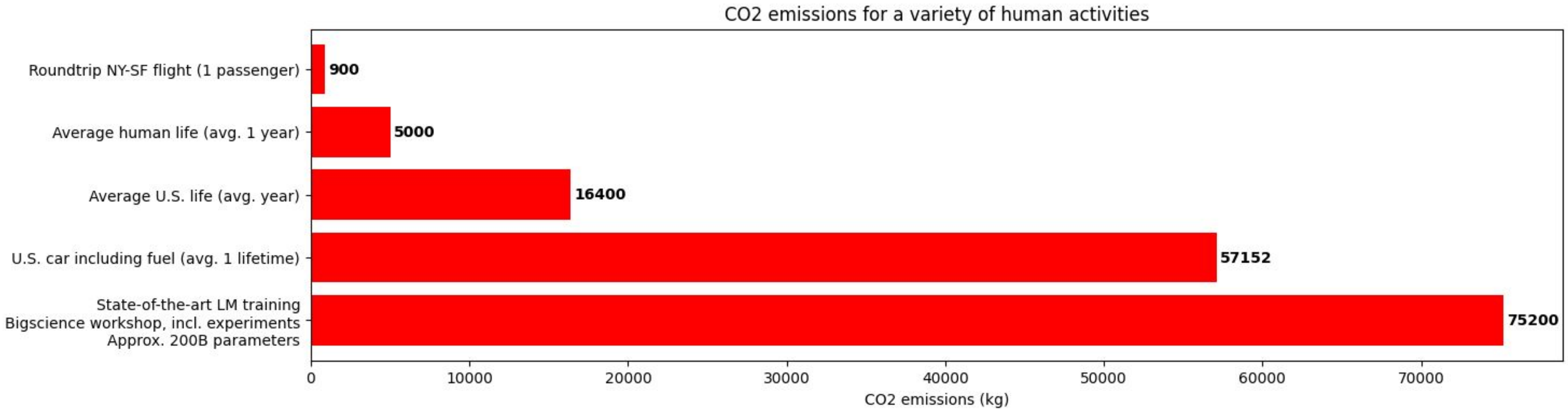
- Those models are everywhere -> Google search, Auto-completitition, etc.

- Transformer models have been trained as *language models*:
  - on a huge amount of raw text
  - self-supervised learning (no human labeling needed)
- Transfer learning paradigm:
  - Pre-training
  - Fine-tuning

Transformer models are huge (!)



CO2 emissions



Conditional random field (CRF)

Standard method for named entity recognition tasks prior to BERT

$$p(y \mid x, \theta) = \frac{1}{Z(x)} \exp\left\{ \sum_{t=1}^{T} \sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, x_t) \right\}$$

$f_2(y_t, y_{t-1}, x_t) = 1$ if $x_t = capitalized;$ 0 otherwise

Objective

Objective

Identification of both explicit and implicit multi-word cause and corresponding effect relations in diabetes related tweets

Causality in text

Causal extraction in text

Two types of causality:

- *Explicit* causality: explicit causal link
  - e.g.: so, hence, because, caused, resulted in, or conditional (if …, then...), etc.
  - e.g.: “Diabetes causes hypoglycaemia” => causal link: “causes”
  - Large majority of approaches tackle *explicit causality*

- *Implicit causality:* no explicit causal link
  - e.g. “Cannot sleep **#insomnia, #overthinking**”
  - e.g. ”I think the **greater danger** is from the **unelected justices** than from the elected Congress and the elected president.”

Causal extraction in text

Two types of approaches:

Rule-based approaches	Machine learning approaches
- hand-coded linguistic and syntactic rules (pattern matching)	- train algorithm based on small training set
- mostly tested on text from a similar domain	- works on texts containing sentences from different domains
- meant to work only for a specific type of text	- generalizes better
- mainly focuses on *explicit* causality	- tackles both *explicit* and *implicit* causality

Examples

- Typical example: Rule-based + focus on *explicit* causality

TABLE I. RULE SET TO EXTRACT CAUSAL RELATIONS FROM TWEETS.

#	Causal relation types	Dependency rules	Examples
1	A (noun) caused B	{}=subj < subj ({ + Clausal verb + }=target >dobj {}=cause)	Stress causes insomnia
2	A (verb-ing) caused B	{}=subj < csubj ({ + Clausal verb + }=target >dobj {}=cause)	Over thinking can increase anxiety and cause insomnia.
3	B was caused by A	{}=ncsubjpass<nsubjpass({ + Clausal verb + }=target >/nmod:agent/ {}=cause)	My insomnia was caused by stress.
4	A is a reason of B	Clausal noun + < nsubj ({}=target > /nmod:of/ {}=cause)	Stress is a reason of my insomnia
5	B was caused by A (verb-ing)	{}=nsubj< nsubjpass ({}=target > /advcl:by/ + Clausal noun)	Insomnia was caused by overthinking
6	A results "in/to/from" B	Clausal verb + <[nc]subj ({}=target> /nmod:(to|in|from)/ {}=cause)	Stress results to insomnia.

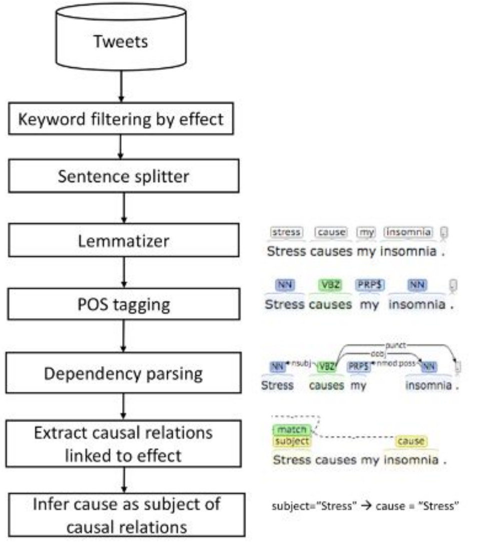


Fig. 1. A general framework to extract causal relations from Twitter messages.

*Doan et al. Extracting health-related causality from twitter messages using natural language processing: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0785-0

epiconcept smart health

Examples

- Typical example: Rule-based + focus on *explicit* causality

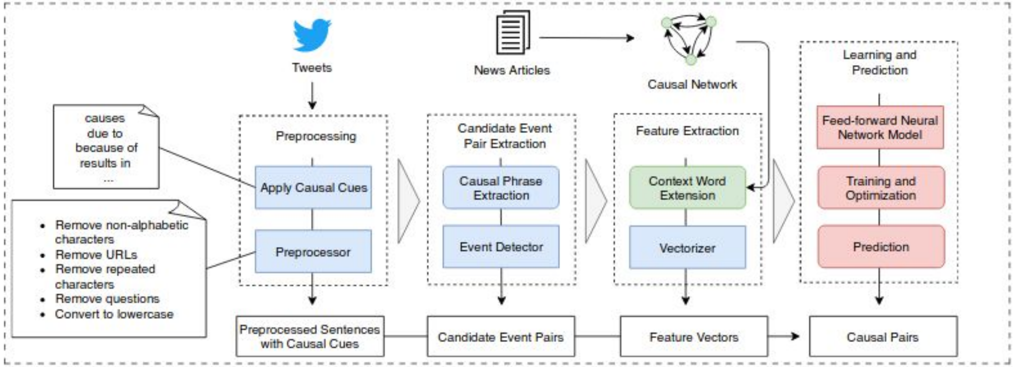


Fig. 2: An overview of the proposed method

TABLE I: Representation of Events

Sentences	Events
Storm **hits** Gold Coast	hit (storm, coast, gold)
Mike **crashed** his car in Gold Coast	crash (mike, car, coast, gold)
Heavy traffic **jam** in Gold Coast today	jam (traffic, today, coast, gold)
A **disruption** in bus service in Gold Coast due to **lack** of communication	disruption (service, bus, coast, gold) lack (communication, organizer, translink)

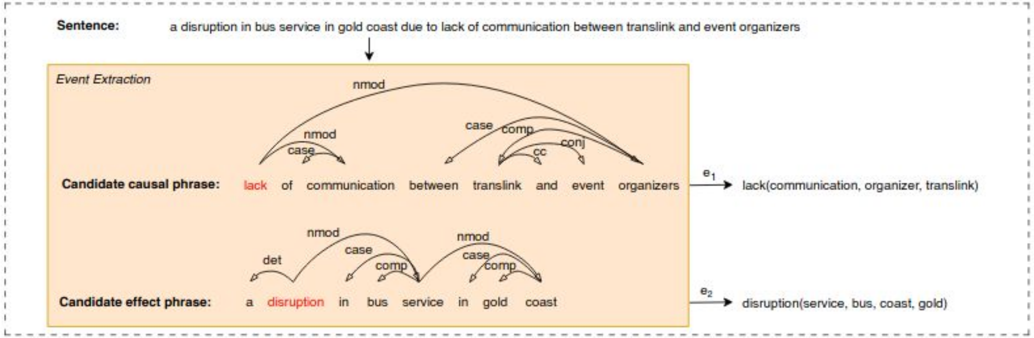


Fig. 3: An example of event pair extraction from a sentence

*Kayesh et al. On event causality detection in tweets, 2019, https://arxiv.org/pdf/1901.03526.pdf

epiconcept
smart health

Examples: Causal-BERT

Khetan et al. *Causal-BERT : Language models for causality detection between events expressed in text*

https://arxiv.org/pdf/2012.05453v1.pdf

- Explicit + implicit causality
- Machine learning approach:
    - Fine-tuning BERT models
- Data: Semeval + Adverse drug effect

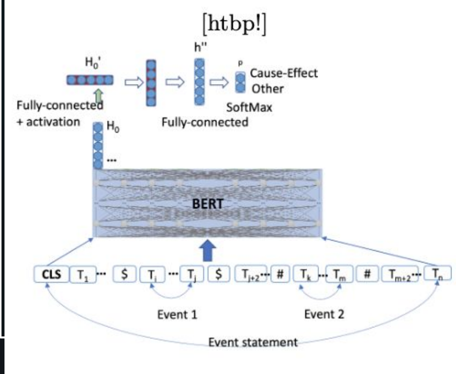


Fig. 2. C-BERT

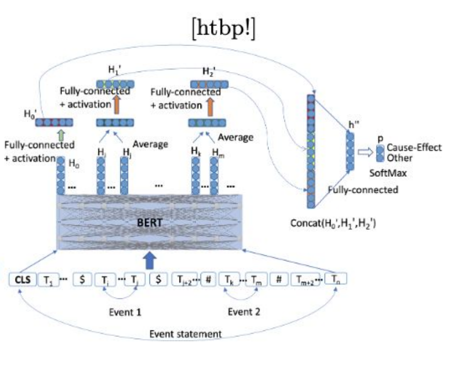
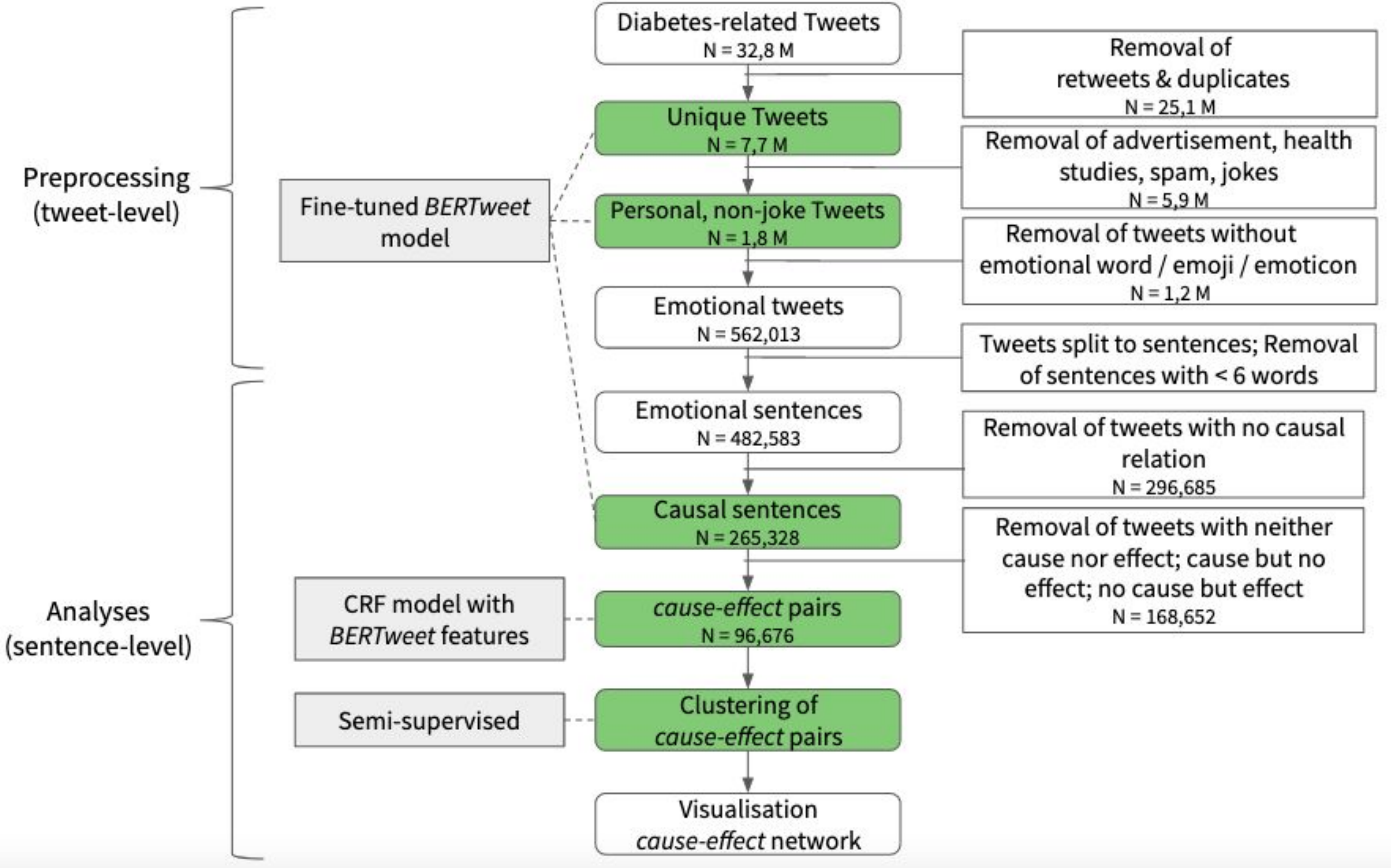


Fig. 3. Event aware C-BERT

Methods/Results

Workflow


Preprocessing (tweet-level)
Analyses (sentence-level)
Fine-tuned *BERTweet* model
CRF model with *BERTweet* features
Semi-supervised
Diabetes-related Tweets N = 32,8 M
Unique Tweets N = 7,7 M
Personal, non-joke Tweets N = 1,8 M
Emotional tweets N = 562,013
Emotional sentences N = 482,583
Causal sentences N = 265,328
cause-effect pairs N = 96,676
Clustering of *cause-effect* pairs
Visualisation *cause-effect* network
Removal of retweets & duplicates N = 25,1 M
Removal of advertisement, health studies, spam, jokes N = 5,9 M
Removal of tweets without emotional word / emoji / emoticon N = 1,2 M
Tweets split to sentences; Removal of sentences with < 6 words
Removal of tweets with no causal relation N = 296,685
Removal of tweets with neither cause nor effect; cause but no effect; no cause but effect N = 168,652

Preprocessing

- **Unique tweets:** Removal of retweets, duplicates (be careful with chatbots)

- **Personal classifier:** Fine-tune *BERTweet* to distinguish
  - *personal content*: feelings, concerns, opinions, etc.
  - *institutional content*: advertisment, health studies
  - 4,403 labeled tweets, adjust for concept drift
  - Acc: 91,2%   Precision: 86,2%   Recall: 90,9%   F1: 88,5%

- **Joke classifier**: Fine-tuned *BERTweet*
  - 1,648 tweets, adjusted for concept drift
  - Acc: 90,4%   Precision: 78,5%   Recall: 90,8%   F1: 84,2%

BERTweet

850 M english tweets (~80GB)

Examples

- "There is too much sugar today. Viewers are about to get diabetes🙆😂😂😂"

- Jah know. Moms just say that I am going to get diabetes because I am always on the phone

epiconcept smart health

Emotional tweets

- Psychologue Parrott : **joy, love, surprise, anger, fear, sadness**
- Diabetes distress-related keywords from diabetes distress questionnaires (PAID, DDS)

- Interest in diabetes distress & emotions

→ Filter tweets containing an *emotional* element

→ **563,013 personal, non-joke tweets with emotional element**

Annotation of cause-effect dataset

Text	Event A (Disease)	Event B (Risk factor / consequence)	Causality (0=no; 1=A->B; 2=B->A)
I missed my workout again, I will certainly get diabetes …….	diabetes	missed my workout	2
Having anxiety before I even start this glucose test because of how sick it made me last time	glucose test	anxiety;sick	1
My family has always had a history of diabetes	diabetes	-	0

Option 1:

+ : easier for model, as events are well separated
- : difficult to define events; to categorise words into both events

Text	Cause	Effect	Causality (0=no; 1=yes)
I missed my workout again, I will certainly get diabetes …….	missed my workout	diabetes	1
diabetes makes me feel sick	diabetes	sick	1

Option 2:

+ : easier to label
- : more difficult to learn, as events can be both *cause* and *effect*

epiconcept
smart health

Annotation of cause-effect dataset

- Labeling *cause-effect* relationships is difficult => Annotation guidelines

Annotation of cause-effect dataset

- Labeling *cause-effect* relationships is difficult => Annotation guidelines
    - Non-diabetes or non-diabetes distress related relationships are <u>ignored</u>
        - “This **virus** can **kill** me as a diabetes patient”

Annotation of cause-effect dataset

- Labeling *cause-effect* relationships is difficult => Annotation guidelines
    - Non-diabetes or non-diabetes distress related relationships are <u>ignored</u>
        - “This **virus** can **kill** me as a diabetes patient”
    - Implicit relations are <u>included</u>
        - “I was sent to fix a guy with **low blood sugar #diabetes**”

Annotation of cause-effect dataset

- Labeling *cause-effect* relationships is difficult => Annotation guidelines
  - Non-diabetes or non-diabetes distress related relationships are ignored
    - “This **virus** can **kill** me as a diabetes patient”
  - Implicit relations are included
    - “I was sent to fix a guy with **low blood sugar #diabetes**”
  - Unclear cause-effect relationships are ignored
    - “My dad has a **diabetes**, cancer, heart problems, and a **weak immune system**”

Annotation of cause-effect dataset

- Labeling *cause-effect* relationships is difficult => Annotation guidelines
  - Non-diabetes or non-diabetes distress related relationships are ignored
    - “This **virus** can **kill** me as a diabetes patient”
  - Implicit relations are included
    - “I was sent to fix a guy with **low blood sugar #diabetes**”
  - Unclear cause-effect relationships are ignored
    - “My dad has a **diabetes**, cancer, heart problems, and a **weak immune system**”
  - Chaining cause-effect relationships: A->B->C => label closest relationship to diabetes
    - “Not sure if I’ve been up since 3:30 for Titanic or because my **anxiety** over my **glucose test** is keeping me up”

Annotation of cause-effect dataset

- Labeling *cause-effect* relationships is difficult => Annotation guidelines
  - Non-diabetes or non-diabetes distress related relationships are <u>ignored</u>
    - “This **virus** can **kill** me as a diabetes patient”
  - Implicit relations are <u>included</u>
    - “I was sent to fix a guy with **low blood sugar #diabetes**”
  - Unclear cause-effect relationships are <u>ignored</u>
    - “My dad has a **diabetes**, cancer, heart problems, and a **weak immune system**”
  - Chaining cause-effect relationships: A->B->C => label closest relationship to diabetes
    - “Not sure if I’ve been up since 3:30 for Titanic or because my **anxiety** over my **glucose test** is keeping me up”
  - Label tweet as negation if it alters the meaning of the tweet
    - “I **cannot afford my insulin** and will **die**” -> no negation
    - “She ‘gave’ herself diabetes by not doing what her doctor told her: Lose weight” -> negation

Annotation of cause-effect dataset

- Chose 5000 random tweets to label

- **Interrater reliability:**
  - First 500 tweets were labeled by two researchers

- Disagreements were discussed and one researchers continued to label 4,500 tweets

Causal models

Two steps to identify *cause-effect pairs*

1) **Causal sentence model:** Train model to detect if a sentence contains causal information (*causal sentences*)
   a) Binary sentence classification

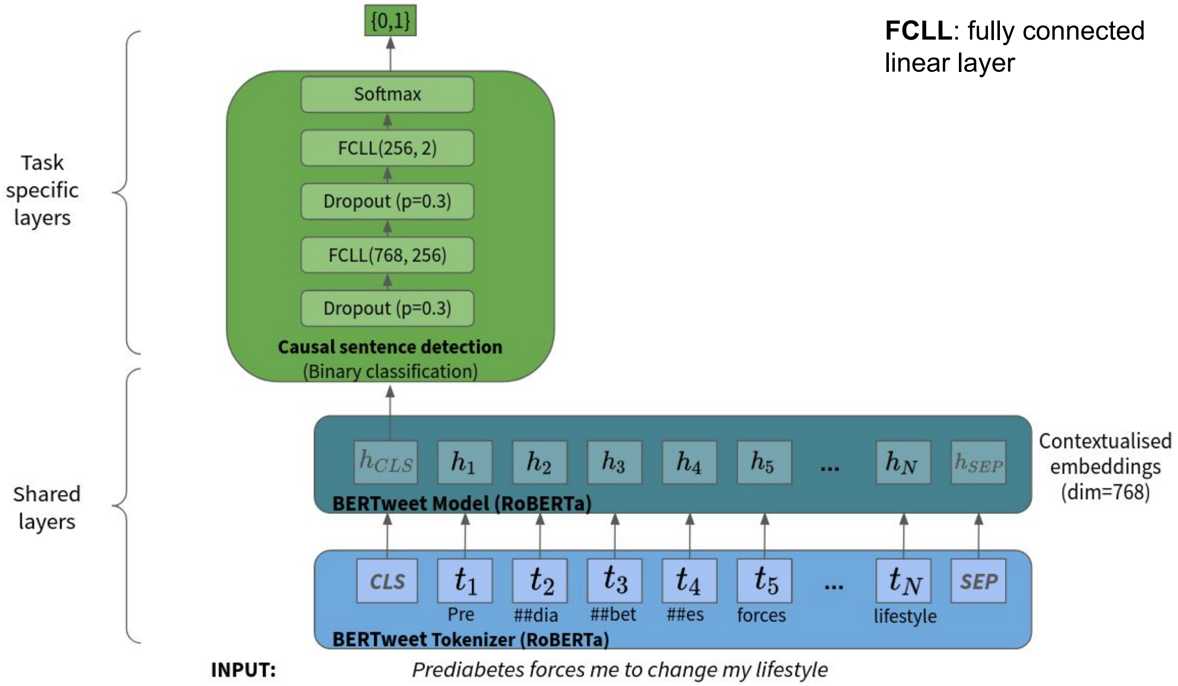
2) **Cause-effect model:** Train model to extract multi-word *cause* and associated *effect* in *causal sentences*
   a) Entity recognition task

Hypotheses:

- *cause-effect* occur only in the same sentence
- sentence with < 6 words were removed due to a lack of context

Causal sentence classifier

- 5000 labeled tweets
  - 7,218 non-causal sentences
  - 1,017 causal sentences
- Class weights added to cross entropy loss function
- Train: 90%
  - 20% validation
- Test: 10%



FCLL: fully connected linear layer

Causal sentence classifier

- Trained model => was ok
- Trained cause-effect model => very bad (only ~1000 training examples)
- Tried/discussed:
  - Different class weights
  - Different loss functions (cross entropy, DICE)
  - simplified from 5 to 3 class tags
  - checked if there are larger pre-trained models (dimension)
  - add more information to the model training, e.g. POS tags

What would you do in such a situation?

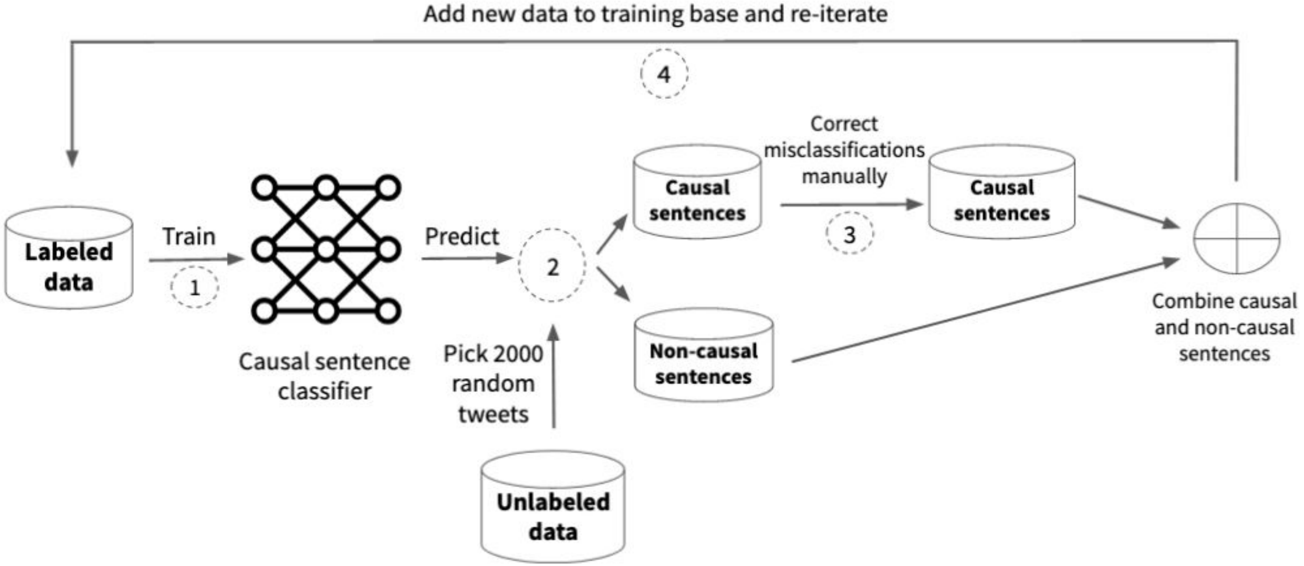
Causal sentence classifier

- Trained model => was ok
- Trained cause-effect model => very bad (only ~1000 training examples)
- Tried/discussed:
  - Different model weights
  - Different loss functions (cross entropy, DICE)
  - simplified from 5 to 3 class tags
  - checked if there are larger pre-trained models (dimension)
  - add more information to the model training, e.g. POS tags
- <u>Problem: </u>too less training data

Data augmentation through active learning

- First predictions of *cause-effect* model very bad

  (1,017 causal sentences)



- Need more data

- Active learning to efficiently increase training data

  (4 iterations) → 2,118 causal sentences

epiconcept
smart health

Cause-effect model - Tagging scheme

- Typically for entity recognition task (e.g. NER): BIO (Beginning, Inside, Outside) tagging

- 5 classes:
  - "B-C" (begin cause), "I-C" (inside cause)
  - "B-E" (begin effect), "I-E" (inside effect)
  - "O" (outside)

- Ex. :

Prediabetes	*forces*	*me*	*to*	*change*	*my*	*lifestyle*
B-C	*O*	*O*	*O*	*B-E*	*I-E*	*I-E*

epiconcept
smart health

Cause-effect model - Tagging scheme

- Typically for entity recognition task (e.g. NER): BIO (Beginning, Inside, Outside) tagging

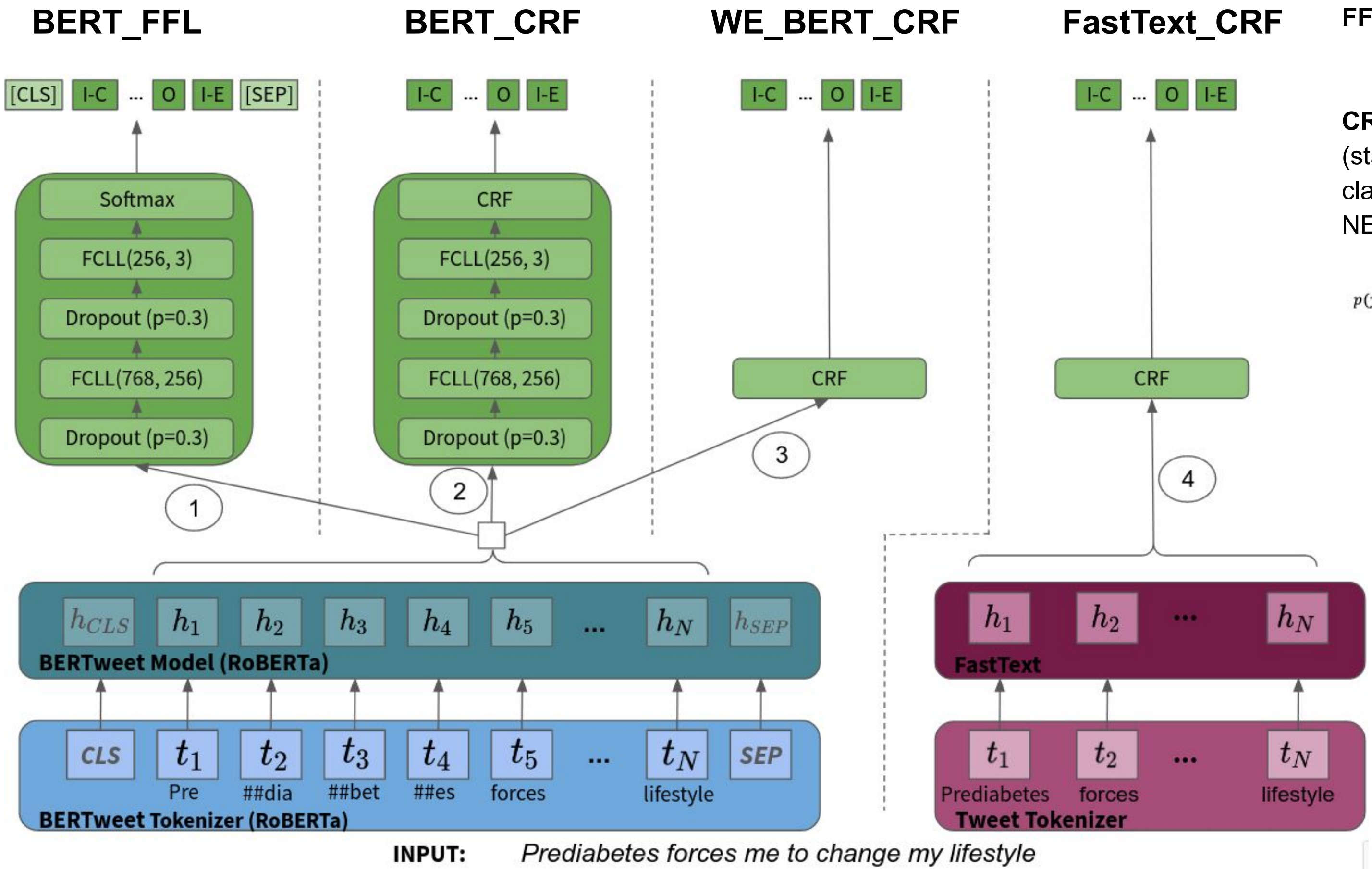
- 5 classes:
  - “B-C” (~~begin~~ cause), “I-C” (inside cause)
  - “B-E” (~~begin~~ effect), “I-E” (inside effect)
  - “O” (outside)

3 Classes simplifies model training

- Ex. :

Prediabetes	*forces*	*me*	*to*	*change*	*my*	*lifestyle*
B-C	*O*	*O*	*O*	*B-E*	*I-E*	*I-E*
I-C	*O*	*O*	*O*	*I-E*	*I-E*	*I-E*

Cause-effect model



FFCL: fully connected linear layer

CRF: conditional random field (standard method for sequential classification methods such as NER)

$$p(y \mid x, \theta) = \frac{1}{Z(x)} \exp\left\{\sum_{t=1}^{T}\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, x_t)\right\}$$

Clustering of *causes* and *effects*

- How to best regroup/cluster and so present/visualise the identified *causes* and *effects?*

Clustering of *causes* and *effects*

- How to best regroup/cluster and so present/visualise the identified *causes* and *effects?*

- Semi-supervised approach:
  - Cluster manually 1000 random causes and effects
  - Compare each *cause/effect* with each cluster using cosine similarity:
    - if sim( ) > 0.55 => associate *cause/effect* to this cluster
    - if sim () <= 0.55 => create new cluster
  - led to 1,751 clusters of *causes/effects*

Cluster name	Synonyms
diabetes	diabetic, #diabetes
T1D	type 1 diabetes
rationing insulin	shortage insulin
insulin price	cost of insulin
retinopathy	lost vision
neuropathy	feet amputation
covid	corona, covid-19
insurance	pharma, Medicare
medication	meds, drug, pill
OGTT	glucose test, 3h drink
stress	mood disorder,
fatigue	exhausted, tired, no power
hypoglycemia	hypo, go low, low blood sugar
overweight	obese, weight gain
physical activity	exercising, sport, walking
insomnia	can't sleep
family	brother, mum, aunt

- To remove noisy (mispredictions) *causes/effect* clusters, only keep clusters with at least 10 cause/effect occurrences in sentences

⇨ 763 clusters in total

Clustering of *causes* and *effects*

- How to best regroup/cluster and so present/visualise the identified *causes* and *effects?*

- Semi-supervised approach:
  - Cluster manually 1000 random causes and effects
  - Compare each *cause/effect* with each cluster using cosine similarity:
    - if sim( ) > 0.55 => associate *cause/effect* to this cluster
    - if sim () <= 0.55 => create new cluster
  - led to 1,751 clusters of *causes/effects*

Cluster name	Synonyms
diabetes	diabetic, #diabetes
T1D	type 1 diabetes
rationing insulin	shortage insulin
insulin price	cost of insulin
retinopathy	lost vision
neuropathy	feet amputation
covid	corona, covid-19
insurance	pharma, Medicare
medication	meds, drug, pill
OGTT	glucose test, 3h drink
stress	mood disorder,
fatigue	exhausted, tired, no power
hypoglycemia	hypo, go low, low blood sugar
overweight	obese, weight gain
physical activity	exercising, sport, walking
insomnia	can't sleep
family	brother, mum, aunt

- To remove noisy (mispredictions) *causes/effect* clusters, only keep clusters with at least 10 cause/effect occurrences in sentences

⇨ 763 clusters in total ⇨ Interactive network visualisation

Results

Causal sentence performance

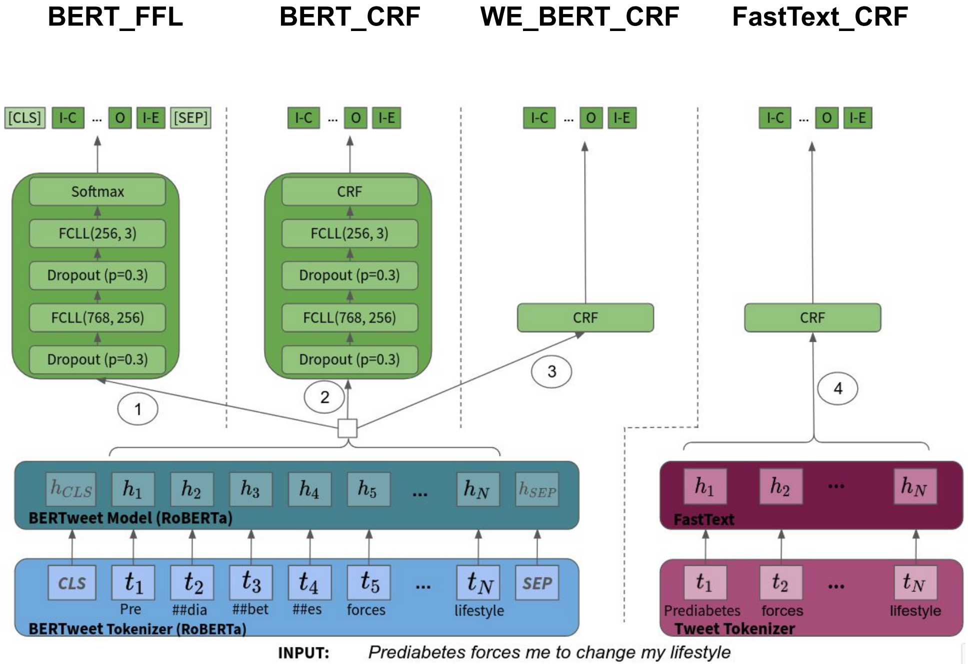
- Oscillating performances in active learning

Round	N° sent. train	N° sent. test	Accuracy	Precision	Recall	F1-Score
0	6,024	837	64.5	58.0	67.4	53.8
1	7,536	1,047	67.7	61.2	71.6	58.4
2	8,804	1,223	67.7	60.3	66.3	56.3
3	10,284	1,429	65.4	60.0	68.8	54.8
4	11,861	1,648	71.0	61.0	67.8	58.3

Table 5.2: Performance measures (macro) for each round of more training data

epiconcept
smart health

Cause-effect classification performance



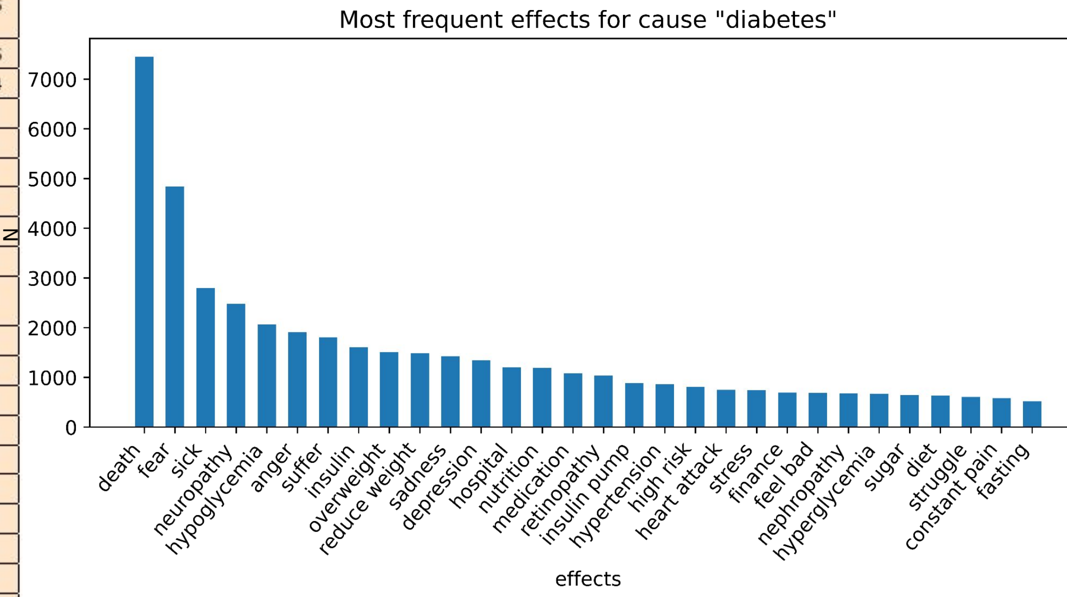
Models		Prec	Rec	F1
BERT_FFL	I-C	0.48	0.46	0.47
	I-E	0.20	0.48	0.29
	O	0.91	0.77	0.83
	macro	0.53	0.57	0.53
BERT_CRF	I-C	0.59	0.20	0.29
	I-E	0.0	0.0	0.0
	O	0.83	0.99	0.90
	macro	0.47	0.39	0.40
WE_BERT_CRF	I-C	0.63	0.61	0.62
	I-E	0.49	0.49	0.49
	O	0.93	0.93	0.93
	macro	0.68	0.68	**0.68**
FastText_CRF	I-C	0.59	0.57	0.58
	I-E	0.45	0.38	0.41
	O	0.92	0.94	0.93
	macro	0.65	0.63	0.64

Table 5.3: Performance measures for each of the four architectures

Cause-effect description

Most frequent clusters		
Parent cluster	**cluster**	**N**
Diabetes	diabetes	66,775
Death	death	16,989
Insulin	insulin	14,148
Diabetes	type 1 diabetes	11,693
Emotions	fear	10,160
Glycemic variability	hypoglycemia	9,547
Symptoms	sick	6,549
Nutrition	overweight	5,186
Diabetes	type 2 diabetes	4,909
Complications & comorbidities	neuropathy	4,481
Healthcare system	medication	4,389
Diabetes Technology	insulin pump	4,307
Nutrition	nutrition	4,230
Emotions	anger	4,149
Health	OGTT*	4,053
Blood pressure	hypertension	3,782
Healthcare system	finance	3,767
Nutrition	reduce weight	3,589
Insulin	unable to afford insulin	3,381
Nutrition	diet	3,325
Emotions	sadness	3,153
Glycemic variability	hyperglycemia	3,144
Diabetes	suffer	3,132
Diabetes Distress	depression	2,810
Healthcare system	hospital	2,721
Diabetes Distress	stress	2,681
Nutrition	sugar	2,369

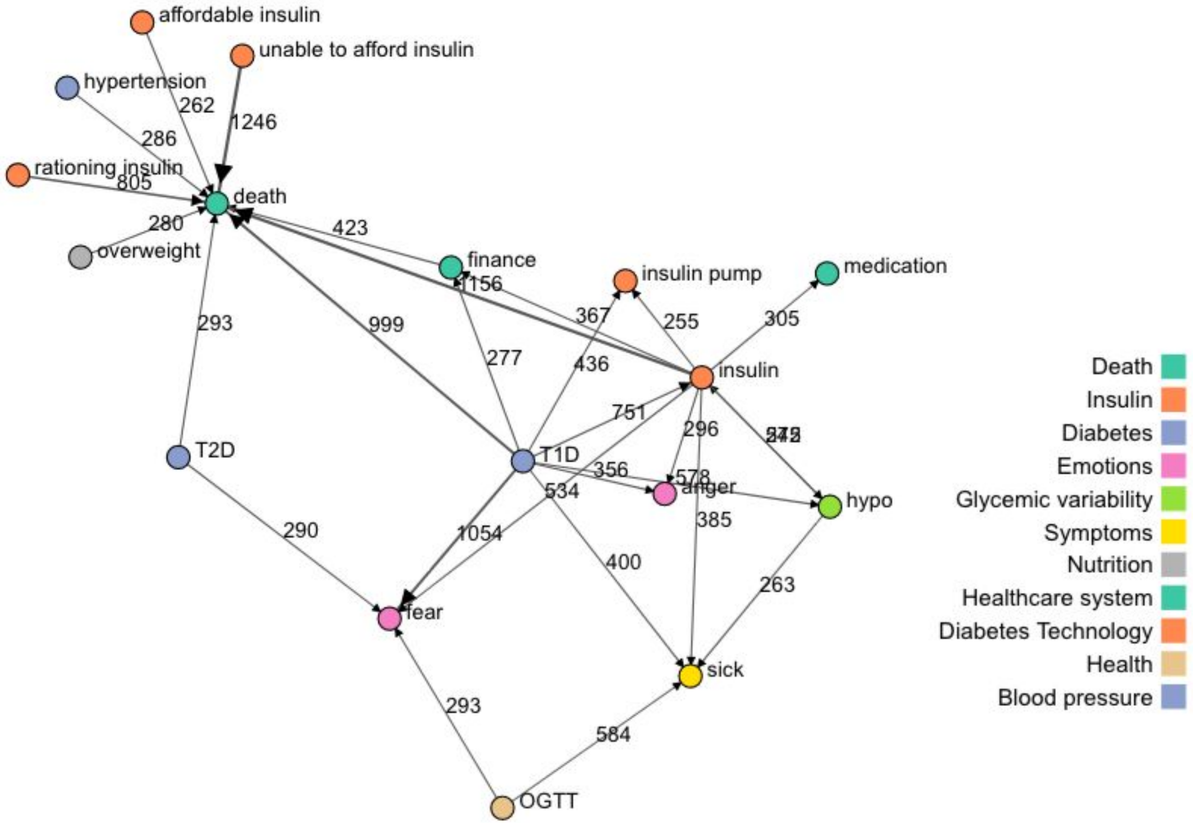
Most frequent cause-effect-associations (excluding cluster "diabetes")		
cause	**effect**	**N**
unable to afford insulin	death	1,246
insulin	death	1,156
type 1 diabetes	fear	1,054
type 1 diabetes	death	999
rationing insulin	death	805
type 1 diabetes	insulin	751
OGTT*	sick	584
type 1 diabetes	hypoglycaemia	578
insulin	hypo	545
insulin	fear	534
type 1 diabetes	insulin pump	436
finance	death	423
type 1 diabetes	sick	400
insulin	sick	385
insulin	finance	367
type 1 diabetes	anger	356
insulin	medication	305
insulin	anger	296
OGTT*	fear	293
type 2 diabetes	death	293
type 2 diabetes	fear	290
hypertension	death	286
overweight	death	280
type 1 diabetes	finance	277
hypoglycaemia	insulin	272
hypoglycaemia	sick	263
affordable insulin	death	262



epiconcept
smart health

Interactive cause effect network

- D3 json



accessible under: https://observablehq.com/@adahne/cause-and-effect-associations-in-diabetes-related-tweets

Discussion

epiconcept
smart health

Strenghts & Limitations

Strength

- detected a large number of *cause-effect* associations compared to other works
- tackled both (multi-word) explicit and implicit causality
- Avoided manually crafting causality rules

Limitations

- no real-world causal inference
- Classification performance
  - lack of recall counterbalanced by sheer amount of data in the first place
  - lack of precision counterbalanced by clustering approach
- Data quality

Big thanks to the co-authors

Identifying causal associations in tweets using deep learning: Use case on diabetes-related tweets from 2017-2021

Adrian Ahne, Vivek Khetan, Xavier Tannier, Md Imbessat Hassan Rizvi, Thomas Czernichow, Francisco Orchard, Charline Bour, Andrew Fano, Guy Fagherazzi

Preprint: https://arxiv.org/abs/2111.01225#

Currently under peer-review

Let’s play!!

Notebooks

1) Basics_Transformers.ipynb
2) Personal tweets.ipynb
3) causal_sentence_classifier.ipynb
4) BERT_cause_effect.ipynb (just looking)
5) FastText_cause_effect.ipynb
6) BERT_CRF.ipynb (just looking)
7) Cluster cause-effects.ipynb