NLP Fundamentals & Longitudinal Language Processing from User Generated Content

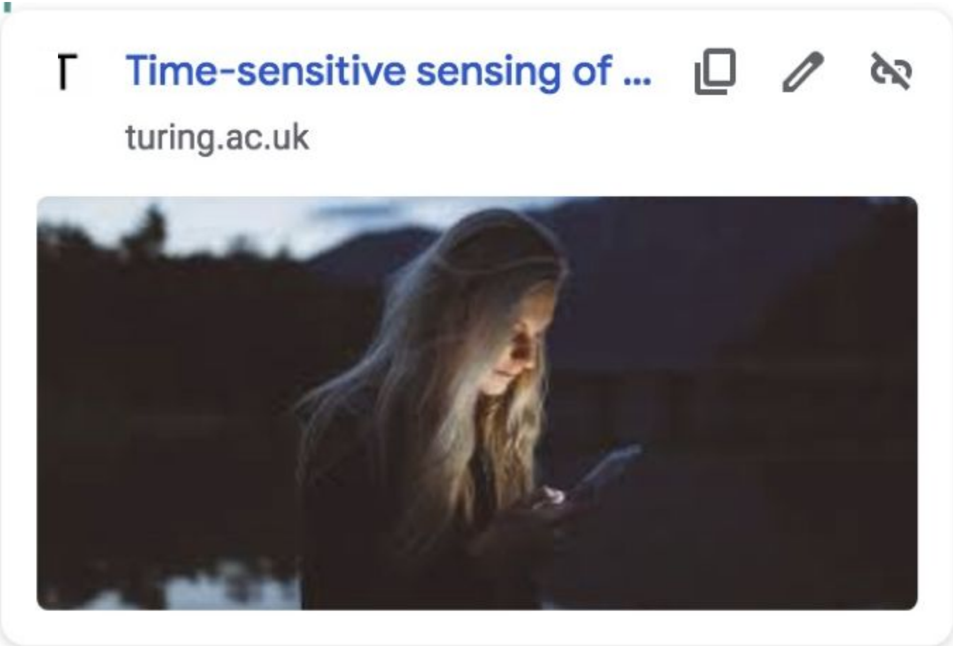
Prof Maria Liakata
UKRI/EPSRC Alan Turing Institute AI fellow
Queen Mary University of London
9th June 2024

Lectures Layout

- **Lecture I:** Introduction to Natural Language Processing (NLP) up to LLMs and associated challenges
- **Lecture II:** Identifying changes in longitudinal user generated content (I) (recurrence and path signatures)
- **Lecture III:** Identifying changes in longitudinal user generated content (II) (transformer based methods)
- **Lecture IV:** Timeline extraction, Timeline summarisation and demo on identifying longitudinal changes.

Overview

AI Fellowship: Creating Time Sensitive Sensors from Language & Heterogeneous UGC



The Alan Turing Institute

EPSRC

Queen Mary University of London

Overview

Addressing sociotechnical limitations of LLMs in medical and social contexts
https://adsolve.github.io/ – Hiring PhDs and postdocs!

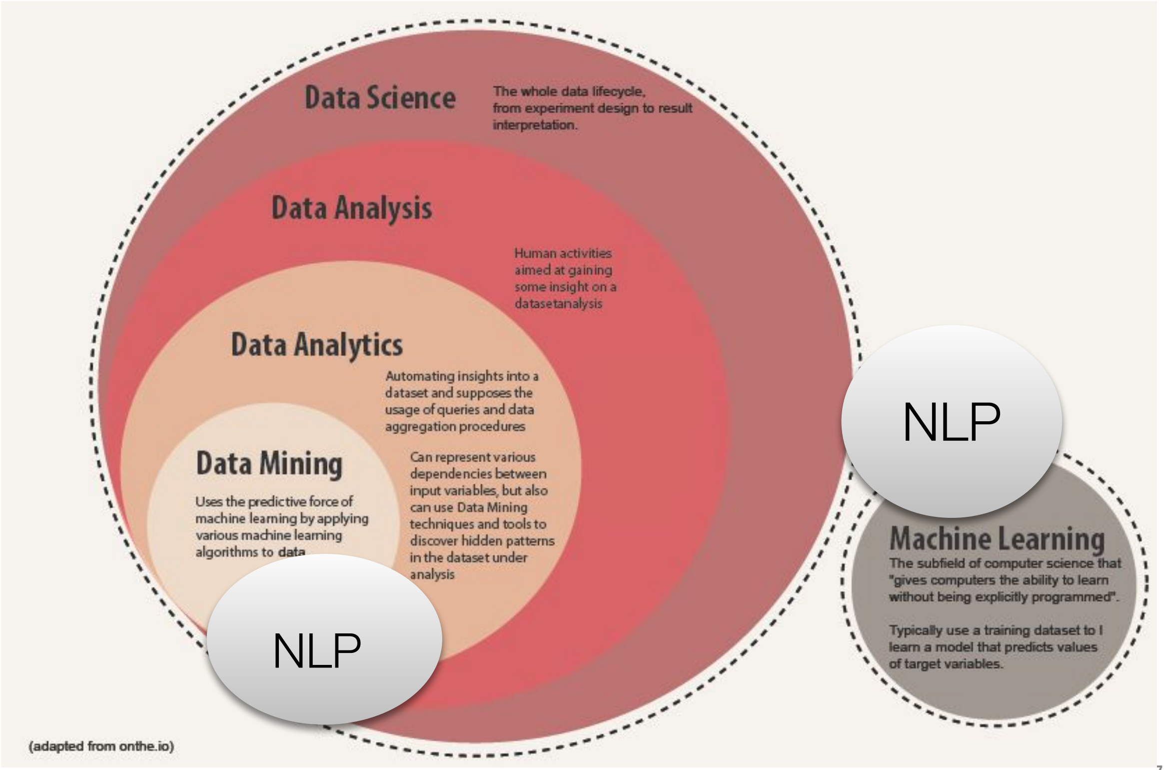
Queen Mary University of London
University of Nottingham UK | CHINA | MALAYSIA
University of Sheffield
WARWICK THE UNIVERSITY OF WARWICK
RAi

Lecture I: Introduction to Natural Language Processing (NLP) up to LLMs and associated challenges

Lecture I Layout

- What is NLP, a brief history of NLP & example applications
- Why is NLP challenging?
- Learning paradigms, tasks & evaluation
- Introduction to RNNs
- Introduction to PLMs
- Limitations of LLMs

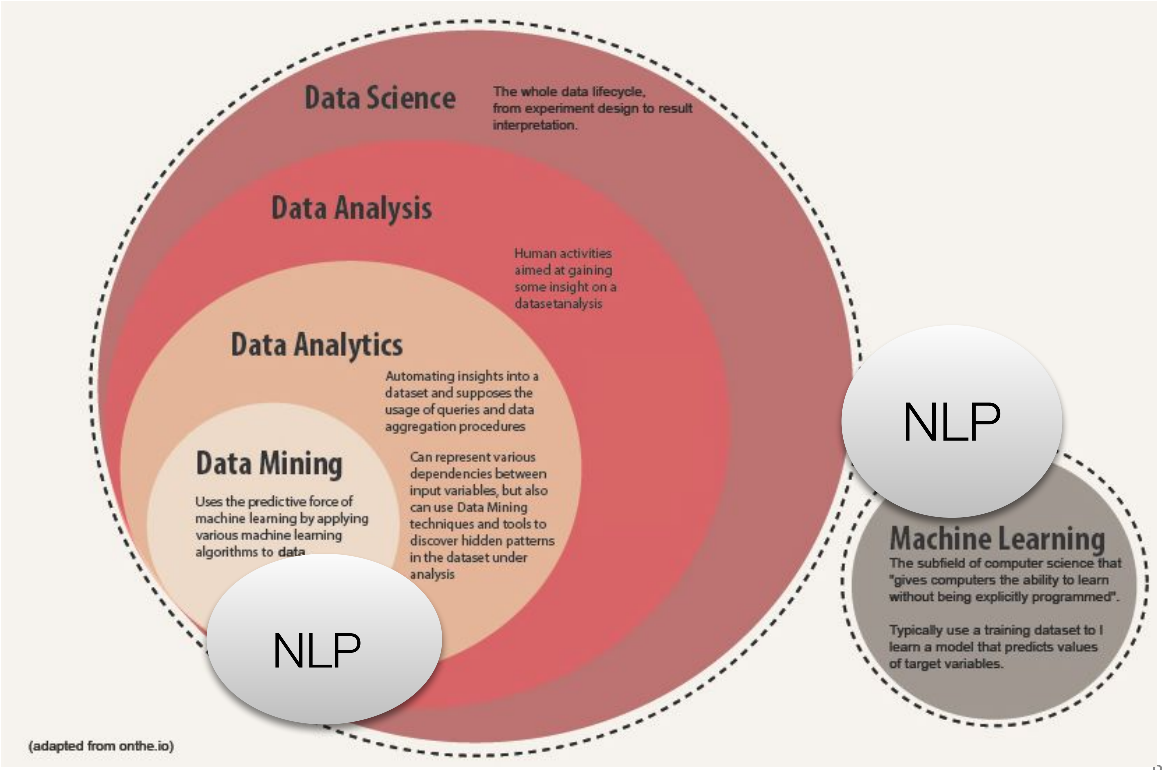
What is Natural Language Processing?



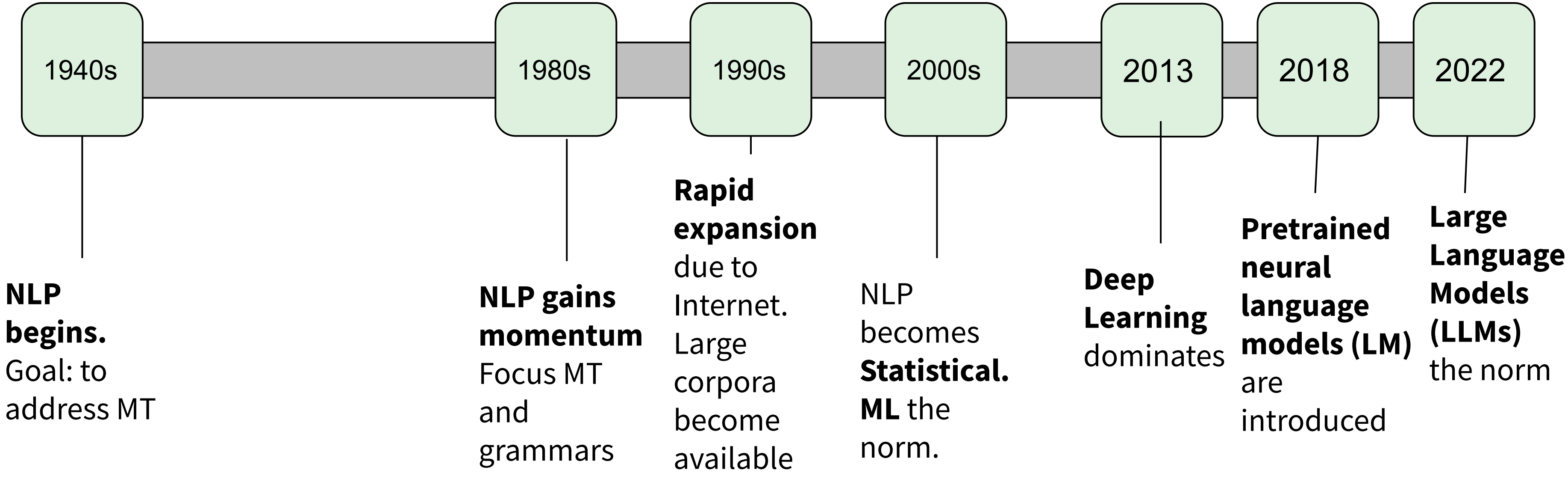
What is Natural Language Processing?

Chris Manning on NLP:

"Our field is the domain science of language technology; it's not about the best method of machine learning—the central issue remains the domain problems. The domain problems will not go away. More of the field's efforts should go into problems, approaches and architectures"



A very brief history of NLP



Why is NLP important?

- An enormous amount of knowledge is available in machine readable form as natural language text (online news, social media, digital archives, wikipedia, etc.)

10

Why is NLP important?

- An enormous amount of knowledge is available in machine readable form as natural language text (online news, social media, digital archives, wikipedia, etc.)

- Conversational agents are becoming an important form of human-computer communication (e.g. siri, alexa, spoken dialogue systems, written Q-A)

Why is NLP important?

- An enormous amount of knowledge is now available in machine readable form as natural language text (online news, social media, digital archives, wikipedia, etc.)

- Conversational agents are becoming an important form of human-computer communication (e.g. siri, alexa, spoken dialogue systems, written Q-A)

- Much of human-human communication is now mediated by computers (e.g. machine translation)

12

What can we do with NLP?

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON ORG LOC

Einstein met with UN officials in Princeton

Tokenization

Sentence segmentation

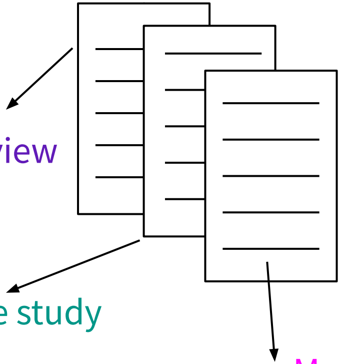
Chunking & Parsing

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Text classification



Review

Case study

Methods

Natural Language Inference

The Dow Jones is up

US Economy doing well

Information Extraction

Einstein met with UN officials in Princeton

Entity 1: Einstein
Entity 2: UN officials
Relation: meeting
Location: Princeton

Text Generation:

Machine Translation
Question Answering
Dialogue Systems
Summarisation
Synthetic Text

Lower level → Higher level

Many commercial applications

Example NLP applications: Machine Translation

Η μεγαλύτερη φωτιά της Ευρώπης στον Εβρο – Κάηκε έκταση μεγαλύτερη από τη Νέα Υόρκη

Η πυρκαγιά έχει καταστρέψει τουλάχιστον 808,7 τετραγωνικά χιλιόμετρα ενώ η πόλη της Νέας Υόρκης καταλαμβάνει 778,2 τετραγωνικά χιλιόμετρα

1' 40" χρόνος ανάγνωσης

Example NLP applications: Machine Translation

Η μεγαλύτερη φωτιά της Ευρώπης στον Εβρο – Κάηκε έκταση μεγαλύτερη από τη Νέα Υόρκη

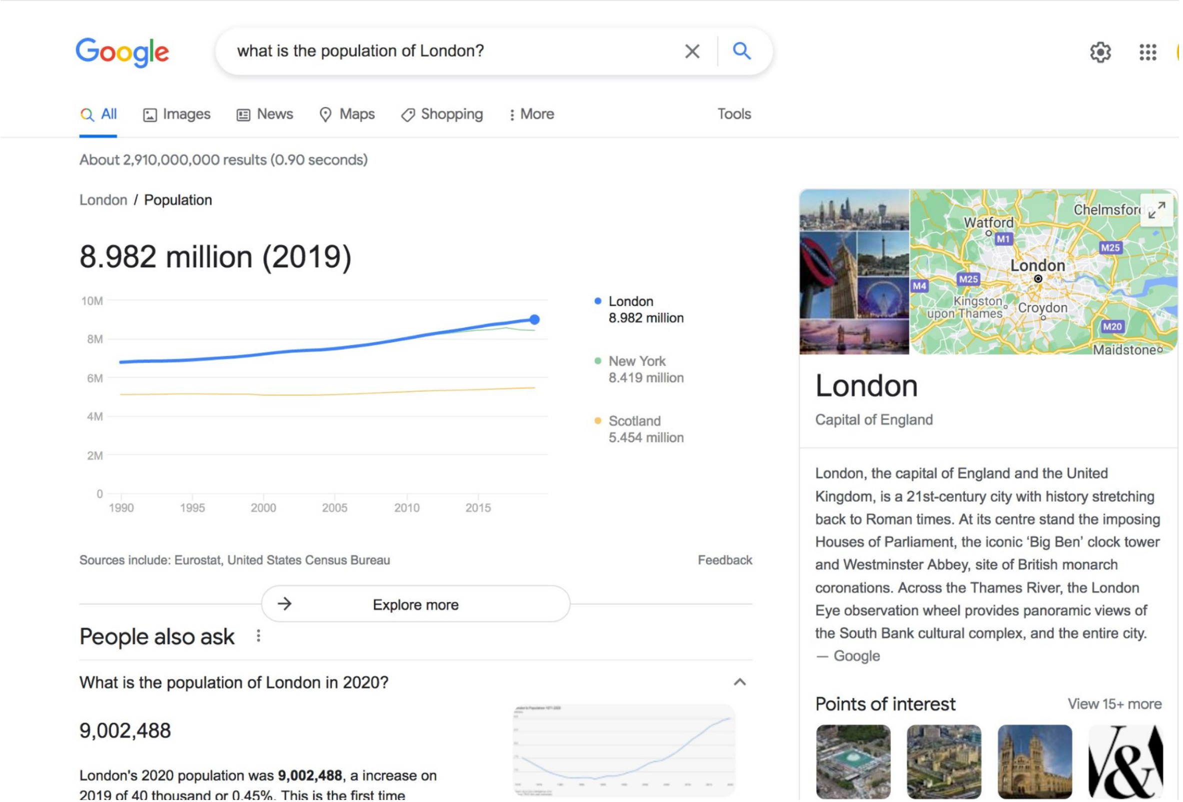
Η πυρκαγιά έχει καταστρέψει τουλάχιστον 808,7 τετραγωνικά χιλιόμετρα ενώ η πόλη της Νέας Υόρκης καταλαμβάνει 778,2 τετραγωνικά χιλιόμετρα

1' 40" χρόνος ανάγνωσης

Europe's biggest fire in Evros – An area larger than New York burned

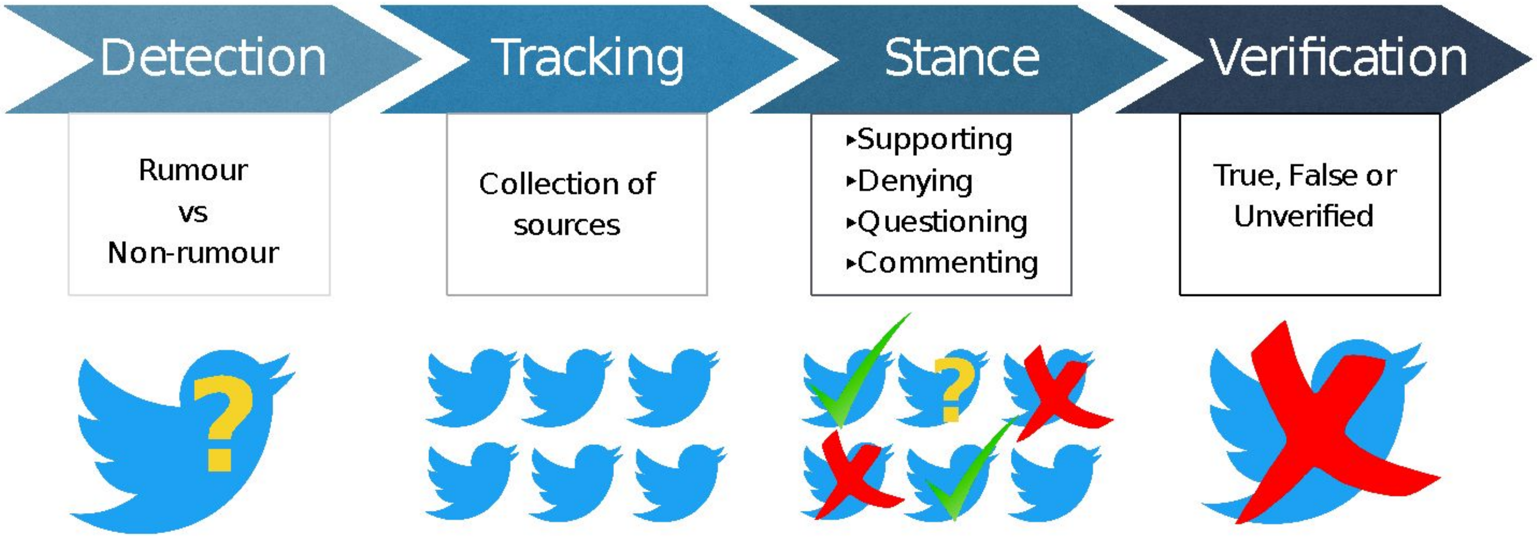
The fire has destroyed at least 808.7 square kilometers while New York City occupies 778.2 square kilometers

Example NLP Applications: Web Q/A



Example NLP Applications: Q-A and dialogue systems

Example NLP application: Rumour verification



Why is NLP Challenging?

Language is ambiguous and it's important to understand the context!

Lexical ambiguity:

"I went to the bank." (The bank could be a place where money is kept, or it could be the edge of a river.)

Syntactic (grammatical) ambiguity:

"Visiting relatives can be exhausting." (What is exhausting: relatives who visit you, or when you visit relatives?)

Why is NLP Challenging?

Language is in flux

Neologisms: “unfriend”, “selfie”
Idioms: “under the weather”
Multi-word expressions: “Let it be” is a good song
Variability: typos, learner mistakes, individual style
Metaphors: “She is an angel for helping me out”

Why is NLP Challenging?

Language is in flux

Neologisms: "unfriend", "selfie"
Idioms: "under the weather"
Multi-word expressions: "Let it be" is a good song
Variability: typos, learner mistakes, individual style
Metaphors: "She is an angel for helping me out"

Language manifests complex phenomena
Bridging: "the student arrived on time but the Professor had already left"
Ellipsis: " I am going and so are you"
Anaphora: co-reference between propositions, across documents and domains

Core Challenges in NLP remaining

- Small Data & Transfer Learning across Domains

Core Challenges in NLP remaining

- Small Data & Transfer Learning across Domains
- Obtaining Useful & Interpretable Representations

Core Challenges in NLP remaining

- Small Data & Transfer Learning across Domains
- Obtaining Useful & Interpretable Representations
- Model fairness & privacy preservation

Core Challenges in NLP remaining

- Small Data & Transfer Learning across Domains
- Obtaining Useful & Interpretable Representations
- Model fairness & privacy preservation
- Collaborative Human-Computer Process

Core Challenges in NLP remaining

- Small Data & Transfer Learning across Domains
- Obtaining Useful & Interpretable Representations and model
- Model bias & privacy preservation
- Collaborative Human-Computer Process
- Resolving Complex Linguistic Phenomena

Context: Core Challenges in NLP

- Small Data & Transfer Learning across Domains
- Obtaining Useful & Interpretable Representations
- Model fairness & privacy preservation
- Collaborative Human-Computer Process
- Resolving Complex Linguistic Phenomena

\+ **Time-sensitivity**

Learning paradigms, tasks & evaluation

Important Concepts

- Like in ML, In NLP a computer **learns to** address a task without being explicitly programmed to do so.

Important Concepts

- In NLP and ML a computer **learns to** address a task without being explicitly programmed to do so.
- **Machine learning algorithms** are procedures that are implemented in code and are run on data.

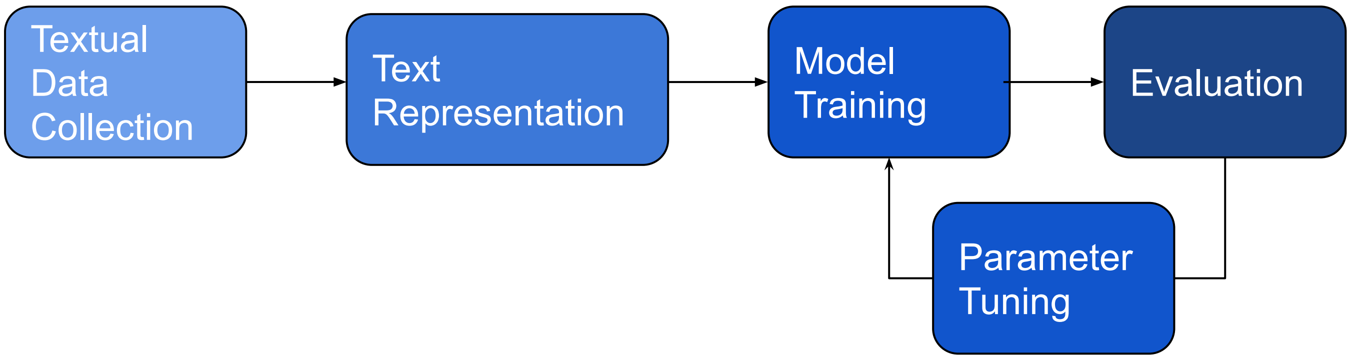
Important Concepts

- In NLP and ML a computer **learns to** address a task without being explicitly programmed to do so.
- **Machine learning algorithms** are procedures that are implemented in code and are run on data.
- **Machine learning models** are output by algorithms and are comprised of model data and a prediction algorithm

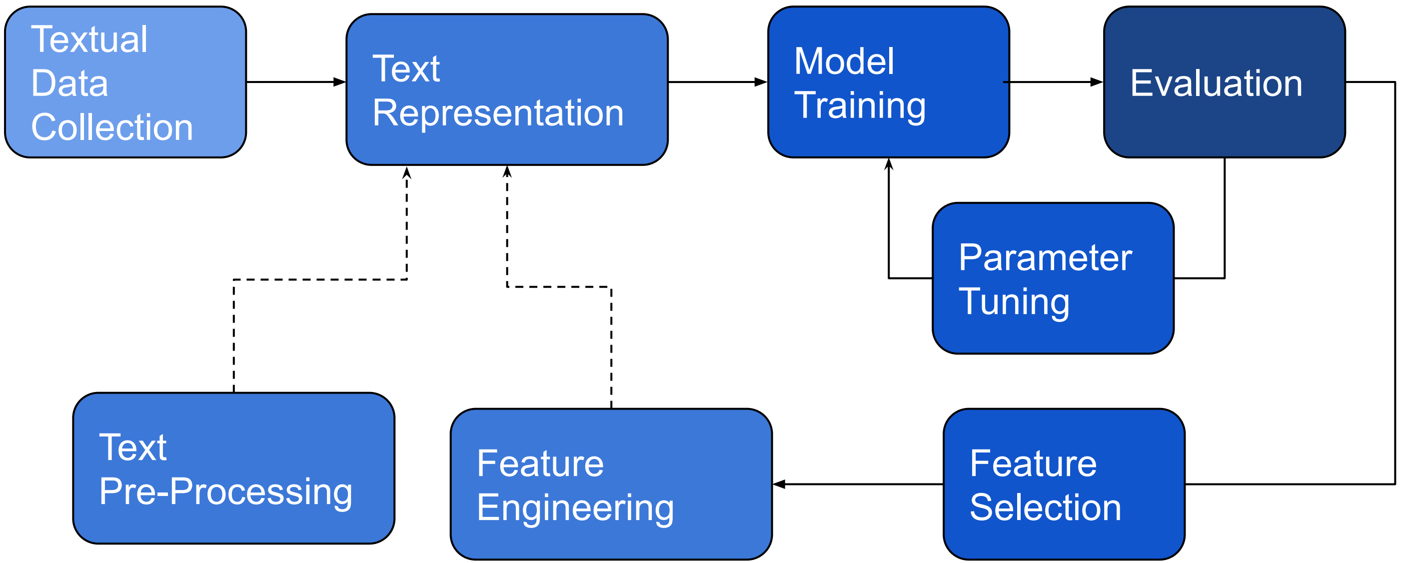
Important Concepts

- In NLP and ML a computer **learns to** address a task without being explicitly programmed to do so.
- **Machine learning algorithms** are procedures that are implemented in code and are run on data.
- **Machine learning models** are output by algorithms and are comprised of model data and a prediction algorithm
- The **learning paradigm** is the learning setup, including data representation, algorithm and model.

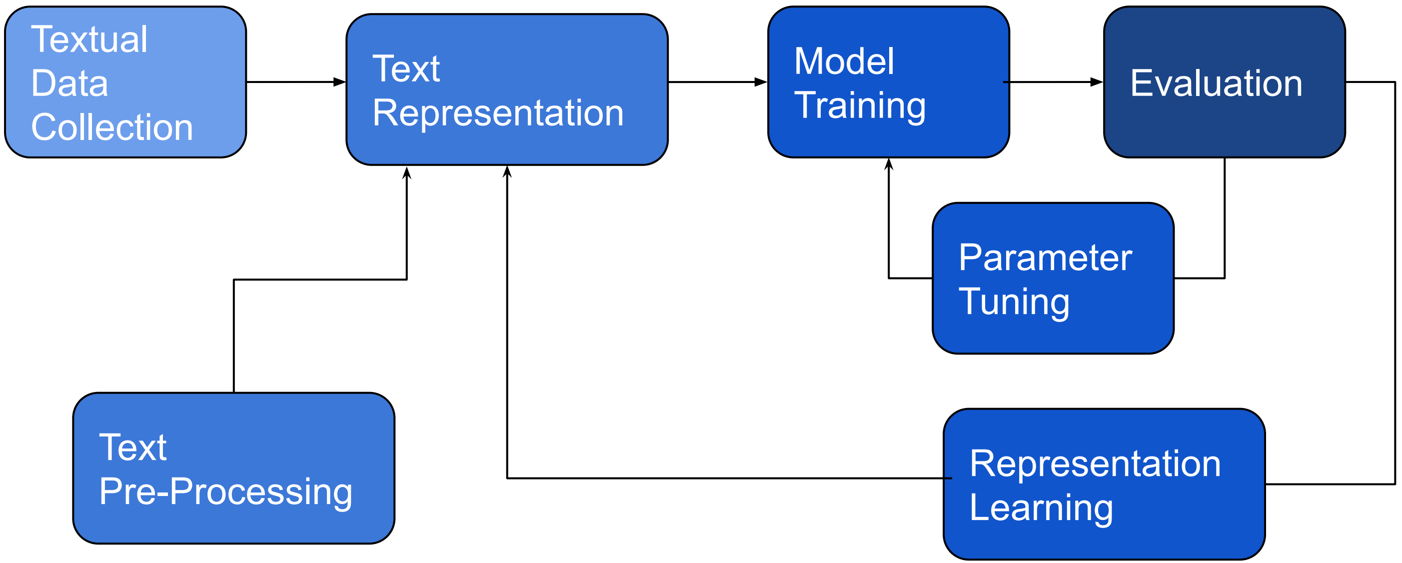
General Statistical NLP Learning Paradigm



General Statistical NLP Learning Paradigm



General Statistical NLP Learning Paradigm



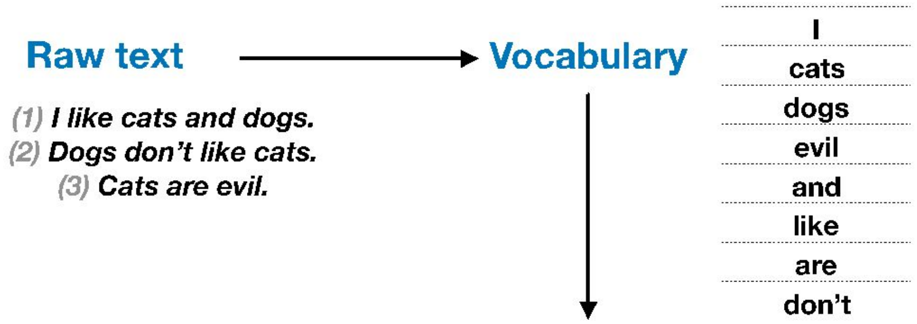
Text Pre-Processing

- Lower-casing
- Stop word removal
- Removal of non alphanumeric characters
- Tokenization
- Stemming/lemmatization
- Parsing
- Part of Speech Tagging

Text Pre-Processing

- Lower-casing
- Stop word removal
- Removal of non alphanumeric characters
- **Tokenization**
- Stemming/lemmatization
- Parsing
- Part of Speech Tagging

Representing textual data



Bag-of-Words Representation

Vocabulary	Cats	Dogs	Sentence 1	Sentence 2	Sentence 3
I	0	0	1	0	0
cats	1	0	1	1	1
dogs	0	1	1	0	1
evil	0	0	0	1	0
and	0	0	1	0	0
like	0	0	1	0	1
are	0	0	0	1	0
don't	0	0	0	0	1

Issues with bag of words representation

- High-dimensional sparse representations
- Dimensionality depends on the vocabulary size
- The semantic relations between words are not captured
- Information about word order is not preserved

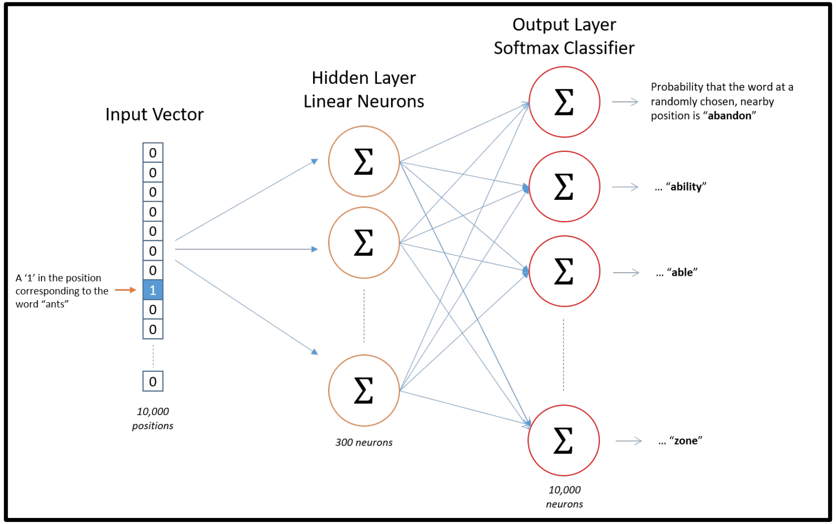
Distributional Hypothesis & Embeddings

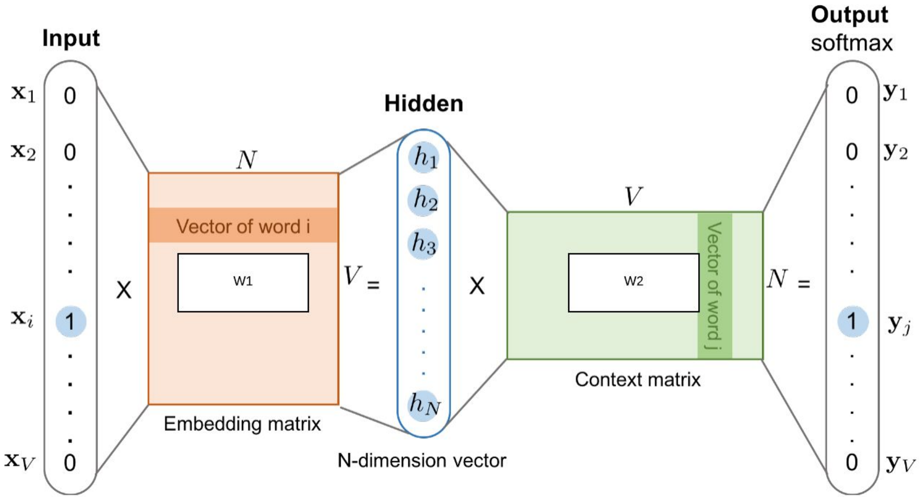
"You shall know a word by the company it keeps"

- Firth, J. R. (1957:11)

Embeddings: Projections of word vectors to a lower dimension, dense semantic space. Based on distributional hypothesis.

Obtaining WORD2VEC Embeddings





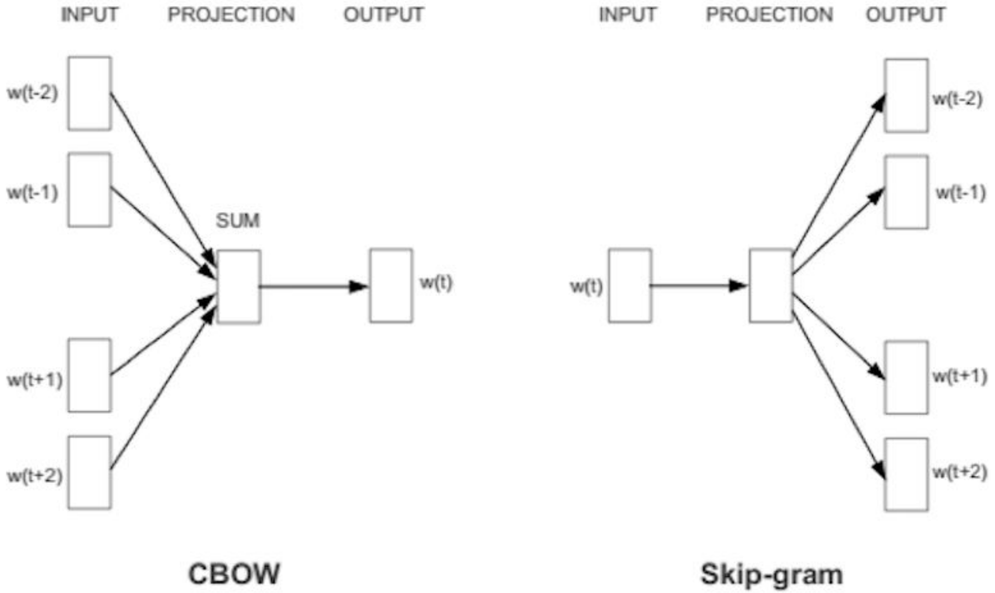
Introduced in a 2013 paper by Mikolov et al.

WORD2VEC

Sliding window

w(t-2) w(t-1) w(t) w(t+1) w(t+2)

An investment in knowledge pays the best interest

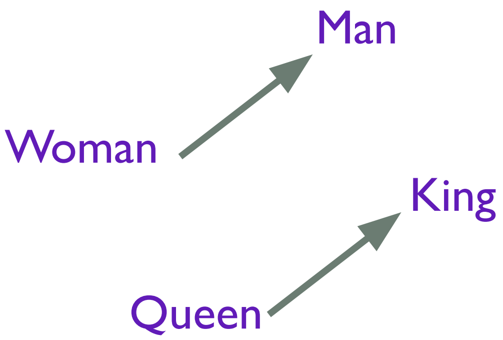


Introduced in a [2013 paper] by Mikolov et al.

Nice properties of WORD2VEC vectors

- Words or phrases from the vocabulary are mapped to vectors of real numbers.
- Each dimension of a feature vector is a latent feature
- Low dimensionality comparing to BoW, and constant wrt the vocabulary size.
- Word embeddings capture the semantic and syntactic similarities between words.

Nice properties of WORD2VEC vectors

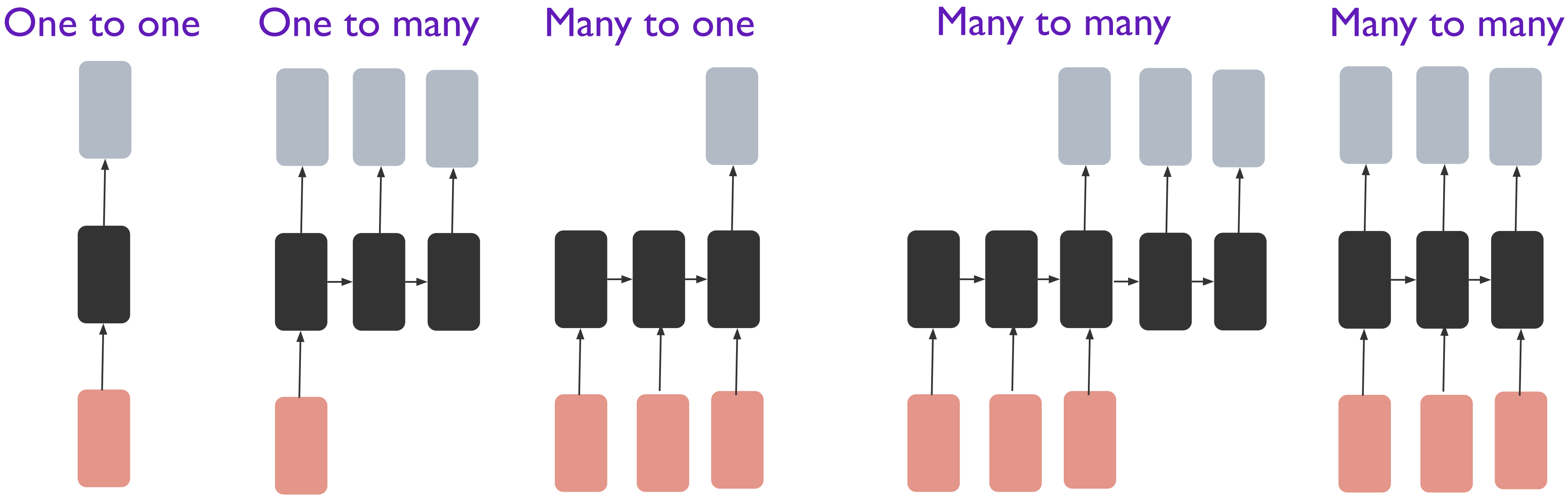


Expression	*Nearest token*
Paris - France + Italy	Rome
bigger - big + cold	colder
sushi - Japan + Germany	bratwurst
Cu - copper + gold	Au
Windows - Microsoft + Google	Android
Montreal Canadiens - Montreal + Toronto	Toronto Maple Leafs

'Queen' – 'Woman' + 'Man' = 'King'

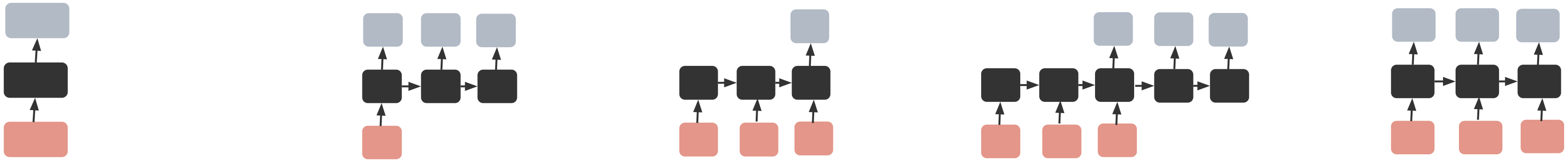
https://lamyiowce.github.io/word2viz/

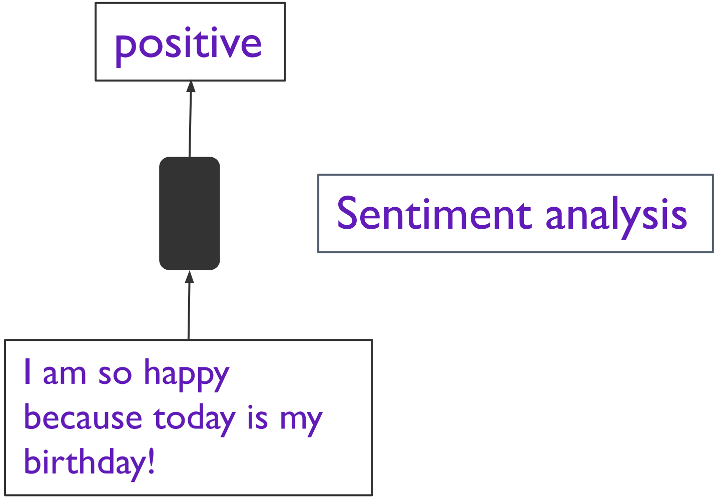
NLP Algorithms: Types of tasks



NLP Algorithms: Types of tasks

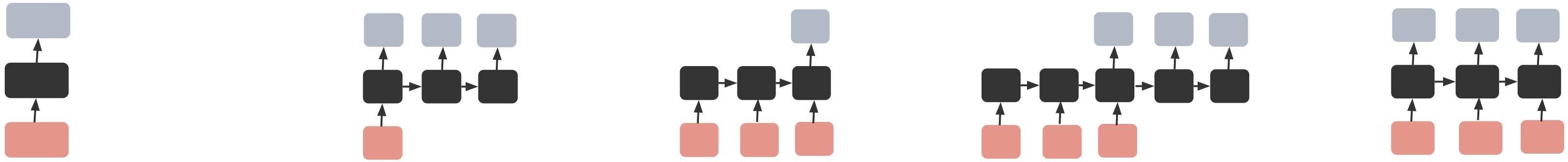
One to one | One to many | Many to one | Many to many | Many to many

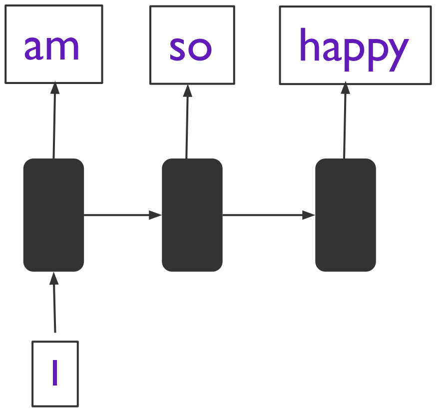




NLP Algorithms: Types of tasks

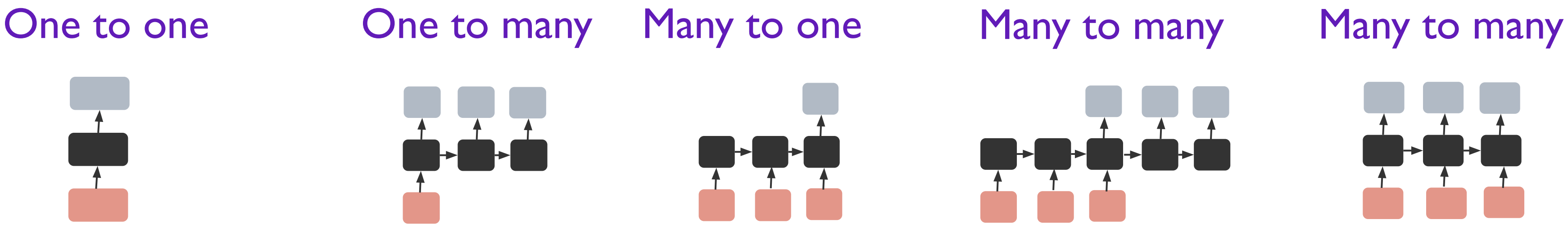
One to one | One to many | Many to one | Many to many | Many to many

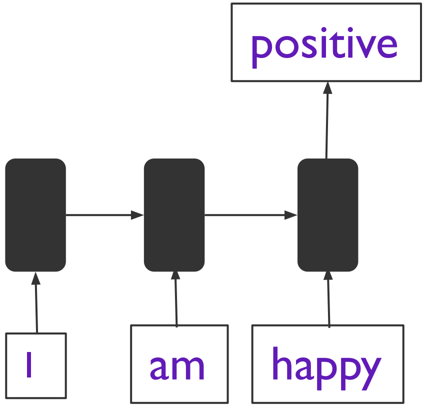




Sentence generation

NLP Algorithms: Types of tasks

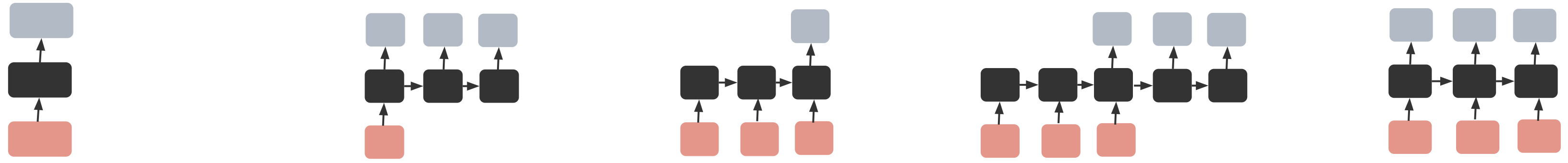


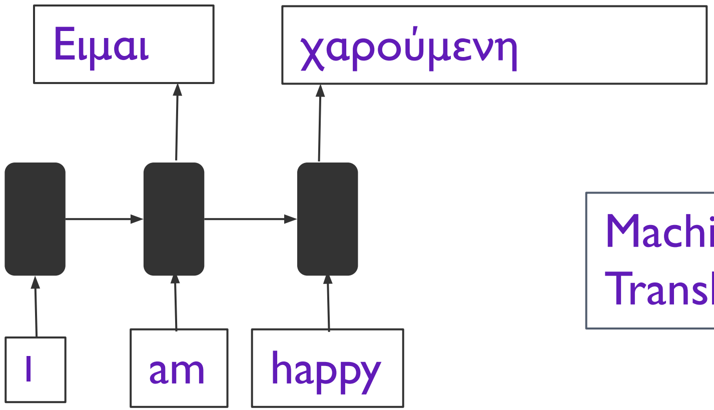


Sentiment analysis

NLP Algorithms: Types of tasks

One to one | One to many | Many to one | Many to many | Many to many

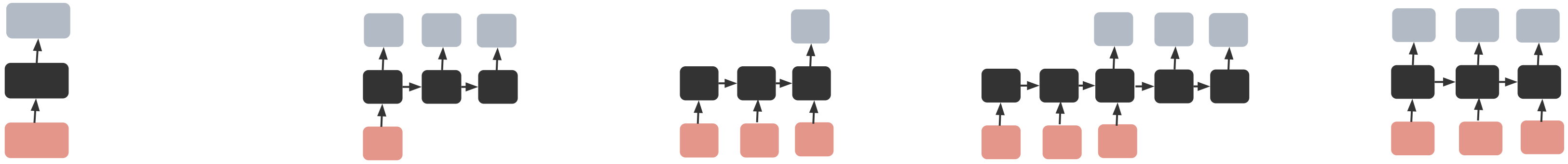




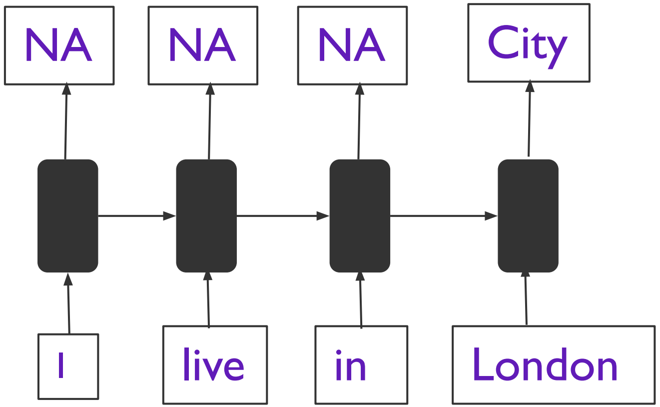
Machine Translation

NLP Algorithms: Types of tasks

One to one | One to many | Many to one | Many to many | Many to many

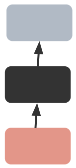


Named Entity Recognition

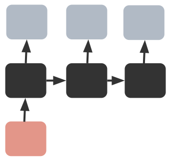


NLP Algorithms: Types of tasks

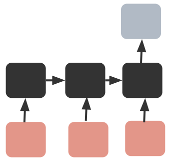
One to one



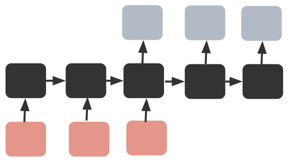
One to many



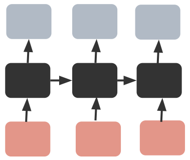
Many to one



Many to many



Many to many



Most common

- SVM
- RF
- LR
- MLP/FFNN

- **BERT**

Sequential classifiers: RNNs: LSTM, GRU; linear-CRF

Structured prediction: Recursive NNs, Tree/Graph NNs, tree-CRF

Different Transformer architectures

Sequential neural models: RNNs and LSTM

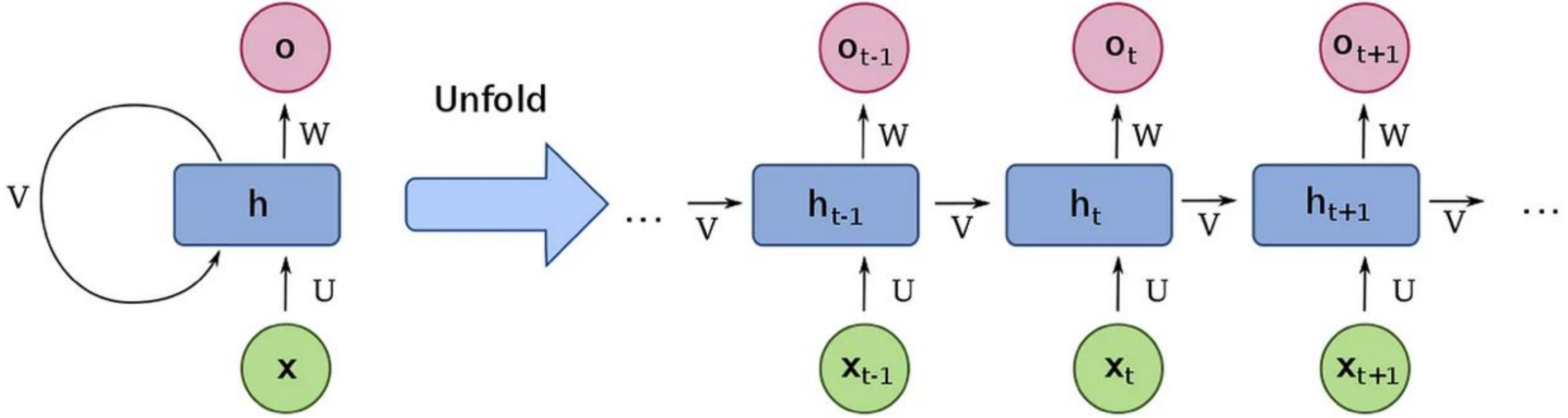
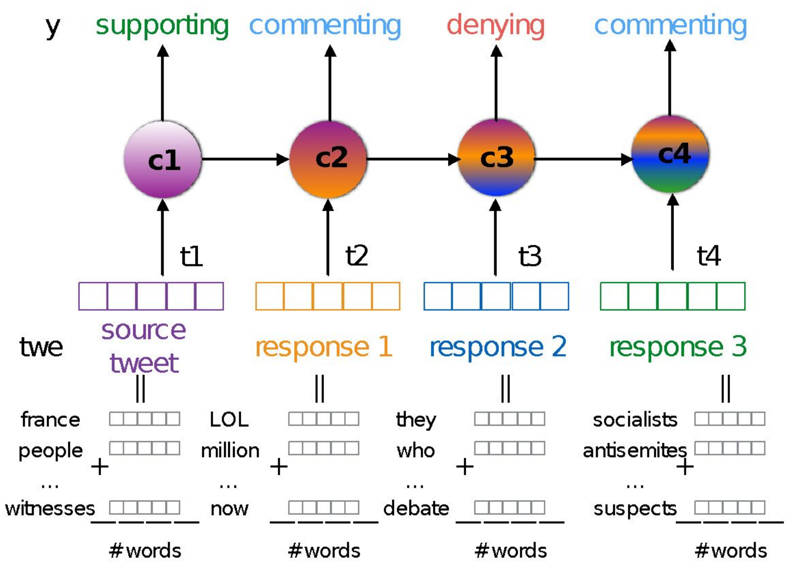
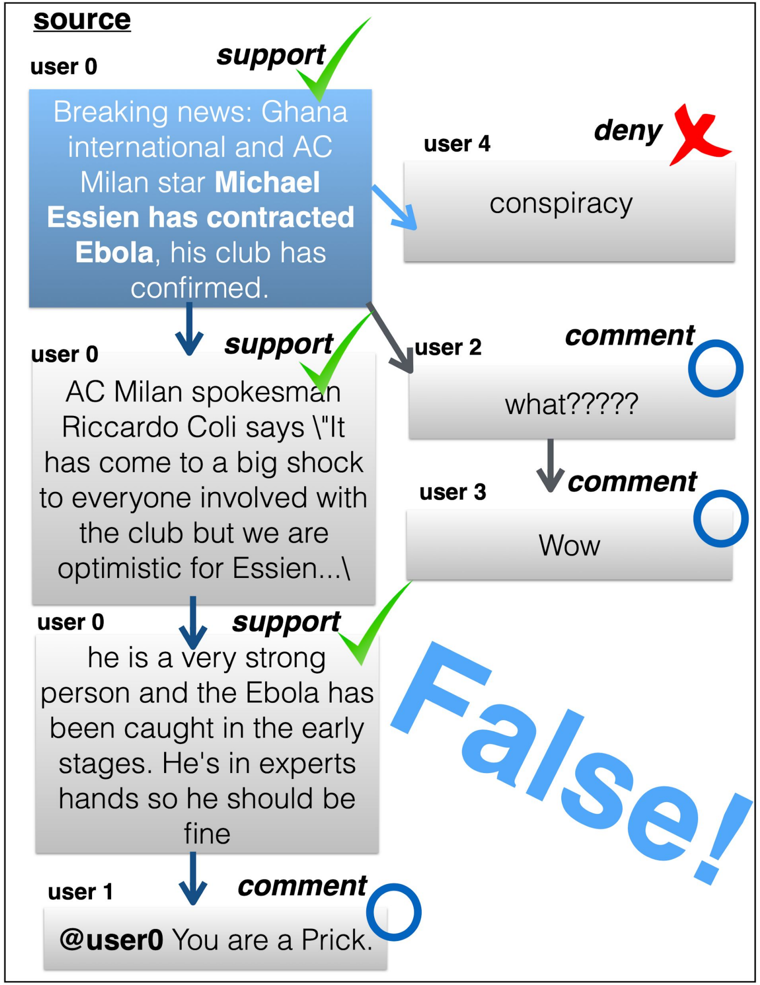


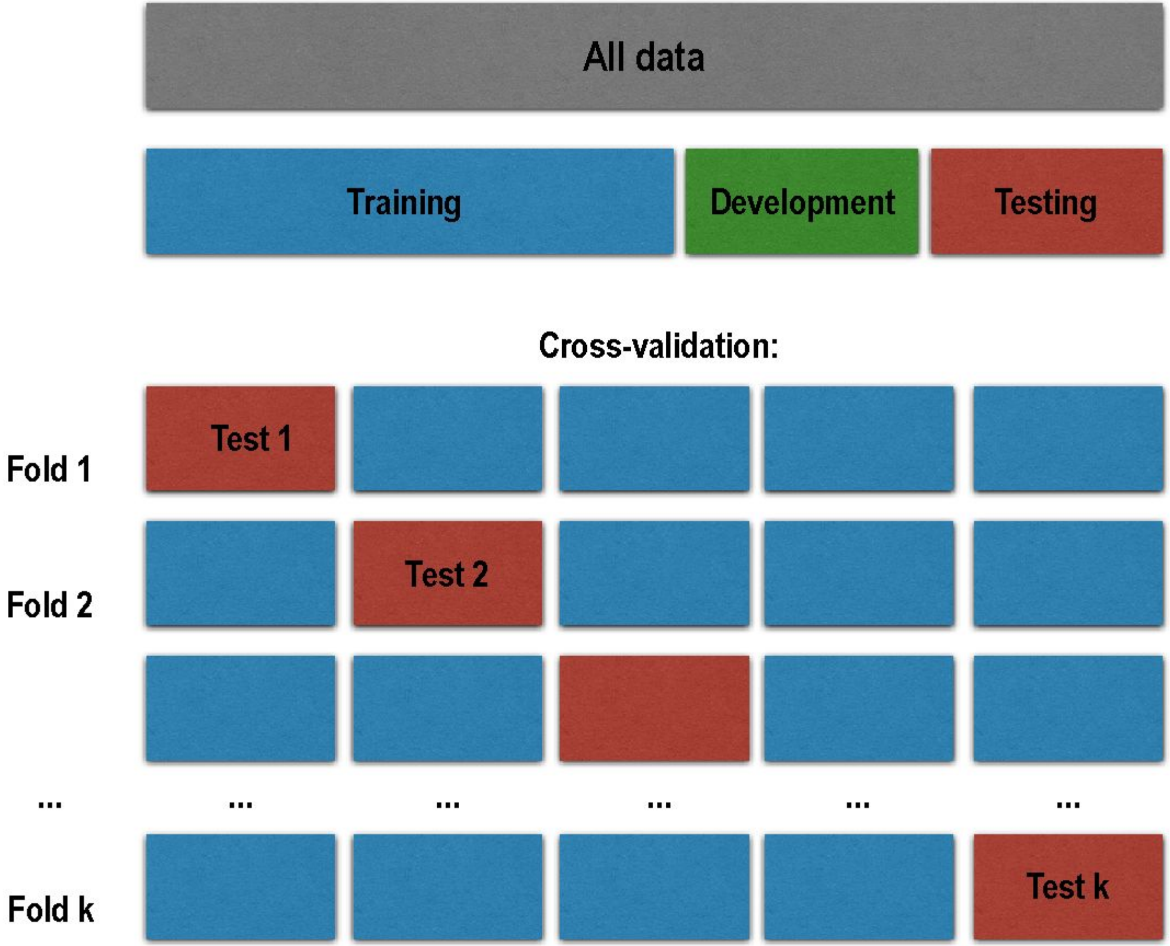
Image from medium.com

Example: Rumour verification





Evaluation Strategies



Evaluation metrics for classification

- The result of classification is a set of predicted probabilities that a certain element belongs to each of the possible classes.

- Classifier decides where a particular instance belongs by choosing a class with the highest probability

- Comparing the predicted output with the true class labels we can create confusion matrix

		Predicted	
		TRUE	FALSE
TRUE	Y	5 TP	3 FN
	N	2 FP	17 TN

Evaluation Metrics

Build confusion matrix per class to obtain TP, TN, FN and FP values, then calculate the following properties and perform averaging.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2TP}{2TP + FN + FP}$$

		Predicted	
		TRUE	FALSE
TRUE	Y	5 TP	3 FN
	N	2 FP	17 TN

- **Micro:** Calculate metrics globally by counting the total true positives, false negatives and false positives.
- **Macro:** Calculate metrics for each label, and find their unweighted mean. This does not take label imbalance into account.

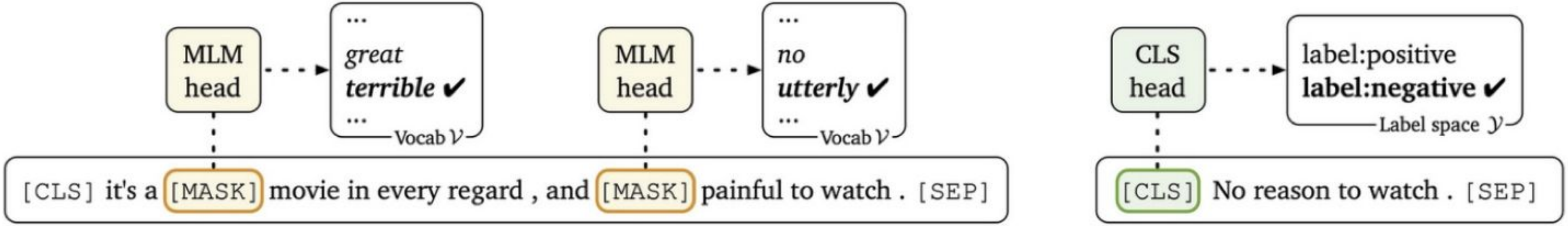
Pre-trained language models

- These days most work in NLP is done on the basis of pre-trained language models

Pre-trained language models

- These days most work in NLP is done on the basis of pre-trained language models

- **Typical training strategies involve Masked Language Modelling (MLM)**.



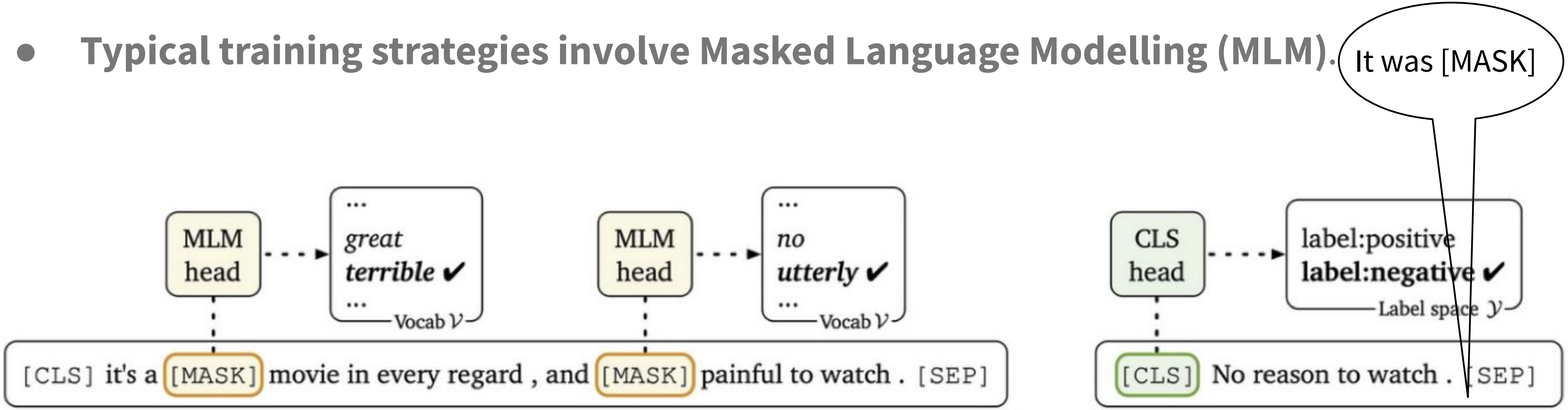
(a) MLM pre-training

(b) Fine-tuning

Pre-trained language models

- These days most work in NLP and vision is done on the basis of pre-trained language models

- **Typical training strategies involve Masked Language Modelling (MLM).**

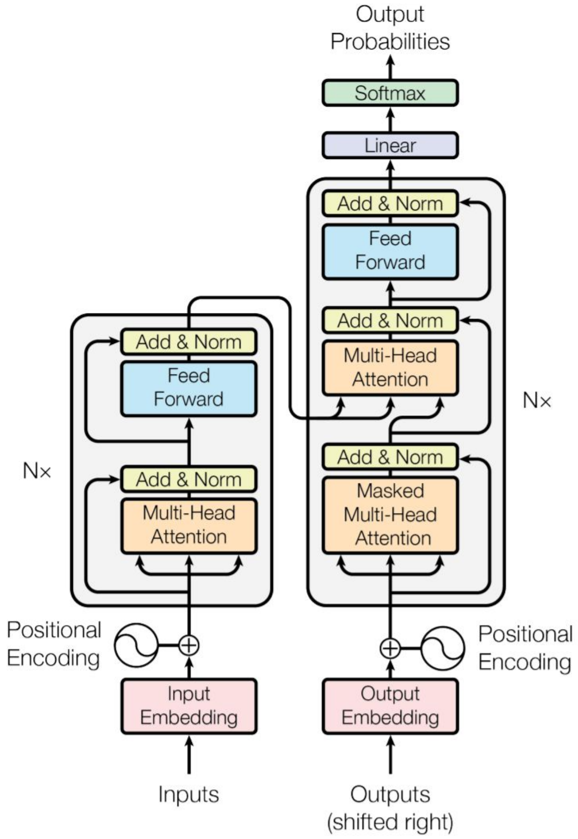


(a) MLM pre-training

(b) Fine-tuning **(prompt-based)**

Transformer Network

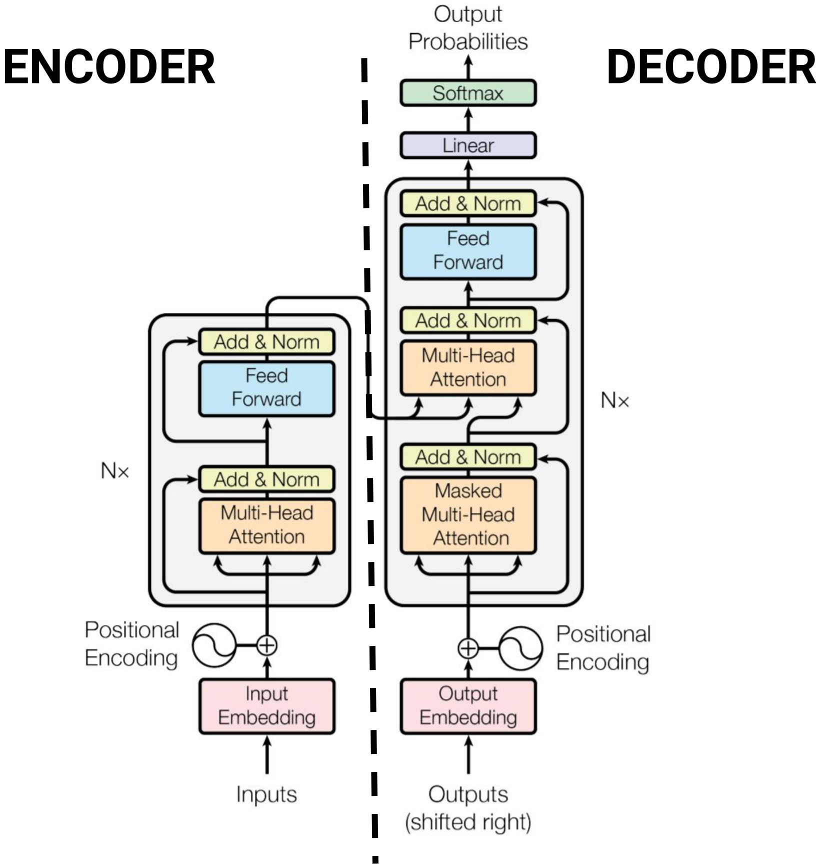
- *Current pre-trained language models are based on a **Transformer network architecture***



[Vaswani et al. 2017. Attention is all you need. NIPS]

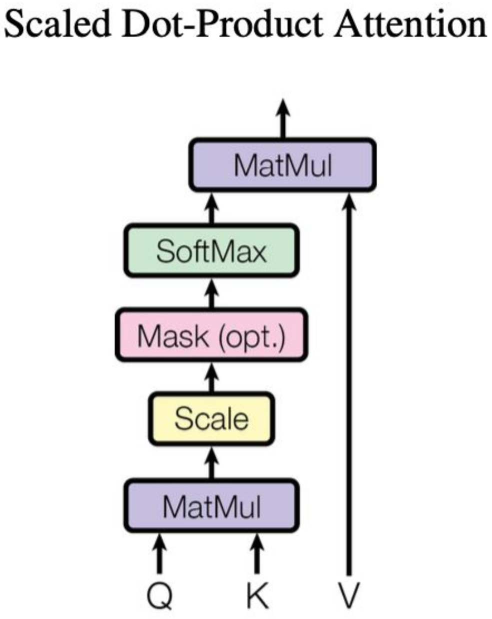
Transformer Network

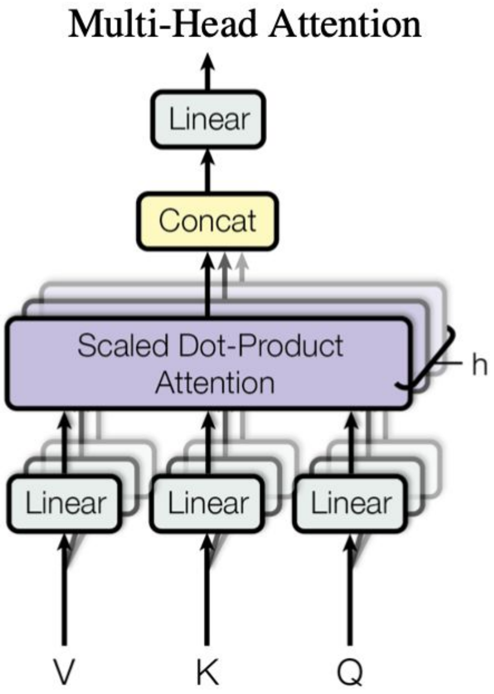
- *Most pre-trained language models are based on a* ***Transformer network architecture***



Vaswani et al. 2017. Attention is all you need. NIPS]

Multi-head attention





[Vaswani et al. 2017. Attention is all you need. NIPS]

Lecture II: Identifying changes in longitudinal user generated content (I) (recurrence)

Lecture II Layout

- Limitations of LLMs and challenges with sensitive data
- Introduction to Personalised Longitudinal NLP and associated objectives
- The task of identifying moments of change from longitudinal user text.
- Sequential signature networks for longitudinal monitoring

Example PLMs

Model	Parameter Size	Function	Other
BERT (Google 2018)	345 million	Embeddable in new applications	Encoder only, best for classification
GPT-2 (Radford et al 2019)	1.5 B	Embeddable in new applications	Decoder only, best for generation
T5 (Raffel et al. 2020)	10m->11B	Embeddable in new applications	Encoder-Decoder
GPT-3,3.5 (Open AI-2020,2022)	175 billion	API, service	Focus on decoder and generation
ChatGPT (Open AI-Nov 2022 based on instruction fine-tuned GPT 3.5 turbo)	117 million	Service, no API	Focus on Q-A & interaction,knowledge up to 2023

What LLMs (such as GPT-3+) are good at

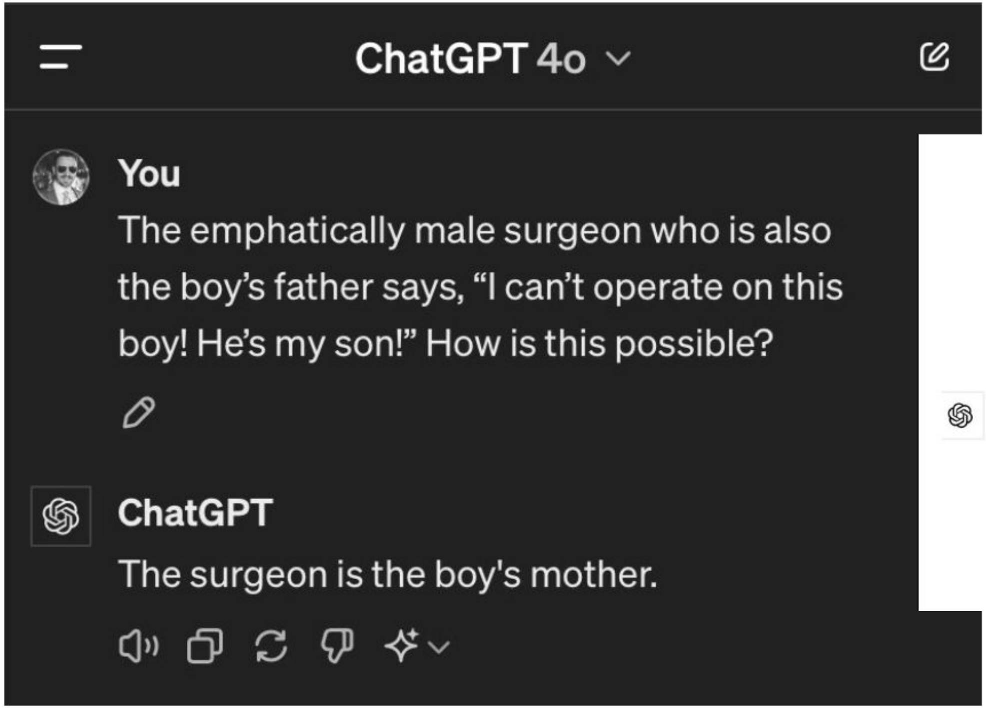
- *Capturing higher order co-occurrences in text*
- *Generation of fluent text in several languages*
- *Paraphrasing*
- *Generation of short coherent summaries from multiple documents*
- *Question Answering*

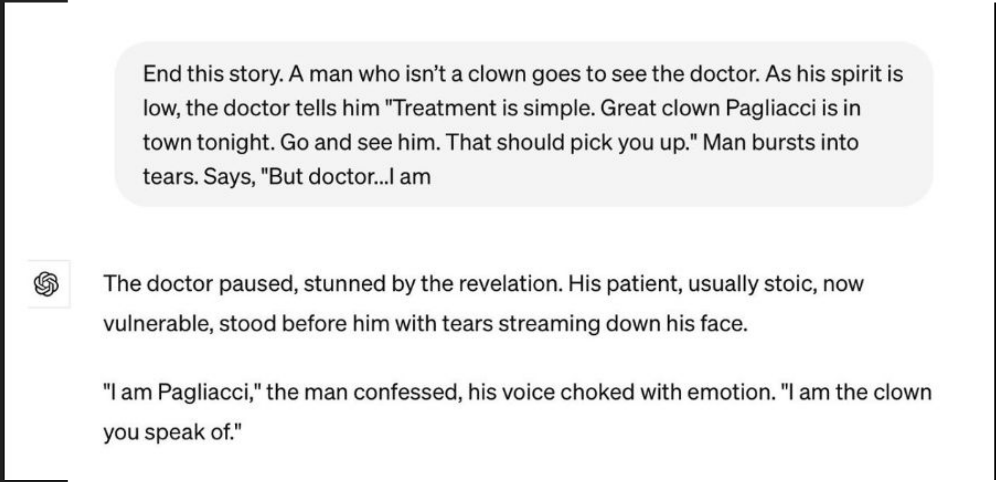
What LLMs (such as GPT-3+) are NOT good at

- Factuality, biases and hallucinations
- Low resource information
- Reasoning and complex inference
- Summarising long documents, temporal reasoning

What are LLMs not good at

Reasoning (here ignoring context and memorising)





What are LLMs not good at

Reasoning (failure in deduction)

You

Just two people, Anna and Beth are in the meeting room. Somebody leaves the meeting room to go to the cafeteria. Beth leaves the meeting room to go to the lab. Where is Anna?

ChatGPT

Based on the information provided, Anna is still in the meeting room. The statement mentions that somebody (not specified as Anna or Beth) leaves the meeting room to go to the cafeteria, and Beth leaves the meeting room to go to the lab. Since Anna's location is not mentioned in the context of leaving the meeting room, she can be presumed to still be in the meeting room until otherwise stated.

What are LLMs not good at

Reasoning (failure in counter-factuals)

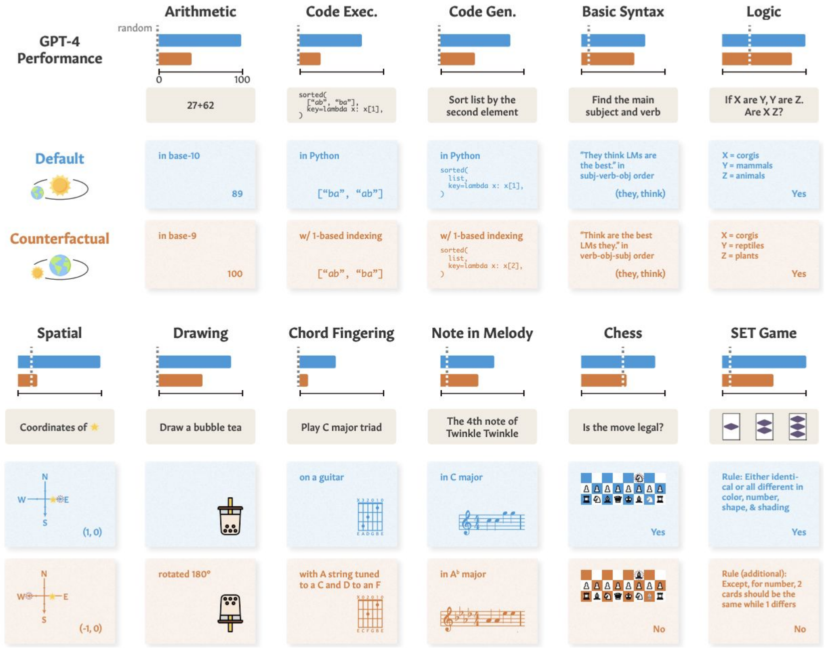


Figure 1: GPT-4's performance on the default version of various tasks (blue) and counterfactual counterparts (orange). The shown results use 0-shot chain-of-thought prompting (§4; Kojima et al., 2023). GPT-4 consistently and substantially underperforms on counterfactual variants compared to default task instantiations.

Wu et al. 2024 Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks. To appear in NAACL 2024.

What are LLMs not good at

Temporal reasoning

Tan et al. 2023. Towards Benchmarking and Improving the Temporal Reasoning Capability of Large Language Models. ACL.

		FLAN-T5-L		ChatGPT		T5-SFT		TempT5		
Question Type	Setting	EM	F1	EM	F1	EM	F1	EM	F1	Δ F1
L1: Time-Time	CBQA	0.0	2.9	30.5	56.7	100	100	100	100	+0.0
L2: Time-Event	CBQA	0.5	9.2	6.5	11.5	1.4	23.2	1.5	23.4	+0.2
	ReasonQA	57.3	66.3	47.5	51.0	82.6	87.1	84.8	88.9	+1.8
	OBQA	9.4	22.5	8.5	16.1	14.8	35.2	15.4	36.3	+1.1
L3: Event-Event	CBQA	0.4	10.5	12.0	21.8	12.1	25.3	12.3	25.4	+0.1
	ReasonQA	36.3	47.5	49.5	52.3	78.2	83.0	81.1	86.1	+3.1
	OBQA	8.1	19.2	17.0	25.3	19.7	31.2	21.1	32.4	+1.2

Level - 1
Time-Time Relation

Abstractive concept, can be viewed as a set of logic rules. Once human understands the concepts, it is easily generalizable.

Example Question:
What is the year after 2010?
Target:
2011

Level - 2
Time-Event Relation

Knowledge-intensive concept, where human will also need to memorize such relations, or leverage external context to deduce such relations.

Example Question:
What team did Leo Messi play for in 2010?
Target:
Barcelona

Level - 3
Event-Event Relation

Requires combination of memorization, deduction and understanding.

Example Question:
What team did Leo Messi play for after Barcelona?
Target:
Paris Saint-Germain

What are LLMs not good at

Reasoning (failure in distinguishing reasoning types)

Case: Wrong veracity wrong reasoning	
Model: GPT-3.5	**Method:** ZS Cot
Prompt: Q: As an expert journalist, classify the following rumour as true or false using only the provided evidence. **RUMOUR:** French media report the Trocadero near the Eiffel Tower in Paris has been evacuated and police have been deployed there **EVIDENCE:** PARIS Reuters - The Eiffel Tower and the surrounding Champ de Mars park were briefly evacuated on Tuesday because of a bomb alert, the fourth in PARIS… A: let's think step by step.	
Ground truth Veracity: False	**Predicted veracity:** True
Ground Truth Reasoning: Deductive	**Predicted Reasoning:** Inductive
Provided Explanation: `The above reasoning can be classified as inductive reasoning. Inductive reasoning involves drawing a general conclusion based on specific evidence or observations. In this case, the conclusion that the rumour is true is based on the specific evidence provided about the Eiffel Tower and Champ de Mars park being evacuated due to a bomb alert.`	

What LLMs are not good at

Summarisation of long documents such as user timelines

The timeline summary of the provided text suggests a series of distressing events and emotions over a period of time

1. Early morning struggle with contract and hunger.
2. Compulsive lying to mother and guilt.
3. Feeling isolated and seeking positivity while alone.
4. Concerns about physical appearance and self-harm.
5. Depressed mood due to weather and family conflicts.
6. Physical health issues, difficulty eating, and vomiting.
7. Mental health struggles, desire for self-harm, and suicidal ideation.
8. Disappointment with treatment and desire for self-harm.
9. Desire for privacy and autonomy.
10. Seeking validation and understanding from others.
11. Friction with family and peers.
12. Frustration with therapy and hospitalization.
13. Persistent desire for self-harm and suicidal thoughts.
14. Discontent with hospitalization and treatment.
15. Desperation for happiness and companionship.

The summarized timeline portrays a cycle of emotional distress, self-harm impulses, dissatisfaction with treatment, and longing for understanding and support.

Personalised Longitudinal Natural Language Processing: Objectives

Challenges with sensitive data

- ***Privacy: cannot use LMs we can’t control ourselves***

Challenges with sensitive data

- ***Privacy: cannot use LLMs we can't control ourselves***

Challenges with sensitive data

- *Privacy: cannot use LLMs we can't control ourselves*
- *Long sequences: generated summaries don't capture most important events or preserve temporal order*

Challenges with sensitive data

- *Privacy: cannot use LLMs we can't control ourselves*
- *Long sequences: generated summaries don't capture most important events or preserve temporal order*
- *PLMs not temporally robust*

Challenges with sensitive data

- *Privacy: cannot use LLMs we can't control ourselves*
- *Long sequences: generated summaries don't capture most important events or preserve temporal order*
- *PLMs not temporally robust*
- *Hallucinated information in generating data*

Challenges with sensitive data

- *Privacy: cannot use LLMs we can't control ourselves*
- *Long sequences: generated summaries don't capture most important events or preserve temporal order*
- *LLMs not temporally robust*
- *Hallucinated information in generating data*
- *No preservation of disfluencies (ums, er, pauses, slurs, flight of ideas)*

User Generated Content (UGC)

Individuals continuously generate large amounts of digital content:

- Social Media (posts, images, videos, location, ...)
- Smart Devices (mobile phones, fitness devices, ...)

User Generated Content (UGC)

Individuals continuously generate large amounts of digital content:

- Social Media (posts, images, videos, location, …)
- Smart Devices (mobile phones, fitness devices, …)

Behavioural Cues

Personalised longitudinal NLP Goals

- Methods for **capturing changes in individuals' language over time**

Personalised longitudinal NLP Goals

- Methods for **capturing changes in individuals' language over time**
- Sensors for capturing digital markers from language & heterogeneous UGC to understand the evolution of an individual over time (**Time-sensitive sensors from UGC**)

Personalised longitudinal NLP Goals

- Methods for **capturing changes in individuals' language over time**
- Sensors for capturing digital biomarkers from language & heterogeneous UGC to understand the evolution of an individual over time (**Time-sensitive sensors from UGC**)
- **Make a significant contribution to mental health**

Goals & Use Cases in Mental Health

Cognition changes

Mood instability

A few months later…

:):

Context: Core Challenges in NLP

- **Small Data & Transfer Learning across Domains**
- **Obtaining Useful & Interpretable Representations**
- **Model fairness & privacy preservation**
- **Collaborative Human-Computer Process**
- Resolving Complex Linguistic Phenomena

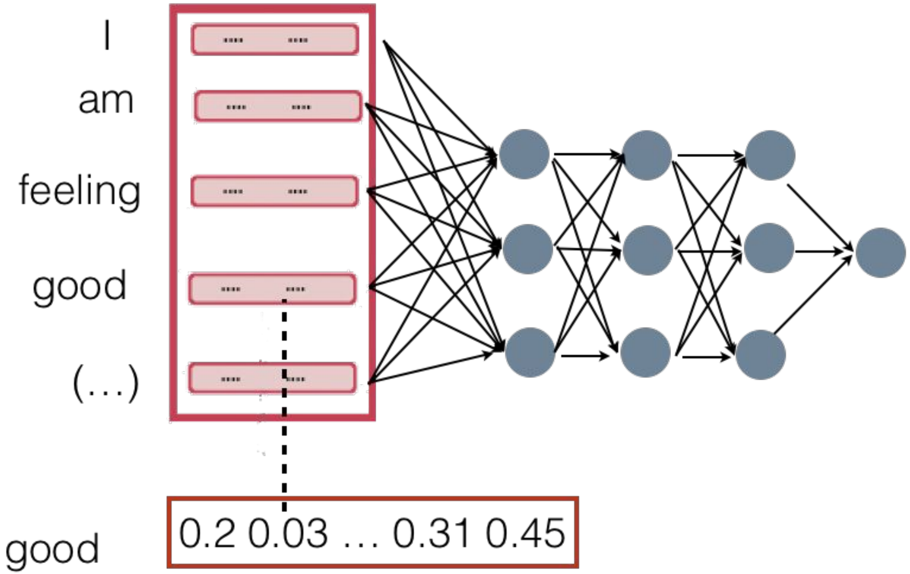
Context: Core Challenges in NLP

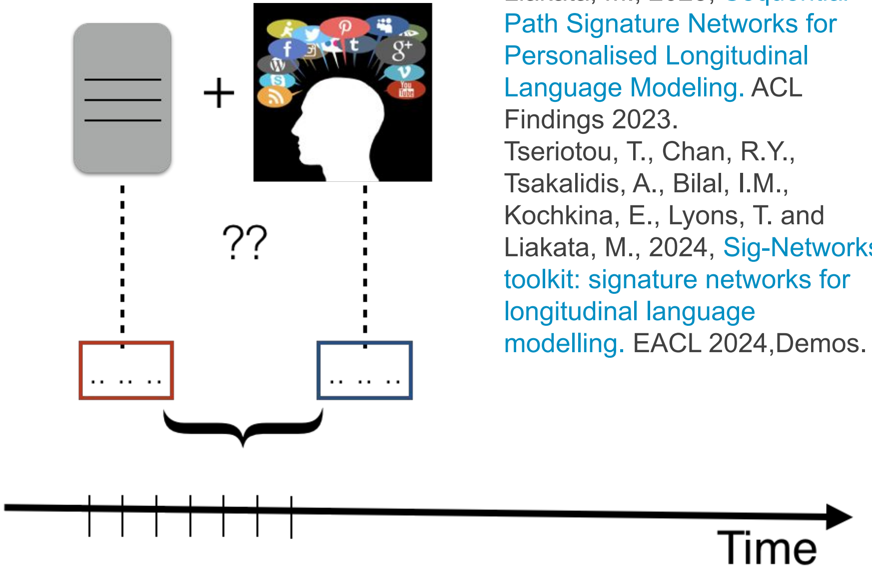
- **Small Data & Transfer Learning across Domains**
- **Obtaining Useful & Interpretable Representations**
- **Model fairness & privacy preservation**
- **Collaborative Human-Computer Process**
- Resolving Complex Linguistic Phenomena

\+ **Time-sensitivity**

User Representations

Objective 1: User representations from asynchronous longitudinal UGC





Tseriotou, T., Tsakalidis, A., Foster, P., Lyons, T. and Liakata, M., 2023, Sequential Path Signature Networks for Personalised Longitudinal Language Modeling. ACL Findings 2023.
Tseriotou, T., Chan, R.Y., Tsakalidis, A., Bilal, I.M., Kochkina, E., Lyons, T. and Liakata, M., 2024, Sig-Networks toolkit: signature networks for longitudinal language modelling. EACL 2024,Demos.

Gkoumas, D., Wang, B., Tsakalidis, A., Wolters, M., Zubiaga, A., Purver, M., and Liakata, M., 2024. A longitudinal multi-modal dataset for dementia monitoring and diagnosis. Journal of Language Resources and Evaluation.

Data annotation & Synthetic Language Generation

Datasets	Size	Time span	Data	Labels
NYU Students	29 users: 110K posts 42GB data	4-5 months p.i.	textual, mobile phone	PANAS, WEMWBS
Dementia Cohort	30 users: ~200GB	12 months p.i.	conversation, textual, extra-linguistic	diagnosis, cognitive tests
TalkLife	> 10m posts > 37m actions	2010-19	posts, demographics, network	self-reported mood MoC for 500 timelines, 18K posts
Mumsnet	> 62m posts, 1m users	2000-19		comment category
Reddit	83K users 15m posts	2015-21		MoC for 256 timelines, 6K+ posts Risk: 186 users

Objective 2: Addressing data privacy, ethics and data sparsity issues in real-world datasets.

Data annotation & Synthetic Language Generation

Datasets	Size	Time span	Data	Labels
NYU Students	29 users: 110K posts 42GB data	4-5 months p.i.	textual, mobile phone	PANAS, WEMWBS
Dementia Cohort	30 users: ~200GB	12 months p.i.	conversation, textual, extra-linguistic	diagnosis, cognitive tests
TalkLife	> 10m posts > 37m actions	2010-19	posts, demographics, network	self-reported mood MoC for 500 timelines, 18K posts
Mumsnet	> 62m posts, 1m users	2000-19		comment category
Reddit	83K users 15m posts	2015-21		MoC for 256 timelines, 6K+ posts Risk: 186 users

Tsakalidis et al. Identifying Moments of Change from Longitudinal User Text. ACL 2022.

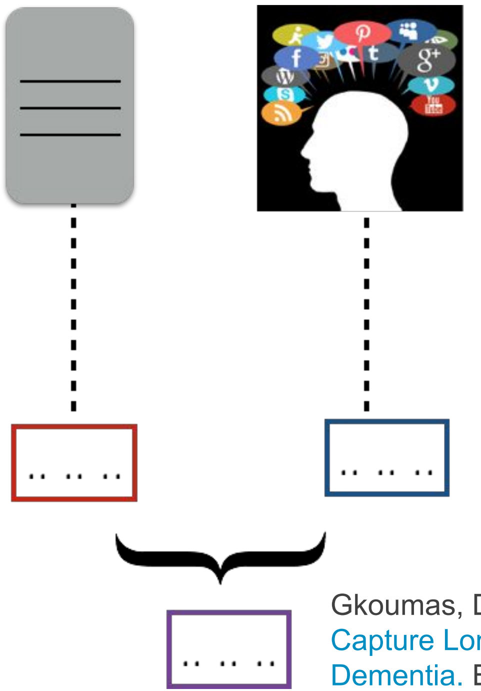
Tsakalidis et al. Overview of the CLPsych 2022 shared task: Capturing moments of change in longitudinal user posts. NAACL 2022.

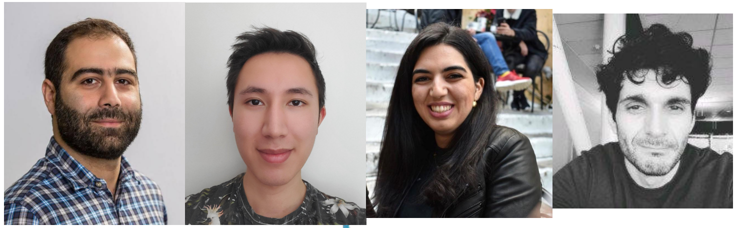
Hills, A., Tsakalidis, A., Nanni, F., Zachos, I., Liakata, M., 2023, May. Creation and evaluation of timelines for longitudinal user posts. EACL 2023.

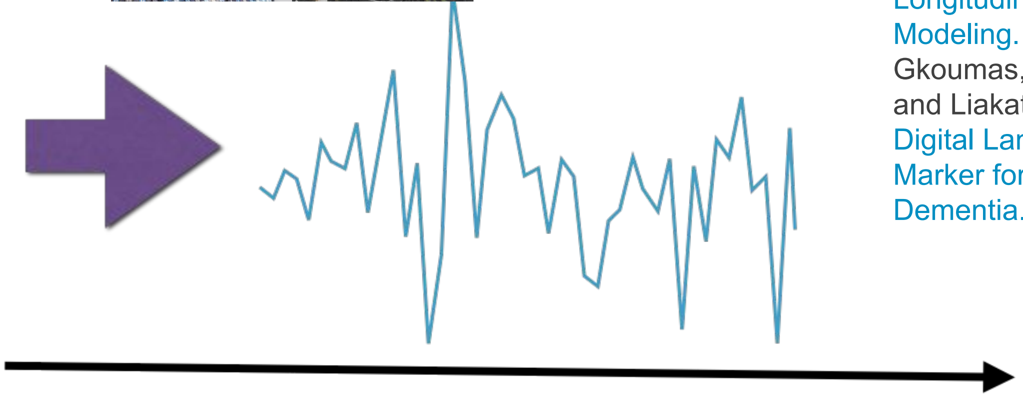
Chim, J., Ive.J. and Liakata,M. 2024. Evaluating Synthetic Language Generations from User Generated Text. Under review.

Normal States & Anomalies Identification

Objective 3: Personalised normal states and anomalies.







Time

Gkoumas, D., Purver, M. and Liakata, M., 2023. Reformulating NLP tasks to Capture Longitudinal Manifestation of Language Disorders in People with Dementia. EMNLP 2023

Tsakalidis et al. 2022. Identifying Moments of Change from Longitudinal User Text. ACL 2022.

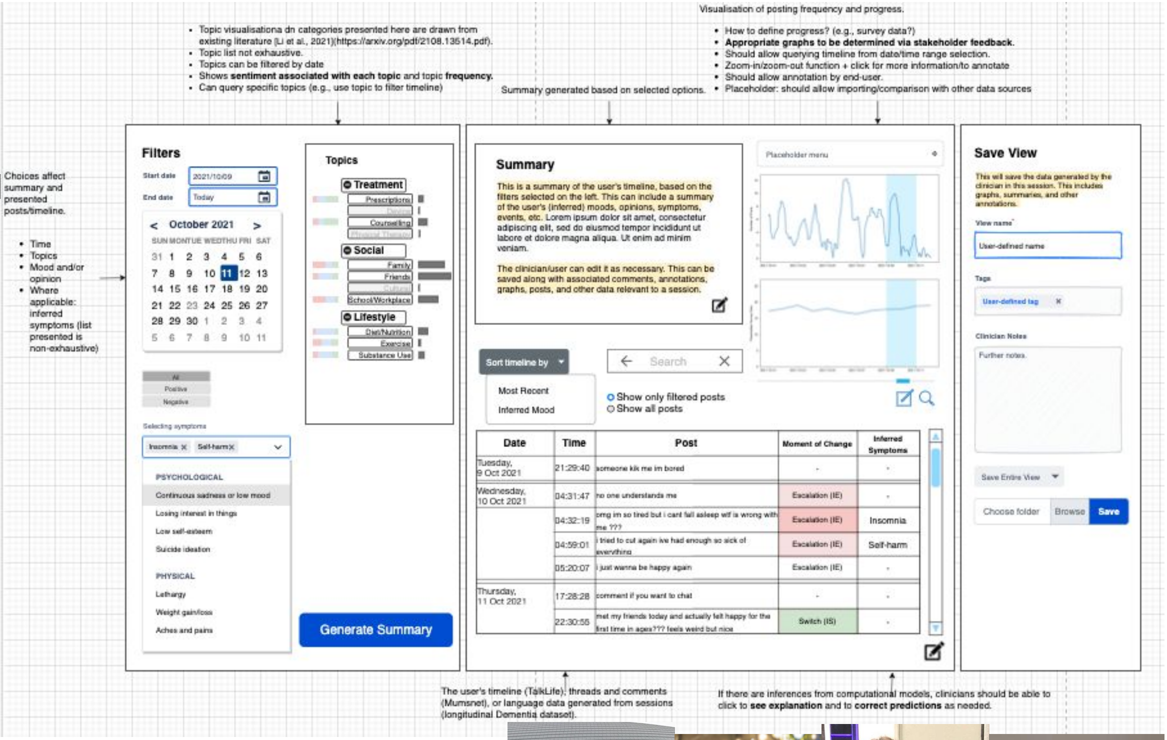
Hills, A., Tsakalidis, A., Nanni, F., Zachos, I., Liakata, M., 2023. Creation and evaluation of timelines for longitudinal user posts. EACL 2023.

Tseriotou, T., et al. 2023. Sequential Path Signature Networks for Personalised Longitudinal Language Modeling. ACL Findings.

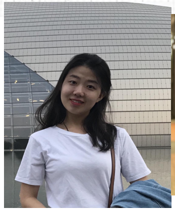
Gkoumas, D., Tsakalidis, A. and Liakata, M., 2023. A Digital Language Coherence Marker for Monitoring Dementia. EMNLP 2023

Summarisation, Explainability & Interventions

Objective 4: Real-world evaluation settings, interpretable summaries, instrument co-creatio



Bilal, I.M., Wang, B., Tsakalidis, A., Nguyen, D., Procter, R. and Liakata, M., 2022. Template-based Abstractive Microblog Opinion Summarization. TACL.
Song, J., Bilal, I.M., Tsakalidis, A., Procter R. and Liakata, M., 2022. Unsupervised Opinion Summarisation in the Wasserstein Space. EMNLP.
Bilal, I.M., Nakov, P., Procter, R. and Liakata, M., 2024. Generating Zero-shot Abstractive Explanations for Rumour Verification. (under review)
Chim, J. et al. Overview of the CLPsych 2024 Shared Task: Leveraging Large Language Models to Identify Evidence of Suicidality Risk in Online Posts. EACL 2024.
Song, J., Chim, J., Tsakalidis, A., Ive, J., Atzil-Slonim, D. and Liakata, M., 2024. Combining Hierachical VAEs with LLMs for clinically meaningful timeline summarisation in social media. (ACL Findings)




Identifying moments of change from longitudinal user text

Tsakalidis, A., Nanni, F., Hills, A., Chim, J., Song, J., & Liakata, M. (2022, May). Identifying Moments of Change from Longitudinal User Text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 4647-4660).

Tsakalidis, A., Chim, J., Bilal, I. M., Zirikly, A., Atzil-Slonim, D., Nanni, F., ... & Liakata, M. (2022). Overview of the CLPsych 2022 Shared Task: Capturing Moments of Change in Longitudinal User Posts. *CLPsych 2022*, 184.

What is a Moment of Change (MoC)?

Timeline: a sequence of posts/interactions between two moments in time (dates).

MoC: A change in an individual's behaviour or mood, **after a particular event** or the identifiable start of a trend such as **symptom onset**.

A MoC **may be a particular point in time or a longer period**. It can be an obvious change or a more subtle change in one's mood.

Switches & Escalations

Two types of mood changes: **drastic** (a switch) or **gradual** (an escalation).

Switches & Escalations

Two types of mood changes: **drastic** (a switch) or **gradual** (an escalation).

A ***Switch*** is a drastic change, i.e. one's mood goes from relatively neutral or positive to negative and vice versa.

Switches & Escalations

Two types of mood changes: **drastic** (a switch) or **gradual** (an escalation).

A ***Switch*** is a drastic change, i.e. one's mood goes from relatively neutral or positive to negative and vice versa.

An ***Escalation*** is a gradual intensification, i.e. one's mood goes from bad to worse or from positive to more positive.

Switches & Escalations

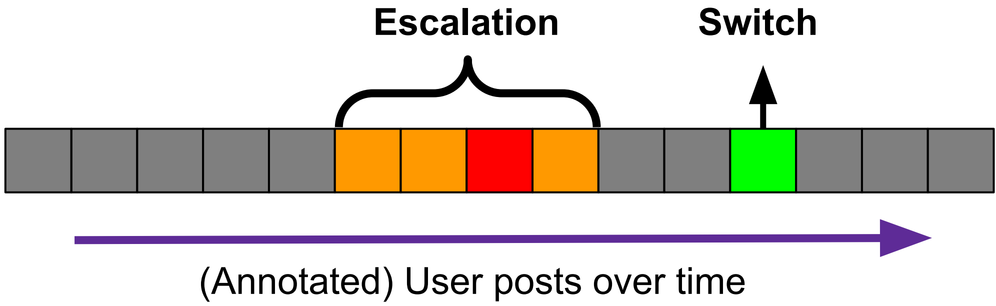
Two types of mood changes: **drastic** (a switch) or **gradual** (an escalation).

A ***Switch*** is a drastic change, i.e. one's mood goes from relatively neutral or positive to negative and vice versa.

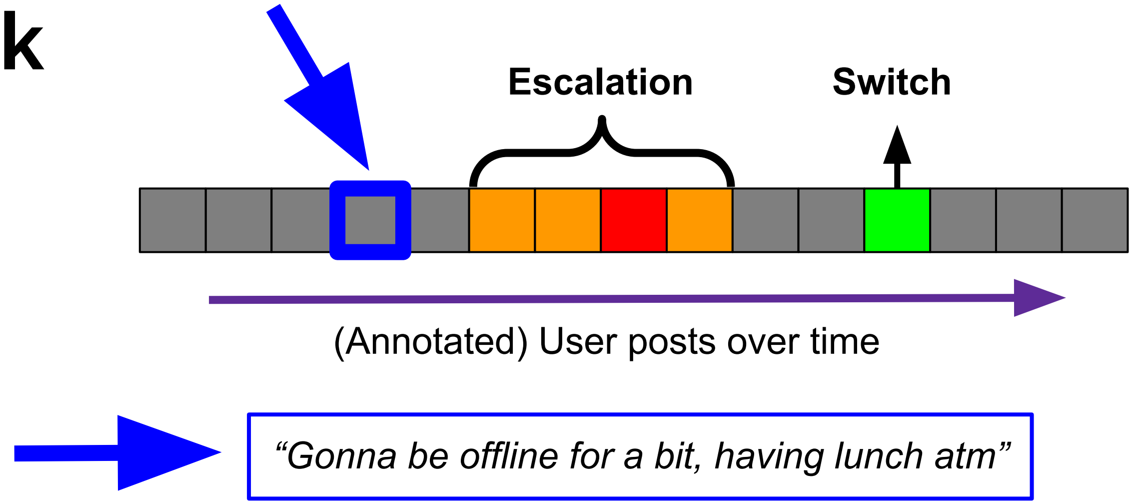
An ***Escalation*** is a gradual intensification, i.e. one's mood goes from bad to worse or from positive to more positive.

Escalations and switches can take place over several posts or days or within a single post. Therefore we are interested in **the range of posts** denoting a particular switch or escalation.

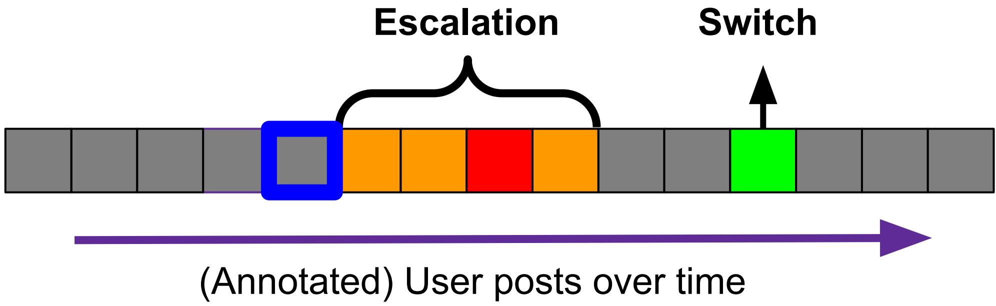
Capturing MoC: Task



Capturing MoC: Task



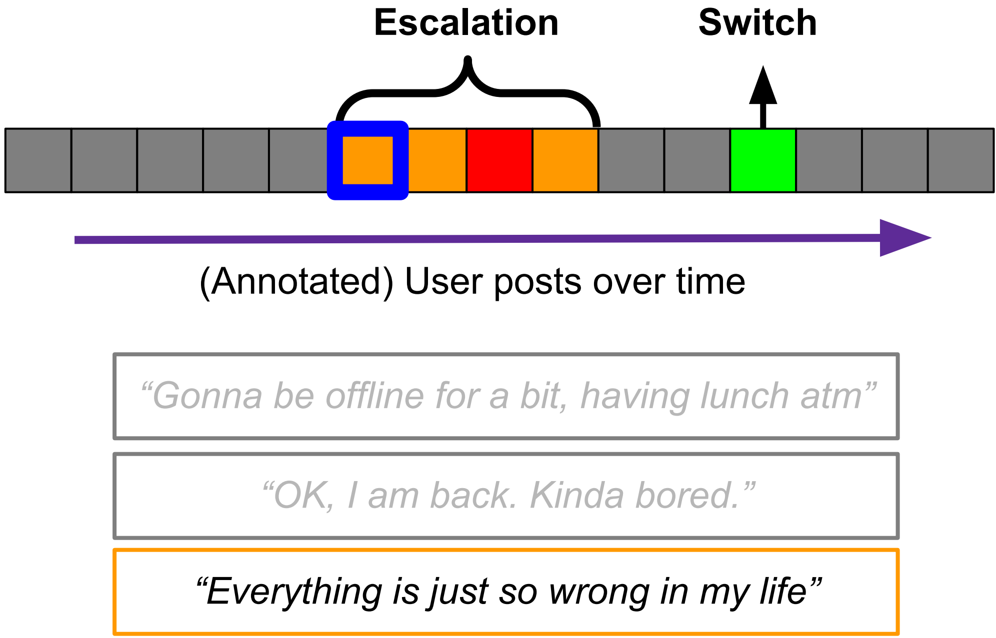
Capturing MoC: Task



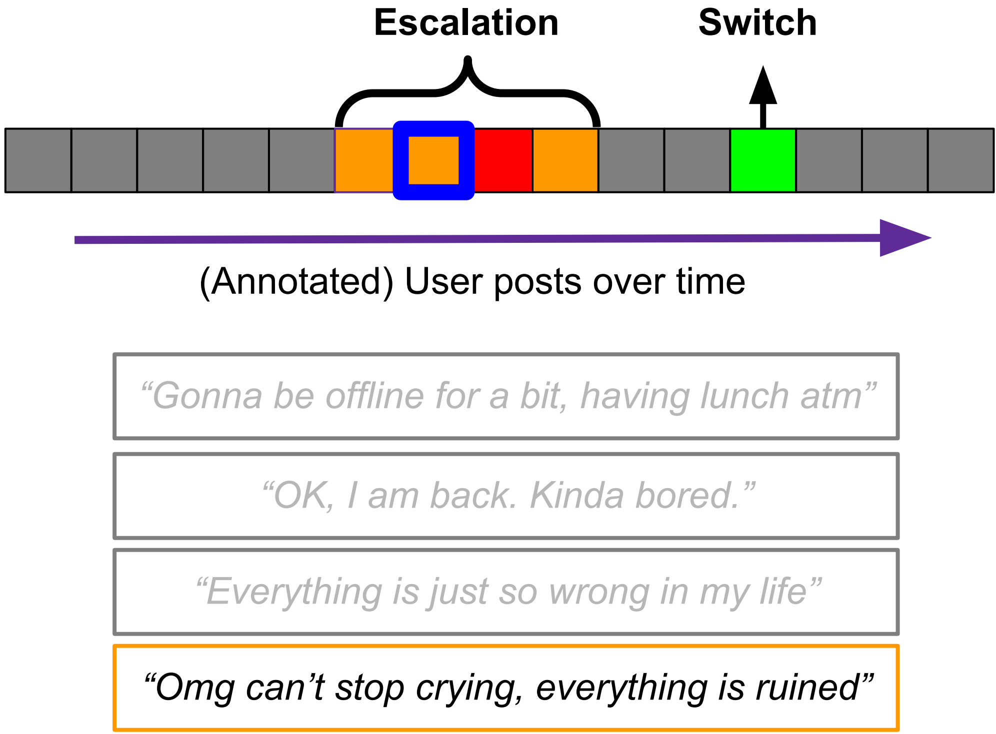
“Gonna be offline for a bit, having lunch atm”

“OK, I am back. Kinda bored.”

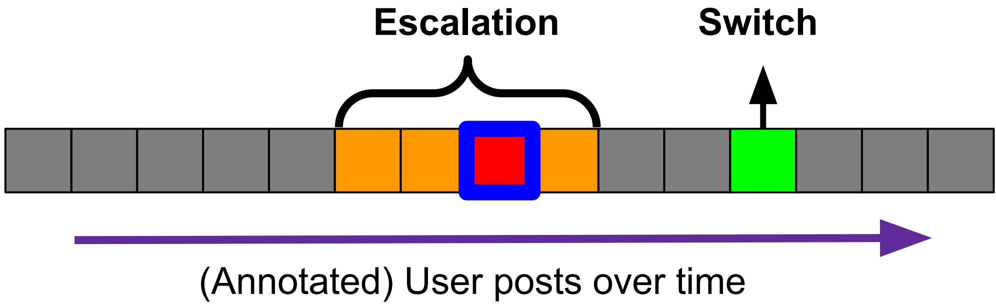
Capturing MoC: Task



Capturing MoC: Task



Capturing MoC: Task



"Gonna be offline for a bit, having lunch atm"

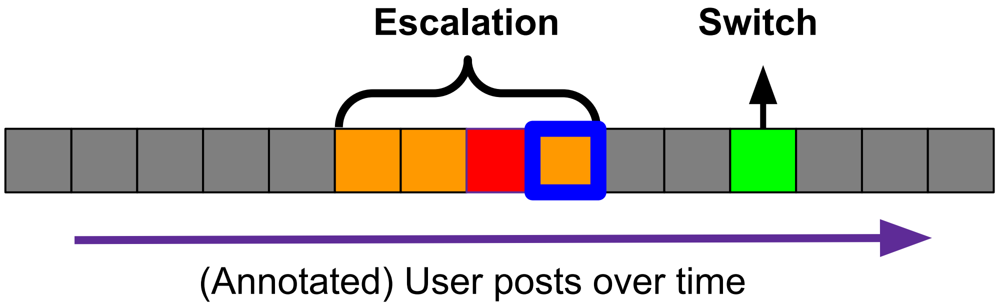
"OK, I am back. Kinda bored."

"Everything is just so wrong in my life"

"Omg can't stop crying, everything is ruined"

"Need someone before I do something stupid !PLEASE HELP!"

Capturing MoC: Task



"OK, I am back. Kinda bored."

"Everything is just so wrong in my life"

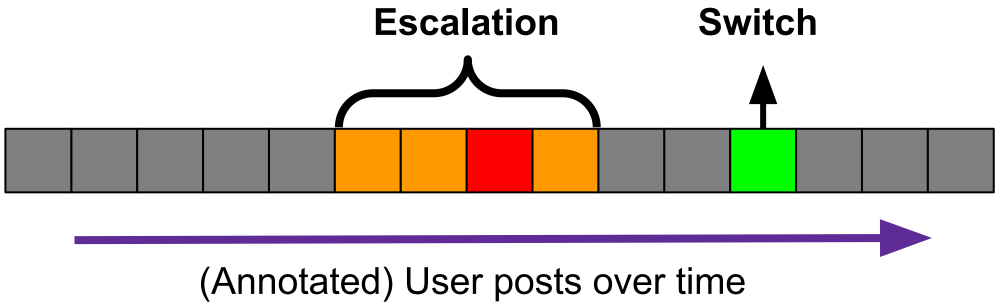
"Omg can't stop crying, everything is ruined"

"Need someone before I do something stupid !PLEASE HELP!"

"Wish things were differently.. Miss my gf.."

Capturing MoC: Task

Aim: Detect changes in a user's mood based on the content they share online.

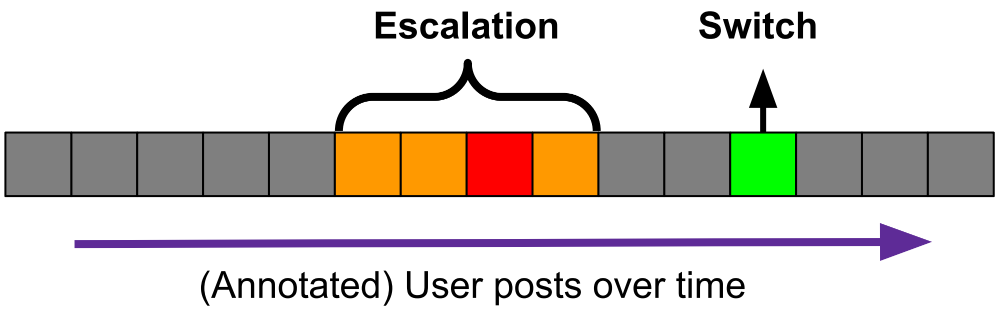


Capturing MoC: Task

Aim: Detect changes in a user's mood based on the content they share online.

Three-label, post-level task:

- IS (In-Switch)
- IE (In-Escalation)
- O (None)



Capturing MoC: Task

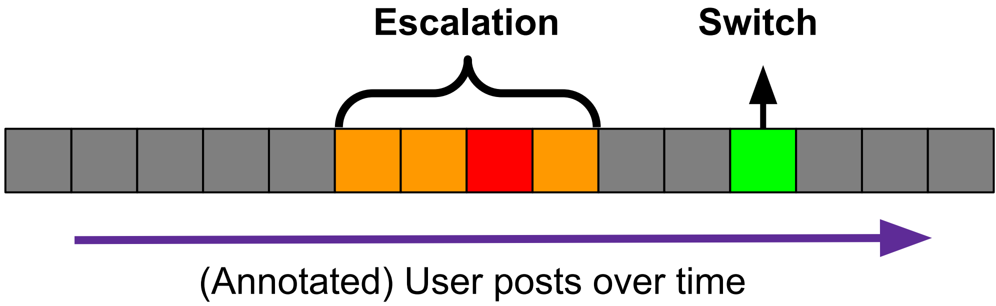
Aim: Detect changes in a user's mood based on the content they share online.

Three-label, post-level task:

- IS (In-Switch)
- IE (In-Escalation)
- O (None)

Post-level Classification Setting:

- Precision/Recall/F1 (macro)
- Timeline-sensitive metrics



Capturing MoC: Task

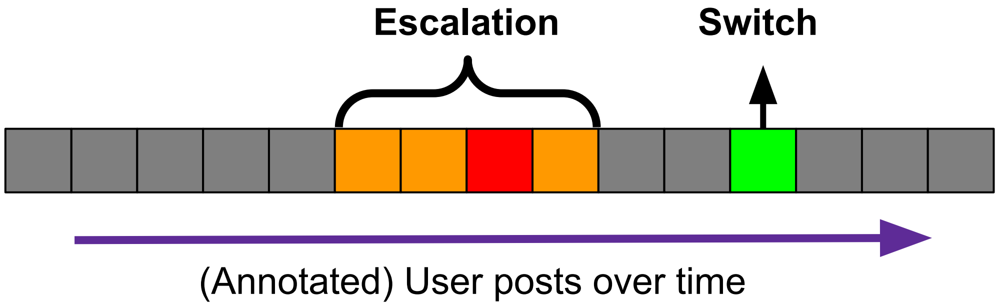
Aim: Detect changes in a user's mood based on the content they share online.

Three-label, post-level task:

- IS (In-Switch)
- IE (In-Escalation)
- O (None)

Post-level Classification Setting:

- Precision/Recall/F1 (macro)
- Timeline-sensitive metrics



- ACL '22 paper
- CLPsych '22 Shared Task

Capturing MoC: Datasets

	Source	#users	#timelines (duration)	#posts	Split	O	IE	IS
ACL	TalkLife	500	500 (2 weeks)	18,702	5-fold CV	84.5%	10.8%	4.7%
CLPsych	Reddit	185	255 (2 months)	6,195	80/20	77.6%	15.8%	6.6%

CLPsych ST additional task: Identify risk level of a user (user-level classification task with 3 classes)

Capturing MoC: Datasets

	Source	#users	#timelines (duration)	#posts	Split	O	IE	IS
ACL	TalkLife	500	500 (2 weeks)	18,702	5-fold CV	84.5%	10.8%	4.7%
CLPsych	Reddit	185	255 (2 months)	6,195	80/20	77.6%	15.8%	6.6%

CLPsych ST additional task: Identify risk level of a user (user-level classification task with 3 classes)

Task-specific Requirements

Imbalanced Data(set):

- Problem: More like an outlier/anomaly detection task
- Solution: Focal Loss *vs* Cross-Entropy Loss

Task-specific Requirements

Imbalanced Data(set):

- Problem: More like an outlier/anomaly detection task
- Solution: Focal Loss *vs* Cross-Entropy Loss

Contextual Information:

- Problem: Single-post classification lacks the context of the post
- Solution: Extract task-oriented representations from isolated posts; feed RNN

Task-specific Requirements

Imbalanced Data(set):

- Problem: More like an outlier/anomaly detection task
- Solution: Focal Loss *vs* Cross-Entropy Loss

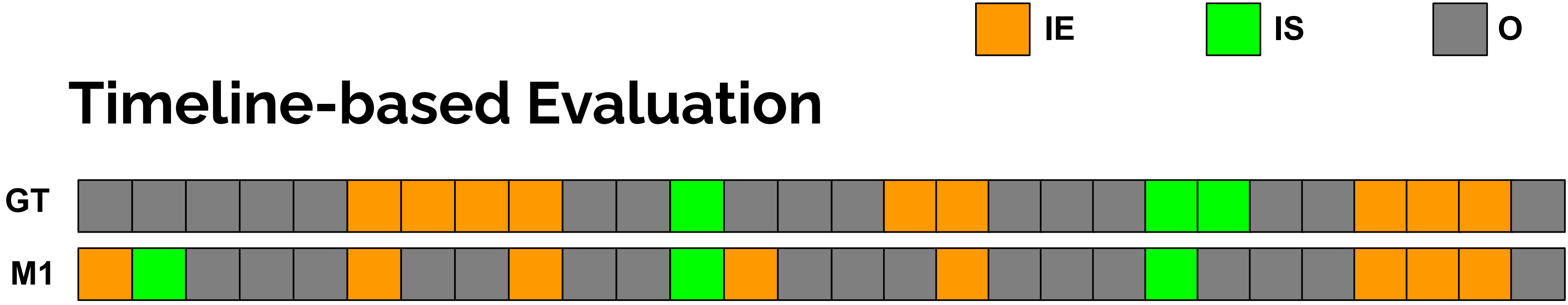
Contextual Information:

- Problem: Single-post classification lacks the context of the post
- Solution: Extract task-oriented representations from isolated posts; feed RNN

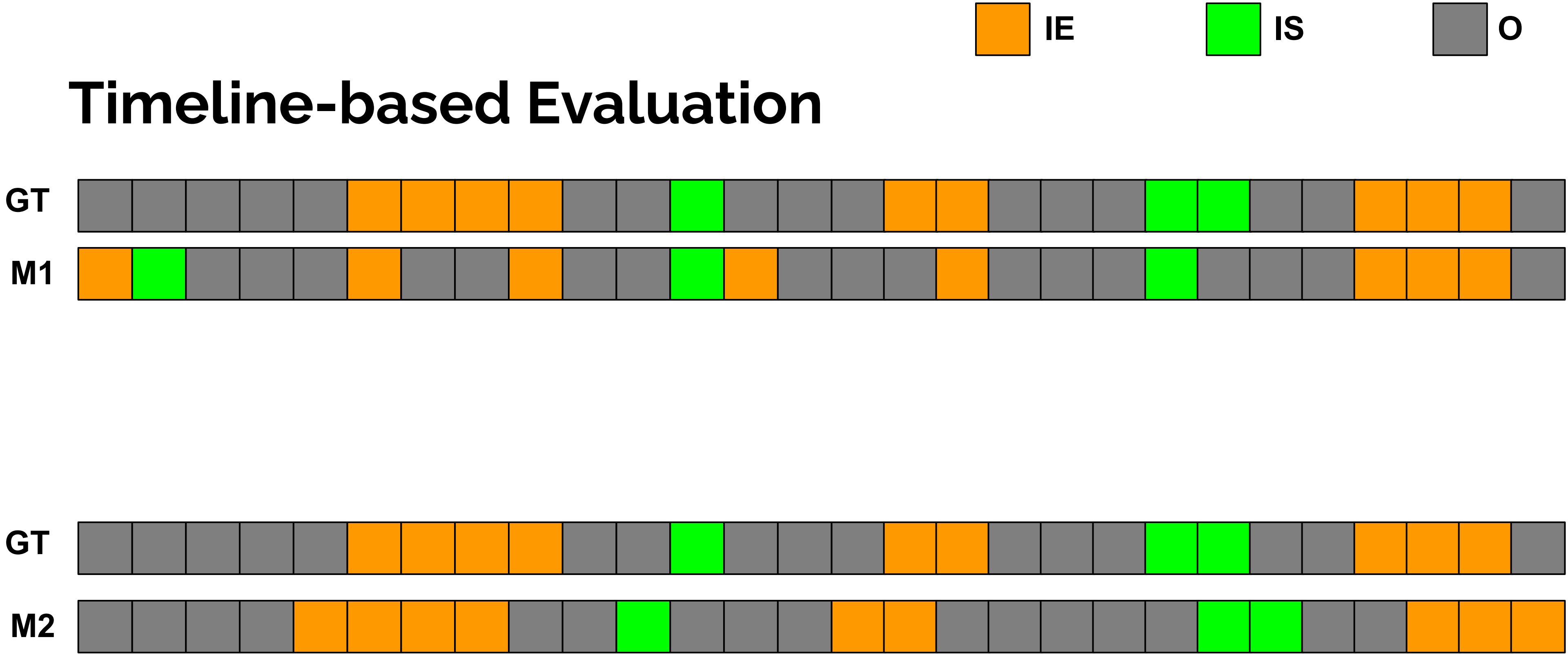
Evaluation:

- Problem: Single-post evaluation, timeline-based task
- Solution: Adjusting existing metrics operating on a window basis

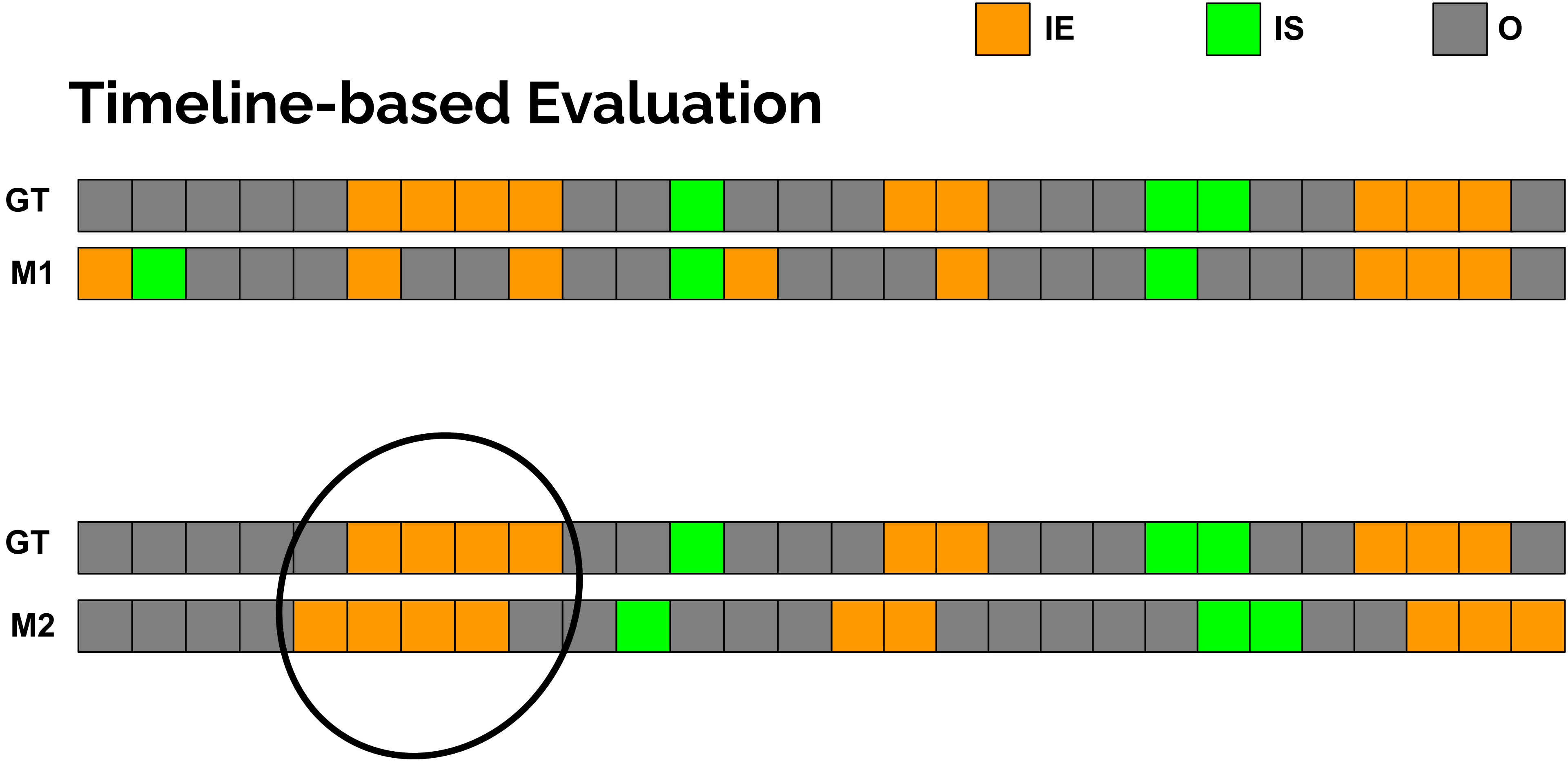
Timeline-based Evaluation

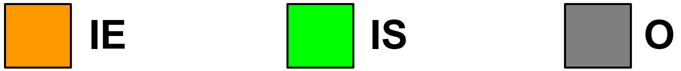


Timeline-based Evaluation



Timeline-based Evaluation

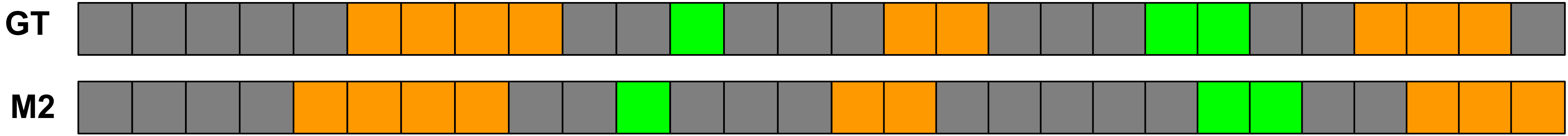




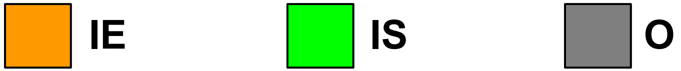
Timeline-based Evaluation

$r(IE) = \frac{6}{9}$ $p(IE) = \frac{6}{8}$ $r(IS) = \frac{2}{3}$ $p(IS) = \frac{2}{3}$

M1 outperforms M2 in almost all metrics!



$r(IE) = \frac{6}{9}$ $p(IE) = \frac{6}{9}$ $r(IS) = \frac{1}{3}$ $p(IS) = \frac{1}{3}$

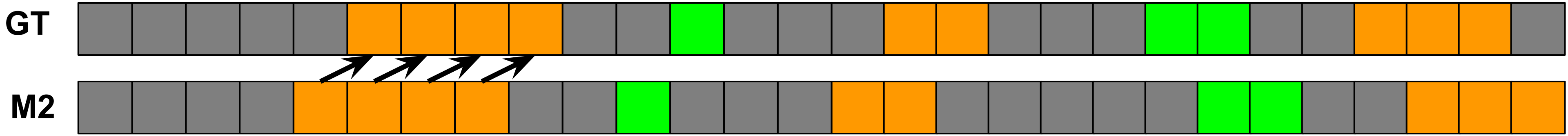


Timeline-based Evaluation

$r(IE) = \frac{6}{9}$ $p(IE) = \frac{6}{8}$

Solution: window-based metrics

$r(IS) = \frac{2}{3}$ $p(IS) = \frac{2}{3}$

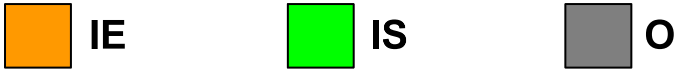


$r(IE) = \frac{9}{9}$ $p(IE) = \frac{9}{9}$

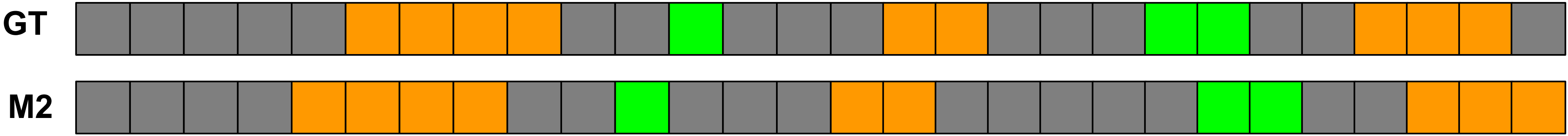
Metrics for *w*=1

$r(IS) = \frac{3}{3}$ $p(IS) = \frac{3}{3}$

(van den Burg and Williams, 2020)

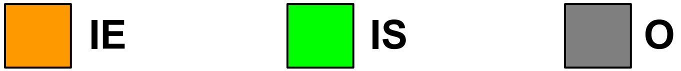


Timeline-based Evaluation



Question: How about detecting regions of change though?

(Arbelaez et al., 2010)



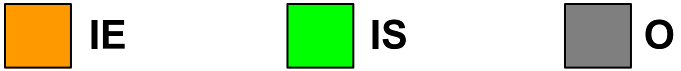
Timeline-based Evaluation

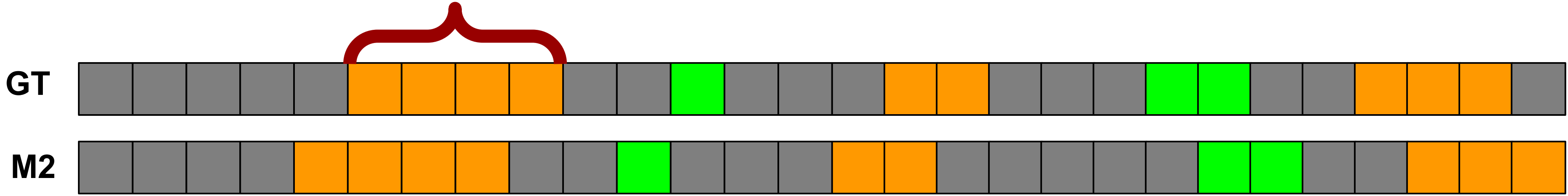
Question: How about detecting regions of change though?

Solution: coverage-based metrics

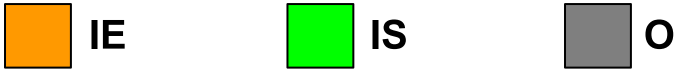
(Arbelaez et al., 2010)

Timeline-based Evaluation

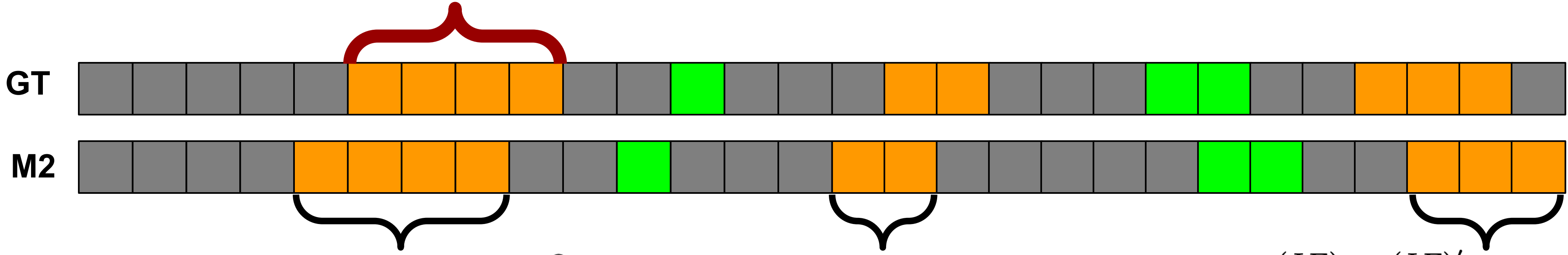




$$O(R, R') = \frac{R \bigcap R'}{R \bigcup R'}$$



Timeline-based Evaluation

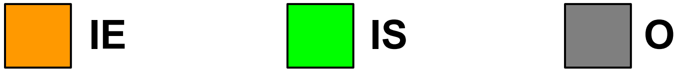


$$O(R_0^{(IE)}, R_0^{(IE)'}) = \frac{3}{5}$$

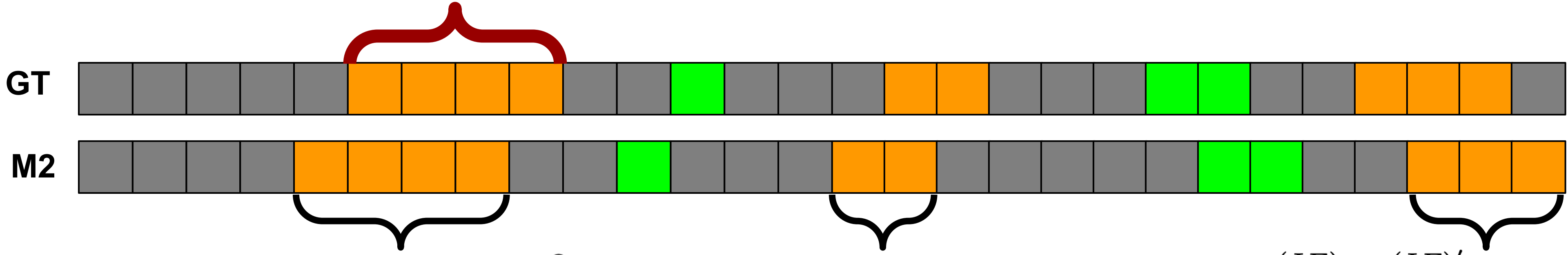
$$O(R_0^{(IE)}, R_1^{(IE)'}) = 0$$

$$O(R_0^{(IE)}, R_2^{(IE)'}) = 0$$

$$O(R, R') = \frac{R \bigcap R'}{R \bigcup R'}$$



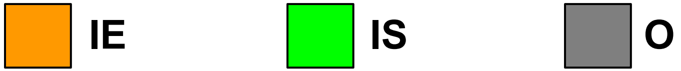
Timeline-based Evaluation



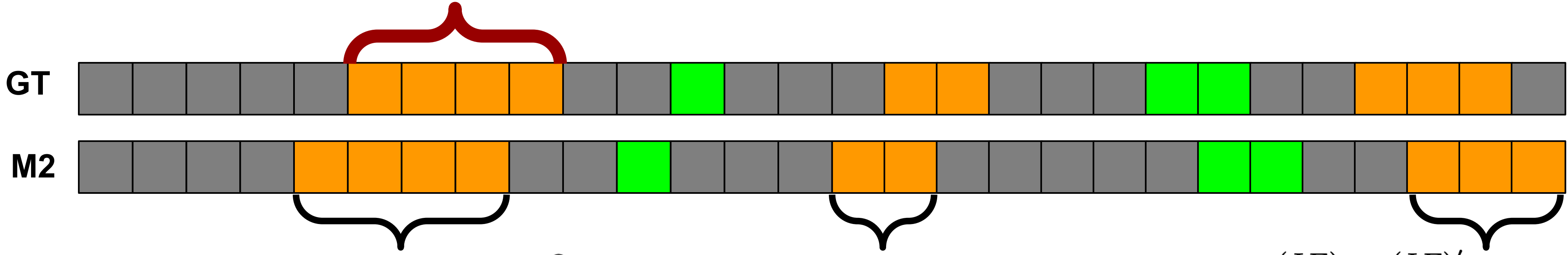
$$O(R_0^{(IE)}, R_0^{(IE)'}) = \frac{3}{5} \qquad O(R_0^{(IE)}, R_1^{(IE)'}) = 0 \qquad O(R_0^{(IE)}, R_2^{(IE)'}) = 0$$

$$C_r^{(L)}(M \rightarrow GT) = \frac{1}{\sum_{R_{GT}^{(L)}} |R_{GT}^{(L)}|} \sum_{R_{GT}^{(L)}} |R_{GT}^{(L)}| \cdot max_{R_M^{(L)}} \{O(R_{GT}^{(L)}, R_M^{(L)})\}$$

$$O(R, R') = \frac{R \bigcap R'}{R \bigcup R'}$$



Timeline-based Evaluation

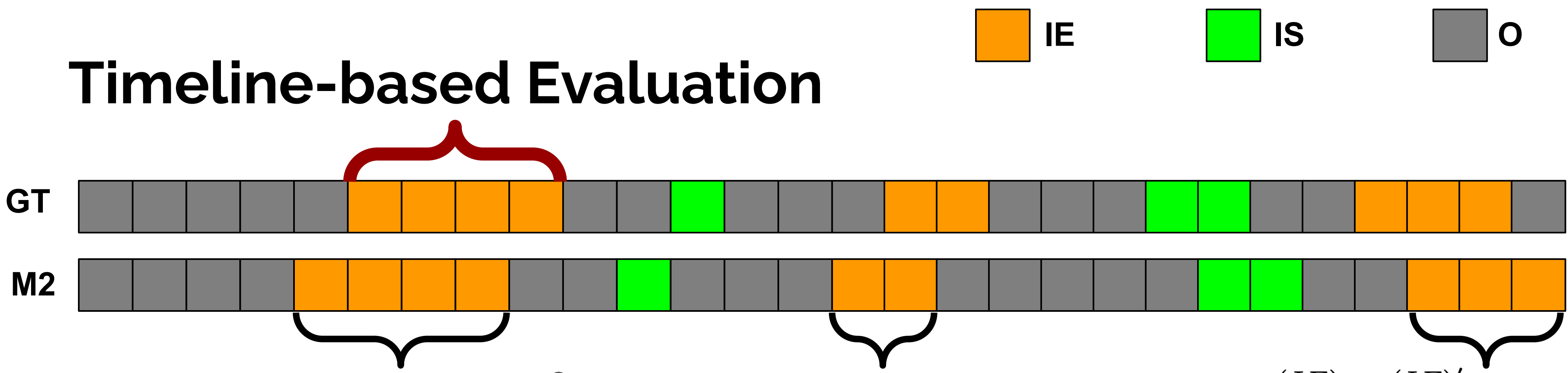


$$O(R_0^{(IE)}, R_0^{(IE)'}) = \frac{3}{5} \qquad O(R_0^{(IE)}, R_1^{(IE)'}) = 0 \qquad O(R_0^{(IE)}, R_2^{(IE)'}) = 0$$

$$C_r^{(L)}(M \to GT) = \frac{1}{\sum_{R_{GT}^{(L)}} |R_{GT}^{(L)}|} \sum_{R_{GT}^{(L)}} |R_{GT}^{(L)}| \cdot max_{R_M^{(L)}} \{O(R_{GT}^{(L)}, R_M^{(L)})\}$$

$$O(R, R') = \frac{R \bigcap R'}{R \bigcup R'}$$

Timeline-based Evaluation



$$O(R_0^{(IE)}, R_0^{(IE)'}) = \frac{3}{5} \qquad O(R_0^{(IE)}, R_1^{(IE)'}) = 0 \qquad O(R_0^{(IE)}, R_2^{(IE)'}) = 0$$

$$C_r^{(L)}(M \rightarrow GT) = \frac{1}{\sum_{R_{GT}^{(L)}} |R_{GT}^{(L)}|} \sum_{R_{GT}^{(L)}} |R_{GT}^{(L)}| \cdot max_{R_M^{(L)}} \{O(R_{GT}^{(L)}, R_M^{(L)})\}$$

$$C_p^{(L)}(M \rightarrow GT) = \frac{1}{\sum_{R_M^{(L)}} |R_M^{(L)}|} \sum_{R_M^{(L)}} |R_M^{(L)}| \cdot max_{R_{GT}^{(L)}} \{O(R_{GT}^{(L)}, R_M^{(L)})\}$$

Results (ACL'22)

		Post-level Evaluation												Coverage-based Metrics							
		IS			**IE**			**O**			**macro-avg**			**IS**		**IE**		**O**		**macro-avg**	
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	C_p	C_r	C_p	C_r	C_p	C_r	C_p	C_r
Naïve	**Majority**																				
	Random																				
Post-level	**RF-tfidf**																				
	BiLSTM-we																				
	BERT(ce)																				
	BERT(f)																				
Timeline-level	**FSD**																				
	EM-TR																				
	EM-DM																				
	SCD-OP																				
	SCD-FP																				
	BiLSTM-bert																				

Results (ACL’22)

		Post-level Evaluation												Coverage-based Metrics							
		IS			IE			O			macro-avg			IS		IE		O		macro-avg	
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	C_p	C_r	C_p	C_r	C_p	C_r	C_p	C_r
Naïve	**Majority**	–	.000	.000	–	.000	.000	.845	1.000	.916	.282	.333	.305	–	.000	–	.000	.619	.559	.206	.186
	Random	.047	.047	.047	.108	.108	.108	.845	.845	.845	.333	.333	.333	.031	.045	.033	.096	.386	.452	.150	.198
Post-level	**RF-tfidf**																				
	BiLSTM-we																				
	BERT(ce)																				
	BERT(f)																				
Timeline-level	**FSD**																				
	EM-TR																				
	EM-DM																				
	SCD-OP																				
	SCD-FP																				
	BiLSTM-bert																				

Results (ACL'22)

		Post-level Evaluation												Coverage-based Metrics							
		IS			IE			O			macro-avg			IS		IE		O		macro-avg	
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	C_p	C_r	C_p	C_r	C_p	C_r	C_p	C_r
Naïve	**Majority**	–	.000	.000	–	.000	.000	.845	1.000	.916	.282	.333	.305	–	.000	–	.000	.619	.559	.206	.186
	Random	.047	.047	.047	.108	.108	.108	.845	.845	.845	.333	.333	.333	.031	.045	.033	.096	.386	.452	.150	.198
Post-level	**RF-tfidf**	.294	.006	.011	**.568**	.087	.151	.852	**.991**	**.917**	.571	.361	.360	.250	.005	.152	.087	.632	.602	.345	.231
	BiLSTM-we	.245	.119	.160	.416	.347	.378	.878	.923	.900	.513	.463	.479	.173	.091	.138	.330	.557	.606	.289	.342
	BERT(ce)	.285	.186	.222	.454	.368	.406	.883	.921	.901	.540	.492	.510	.247	.163	.172	**.344**	.578	**.621**	.332	.376
	BERT(f)	.260	**.321**	**.287**	.401	**.478**	.436	**.898**	.864	.881	.520	**.554**	**.534**	.227	**.269**	.160	**.423**	.503	.567	.297	**.420**
Timeline-level	**FSD**																				
	EM-TR																				
	EM-DM																				
	SCD-OP																				
	SCD-FP																				
	BiLSTM-bert																				

Results (ACL'22)

		Post-level Evaluation												Coverage-based Metrics							
		IS			IE			O			macro-avg			IS		IE		O		macro-avg	
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	C_p	C_r	C_p	C_r	C_p	C_r	C_p	C_r
Naïve	**Majority**	–	.000	.000	–	.000	.000	.845	1.000	.916	.282	.333	.305	–	.000	–	.000	.619	.559	.206	.186
	Random	.047	.047	.047	.108	.108	.108	.845	.845	.845	.333	.333	.333	.031	.045	.033	.096	.386	.452	.150	.198
Post-level	**RF-tfidf**	.294	.006	.011	**.568**	.087	.151	.852	**.991**	**.917**	.571	.361	.360	.250	.005	.152	.087	.632	.602	.345	.231
	BiLSTM-we	.245	.119	.160	.416	.347	.378	.878	.923	.900	.513	.463	.479	.173	.091	.138	.330	.557	.606	.289	.342
	BERT(ce)	.285	.186	.222	.454	.368	.406	.883	.921	.901	.540	.492	.510	.247	.163	.172	**.344**	.578	**.621**	.332	.376
	BERT(f)	.260	**.321**	**.287**	.401	**.478**	.436	**.898**	.864	.881	.520	**.554**	**.534**	.227	**.269**	.160	**.423**	.503	.567	.297	**.420**
Timeline-level	**FSD**																				
	EM-TR																				
	EM-DM																				
	SCD-OP																				
	SCD-FP																				
	BiLSTM-bert																				

Results (ACL'22)

		Post-level Evaluation												Coverage-based Metrics							
		IS			IE			O			macro-avg			IS		IE		O		macro-avg	
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	C_p	C_r	C_p	C_r	C_p	C_r	C_p	C_r
Naïve	**Majority**	–	.000	.000	–	.000	.000	.845	1.000	.916	.282	.333	.305	–	.000	–	.000	.619	.559	.206	.186
	Random	.047	.047	.047	.108	.108	.108	.845	.845	.845	.333	.333	.333	.031	.045	.033	.096	.386	.452	.150	.198
Post-level	**RF-tfidf**	.294	.006	.011	**.568**	.087	.151	.852	**.991**	**.917**	.571	.361	.360	.250	.005	.152	.087	.632	.602	.345	.231
	BiLSTM-we	.245	.119	.160	.416	.347	.378	.878	.923	.900	.513	.463	.479	.173	.091	.138	.330	.557	.606	.289	.342
	BERT(ce)	.285	.186	.222	.454	.368	.406	.883	.921	.901	.540	.492	.510	.247	.163	.172	**.344**	.578	**.621**	.332	.376
	BERT(f)	.260	**.321**	**.287**	.401	**.478**	.436	**.898**	.864	.881	.520	**.554**	**.534**	.227	**.269**	.160	**.423**	.503	.567	.297	**.420**
Timeline-level	**FSD**	–	.000	.000	–	.000	.000	.845	1.000	.916	.282	.333	.305	–	.000	–	.000	.619	.559	.206	.186
	EM-TR	.344	.036	.065	.444	.248	.318	.865	**.957**	.909	.551	.414	.431	.297	.024	.273	.104	.639	.589	.403	.239
	EM-DM	**.533**	.118	.193	.479	.351	.405	.880	.948	.913	**.631**	.472	.504	**.347**	.023	**.363**	.177	.646	.592	**.452**	.264
	SCD-OP	.200	.005	.009	.478	.408	.440	.882	.947	.913	.520	.453	.454	.167	.001	.344	.180	**.663**	.609	.391	.263
	SCD-FP	.270	.082	.126	.503	.370	.426	.880	.944	.911	.551	.465	.488	.227	.039	.317	.254	.649	.611	.398	.301
	BiLSTM-bert	**.397**	**.264**	**.316**	**.568**	**.461**	**.508**	**.898**	.936	**.917**	**.621**	**.553**	**.580**	**.331**	**.197**	**.345**	.340	**.664**	**.656**	**.447**	**.398**

Results (ACL'22)

		Post-level Evaluation												Coverage-based Metrics							
		IS			IE			O			macro-avg			IS		IE		O		macro-avg	
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	C_p	C_r	C_p	C_r	C_p	C_r	C_p	C_r
Naïve	**Majority**	–	.000	.000	–	.000	.000	.845	1.000	.916	.282	.333	.305	–	.000	–	.000	.619	.559	.206	.186
	Random	.047	.047	.047	.108	.108	.108	.845	.845	.845	.333	.333	.333	.031	.045	.033	.096	.386	.452	.150	.198
Post-level	**RF-tfidf**	.294	.006	.011	**.568**	.087	.151	.852	**.991**	**.917**	.571	.361	.360	.250	.005	.152	.087	.632	.602	.345	.231
	BiLSTM-we	.245	.119	.160	.416	.347	.378	.878	.923	.900	.513	.463	.479	.173	.091	.138	.330	.557	.606	.289	.342
	BERT(ce)	.285	.186	.222	.454	.368	.406	.883	.921	.901	.540	.492	.510	.247	.163	.172	**.344**	.578	**.621**	.332	.376
	BERT(f)	.260	**.321**	**.287**	.401	**.478**	.436	**.898**	.864	.881	.520	**.554**	**.534**	.227	**.269**	.160	**.423**	.503	.567	.297	**.420**
Timeline-level	**FSD**	–	.000	.000	–	.000	.000	.845	1.000	.916	.282	.333	.305	–	.000	–	.000	.619	.559	.206	.186
	EM-TR	.344	.036	.065	.444	.248	.318	.865	**.957**	.909	.551	.414	.431	.297	.024	.273	.104	.639	.589	.403	.239
	EM-DM	**.533**	.118	.193	.479	.351	.405	.880	.948	.913	**.631**	.472	.504	**.347**	.023	**.363**	.177	.646	.592	**.452**	.264
	SCD-OP	.200	.005	.009	.478	.408	.440	.882	.947	.913	.520	.453	.454	.167	.001	.344	.180	**.663**	.609	.391	.263
	SCD-FP	.270	.082	.126	.503	.370	.426	.880	.944	.911	.551	.465	.488	.227	.039	.317	.254	.649	.611	.398	.301
	BiLSTM-bert	**.397**	**.264**	**.316**	**.568**	**.461**	**.508**	**.898**	.936	**.917**	**.621**	**.553**	**.580**	**.331**	**.197**	**.345**	.340	**.664**	**.656**	**.447**	**.398**

Summary:
(a) Importance of longitudinal modelling
(b) Tackling the class imbalance problem
(c) Contextualised representations

Tsakalidis, A., Nanni, F., Hills, A., Chim, J., Song, J., & Liakata, M. (2022, May). Identifying Moments of Change from Longitudinal User Text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 4647-4660).

Results (CLPsych Shared Task ‘22)

Results (CLPsych Shared Task ‘22)

Tsakalidis, A., Chim, J., Bilal, I. M., Zirikly, A., Atzil-Slonim, D., Nanni, F., ... & Liakata, M. (2022). Overview of the CLPsych 2022 Shared Task: Capturing Moments of Change in Longitudinal User Posts. *CLPsych 2022*, 184.

		DE	Post-level Evaluation												Coverage-based Metrics							
			macro-avg			**IS**			**IE**			**O**			**macro-avg**		**IS**		**IE**		**O**	
			P	R	F1	P	R	F1	P	R	F1	P	R	F1	C_p	C_r	C_p	C_r	C_p	C_r	C_p	C_r
Baseline	**Majority**		–	.333	.280	–	.000	.000	–	.000	.000	.724	1.000	.840	–	.142	–	.000	–	.000	.489	.426
	LR-tfidf		.545	.495	.492	.222	.024	.044	.569	.514	.540	.844	.948	.893	.378	.425	.111	.008	.284	.504	.738	.762
	BERT$_f$-TalkLife		.523	.386	.380	.091	.012	.022	.723	.163	.267	.754	.983	.853	.260	.204	.025	.007	.226	.094	.529	.513
System Submissions	**BLUE**		.505	.495	.499	.175	.171	.173	.484	.433	.457	.855	.882	.868	**.499**	.378	**.500**	.028	.299	.395	**.699**	.712
	IIITH		.520	**.600**	.519	.206	**.524**	.296	.402	**.630**	.491	**.954**	.647	.771	.347	.405	.254	**.356**	.249	.373	.536	.486
	LAMA		.552	.535	.524	.166	**.354**	.226	.609	.389	.475	.882	.861	.871	.376	.441	.253	**.373**	.193	.244	.680	.706
	NLP-UNED	✓	.493	.518	.501	.189	.293	.230	.414	.471	.440	.876	.791	.832	.306	.401	.244	.304	.134	.330	.541	.569
	UArizona	✓	.525	.507	.510	.142	.220	.172	.561	.423	.482	.872	.879	.876	.418	.416	.368	.248	.202	.285	.682	.716
	UoS		**.689**	**.625**	**.649**	**.490**	.305	**.376**	**.697**	**.630**	**.662**	.881	**.940**	**.909**	**.506**	**.503**	**.453**	.343	**.369**	**.450**	.695	**.717**
	uOttawa-AI		.505	.530	.512	.213	.244	.227	.402	.553	.466	**.899**	.793	.842	.348	.453	.272	.317	.176	.417	.595	.625
	WResearch	✓	**.625**	.579	**.598**	**.362**	.256	**.300**	**.646**	.553	**.596**	.868	**.929**	**.897**	.472	**.503**	.406	.318	**.307**	**.467**	**.703**	**.725**
	WWBP-SQT-lite		.508	.509	.508	.231	.220	.225	.440	.462	.451	.852	.845	.848	.336	.376	.270	.224	.186	.321	.551	.583

Tsakalidis, A., Chim, J., Bilal, I. M., Zirikly, A., Atzil-Slonim, D., Nanni, F., ... & Liakata, M. (2022). Overview of the CLPsych 2022 Shared Task: Capturing Moments of Change in Longitudinal User Posts. *CLPsych 2022*, 184.

Results (CLPsych Shared Task '22)

		DE	Post-level Evaluation												Coverage-based Metrics							
			macro-avg			IS			IE			O			macro-avg		IS		IE		O	
			P	R	F1	P	R	F1	P	R	F1	P	R	F1	C_p	C_r	C_p	C_r	C_p	C_r	C_p	C_r
Baseline	**Majority**		–	.333	.280	–	.000	.000	–	.000	.000	.724	1.000	.840	–	.142	–	.000	–	.000	.489	.426
	LR-tfidf		.545	.495	.492	.222	.024	.044	.569	.514	.540	.844	.948	.893	.378	.425	.111	.008	.284	.504	.738	.762
	BERT$_f$-TalkLife		.523	.386	.380	.091	.012	.022	.723	.163	.267	.754	.983	.853	.260	.204	.025	.007	.226	.094	.529	.513
System Submissions	**BLUE**		.505	.495	.499	.175	.171	.173	.484	.433	.457	.855	.882	.868	**.499**	.378	**.500**	.028	.299	.395	**.699**	.712
	IIITH		.520	**.600**	.519	.206	**.524**	.296	.402	**.630**	.491	**.954**	.647	.771	.347	.405	.254	**.356**	.249	.373	.536	.486
	LAMA		.552	.535	.524	.166	**.354**	.226	.609	.389	.475	.882	.861	.871	.376	.441	.253	**.373**	.193	.244	.680	.706
	NLP-UNED	✓	.493	.518	.501	.189	.293	.230	.414	.471	.440	.876	.791	.832	.306	.401	.244	.304	.134	.330	.541	.569
	UArizona	✓	.525	.507	.510	.142	.220	.172	.561	.423	.482	.872	.879	.876	.418	.416	.368	.248	.202	.285	.682	.716
	UoS		**.689**	**.625**	**.649**	**.490**	.305	**.376**	**.697**	**.630**	**.662**	.881	**.940**	**.909**	**.506**	**.503**	**.453**	.343	**.369**	**.450**	.695	**.717**
	uOttawa-AI		.505	.530	.512	.213	.244	.227	.402	.553	.466	**.899**	.793	.842	.348	.453	.272	.317	.176	.417	.595	.625
	WResearch	✓	**.625**	.579	**.598**	**.362**	.256	**.300**	**.646**	.553	**.596**	.868	**.929**	**.897**	.472	**.503**	.406	.318	**.307**	**.467**	**.703**	**.725**
	WWBP-SQT-lite		.508	.509	.508	.231	.220	.225	.440	.462	.451	.852	.845	.848	.336	.376	.270	.224	.186	.321	.551	.583

Both best-performing systems used a **longitudinal component** in their modelling
Many teams tackled the issue of **class imbalance**

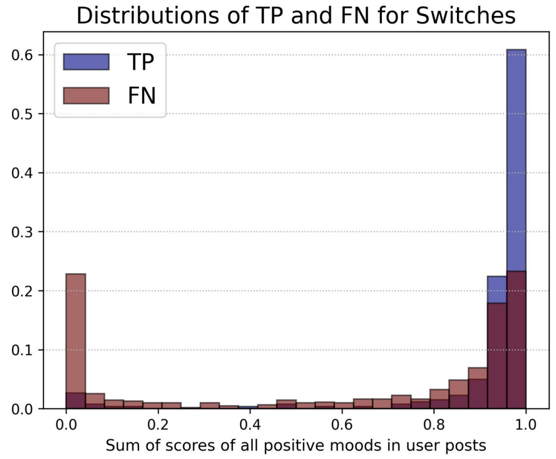
Additional task (user risk identification – not shown here): boost by using info from capturing MoC task

Error Analysis

Text	True	Pred.
Oh, forgot :) Stay safe you lovely people all around the world!	O	IS
Hope you are all having a good night! Stay safe! :D	O	IS
Don’t wanna deal with anyone.. Hope school finishes so I can go home soon	IS	O
Tired of my leg hurting so badly today. I really can’t do any training :(	IS	O
Hope you’re all great! <3 Love you all!	O	IS

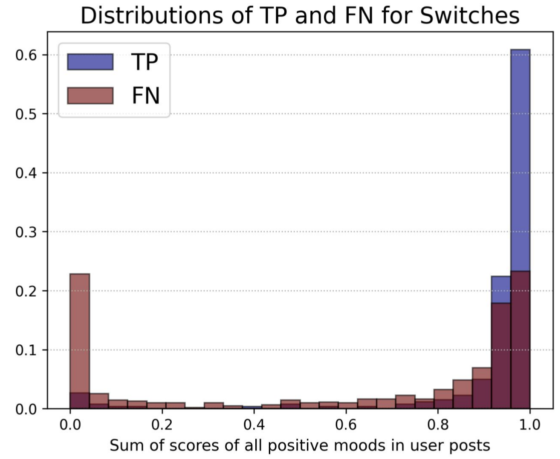
Error Analysis

Text	True	Pred.
Oh, forgot :) Stay safe you lovely people all around the world!	O	IS
Hope you are all having a good night! Stay safe! :D	O	IS
Don’t wanna deal with anyone.. Hope school finishes so I can go home soon	IS	O
Tired of my leg hurting so badly today. I really can’t do any training :(	IS	O
Hope you’re all great! <3 Love you all!	O	IS



Error Analysis

Text	True	Pred.
Oh, forgot :) Stay safe you lovely people all around the world!	O	IS
Hope you are all having a good night! Stay safe! :D	O	IS
Don't wanna deal with anyone.. Hope school finishes so I can go home soon	IS	O
Tired of my leg hurting so badly today. I really can't do any training :(	IS	O
Hope you're all great! <3 Love you all!	O	IS



Issue: Our models fail to learn what constitutes the "user's baseline" mood!

Proposed Solution: Personalisation by learning across-user similarities across-time

Lecture III: Identifying changes in longitudinal user generated content (II) (signature and transformer based methods)

Sequential Signature Networks for Longitudinal Monitoring

Tseriotou, T., Tsakalidis, A., Foster, P., Lyons, T. & Liakata, M. (2023). Sequential Signature Networks for Longitudinal Monitoring. ACL 2023 (Findings Volume)

Research Motivation

- PLMs are powerful in producing **static** word embeddings.
- **Limited work** on dynamic user representations.
- **Current research**:
  - applied only off-line.
  - lacks generalisability.

Research Motivation

- PLMs are powerful in producing **static** word embeddings.
- **Limited work** on dynamic user representations.
- **Current research**:
  - applied only off-line.
  - lacks generalisability.

Build efficient and compressed temporal user representations to address user-specific changes over time

Data

	TalkLife	Reddit
Timelines	500	256
Timeline length	≤ 2-week	~ 2-month
Avg posts per timeline	37	24

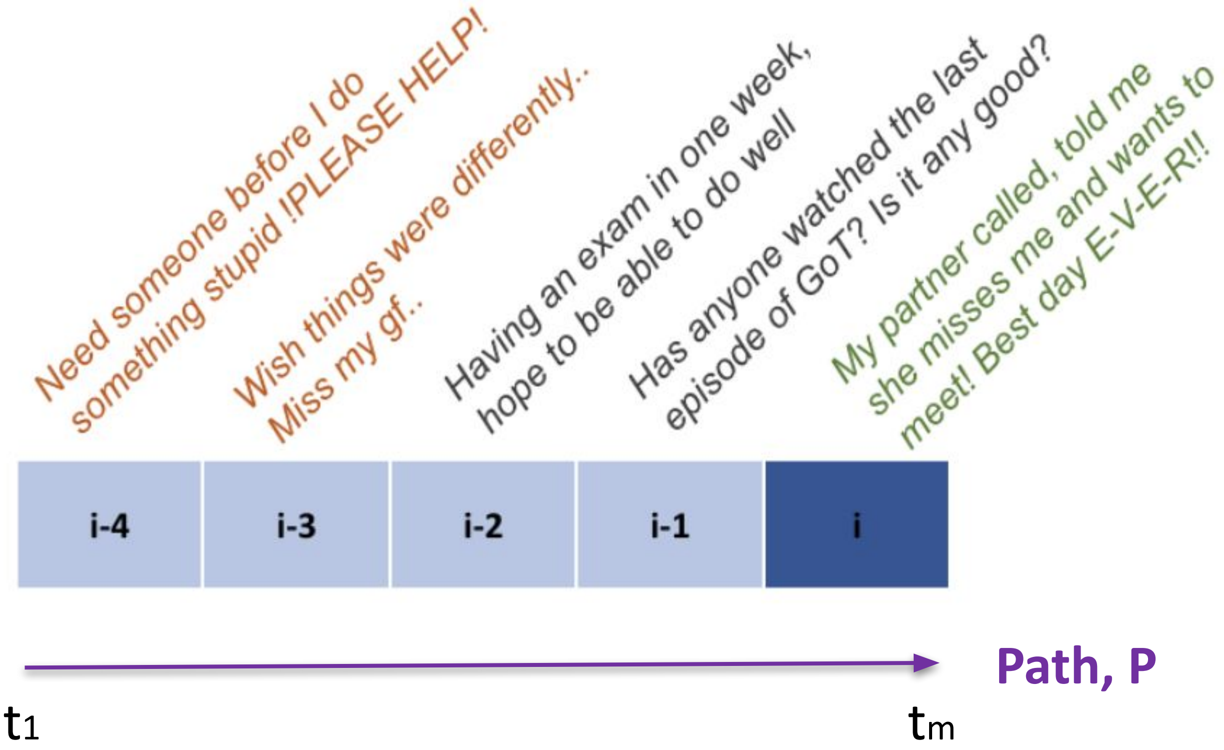
Path Signatures

What are Path Signatures?

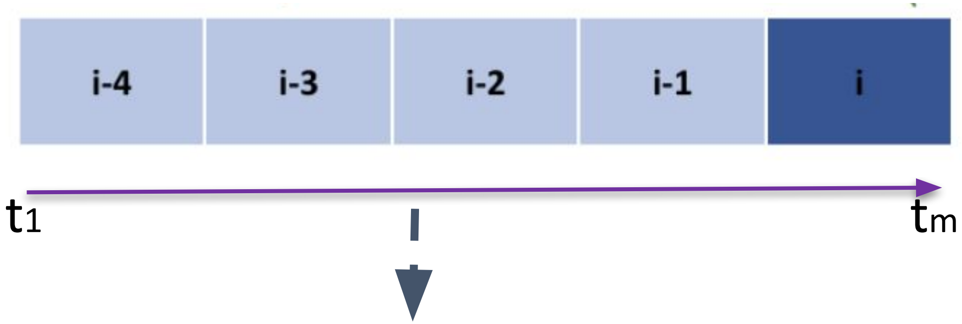
In theory: Path-wise definition to the solution of differential equation driven by irregular signals.

In practice: Produce a collection of statistics summarizing uniquely important information of the path.

Path Signatures - Explained



Path Signatures - Explained



Path, P

$$S(P)_{t_1,t_m} = (1, S(P)^{1}_{t_1,t_m}, \ldots, S(P)^{c}_{t_1,t_m},$$
$$S(P)^{1,1}_{t_1,t_m}, S(P)^{1,2}_{t_1,t_m}, \ldots, S(P)^{c,c}_{t_1,t_m}$$
$$\ldots, S(P)^{i_1,i_2,\cdots,i_r}_{t_1,t_m}, \ldots)$$

$$S(P)^{i_1,i_2,\cdots,i_r}_{t_1,t_m} = \int_{g_r} \ldots \int_{g_1} dP^{i_1}_{g_1} \otimes \ldots \otimes dP^{i_r}_{g_r}$$

Path Signature, S(P)

Path Signatures - Explained

$$S(P)_{t_1,t_m} = (1, \boxed{S(P)^1_{t_1,t_m}, ..., S(P)^c_{t_1,t_m}},$$

degree 1

$$S(P)^{1,1}_{t_1,t_m}, S(P)^{1,2}_{t_1,t_m}, ..., S(P)^{c,c}_{t_1,t_m}$$

$$..., S(P)^{i_1,i_2,\cdots,i_r}_{t_1,t_m}, ...)$$

Path Signature, S(P)

Path Signatures - Explained

$$S(P)_{t_1,t_m} = (1, S(P)^1_{t_1,t_m}, ..., S(P)^c_{t_1,t_m},$$
$$\boxed{S(P)^{1,1}_{t_1,t_m}, S(P)^{1,2}_{t_1,t_m}, ..., S(P)^{c,c}_{t_1,t_m}}$$ degree 2
$$..., S(P)^{i_1,i_2,\cdots,i_r}_{t_1,t_m}, ...)$$

Path Signature, S(P)

Path Signatures - Explained

$$S(P)_{t_1,t_m} = (1, S(P)^1_{t_1,t_m}, \ldots, S(P)^c_{t_1,t_m},$$
$$S(P)^{1,1}_{t_1,t_m}, S(P)^{1,2}_{t_1,t_m}, \ldots, S(P)^{c,c}_{t_1,t_m}$$
$$\ldots, S(P)^{i_1,i_2,\cdots,i_r}_{t_1,t_m}, \ldots)$$

Path Signature, S(P)

Number of output dimensions used as features:

$$(c^{N+1} - c)(c-1)^{-1}$$

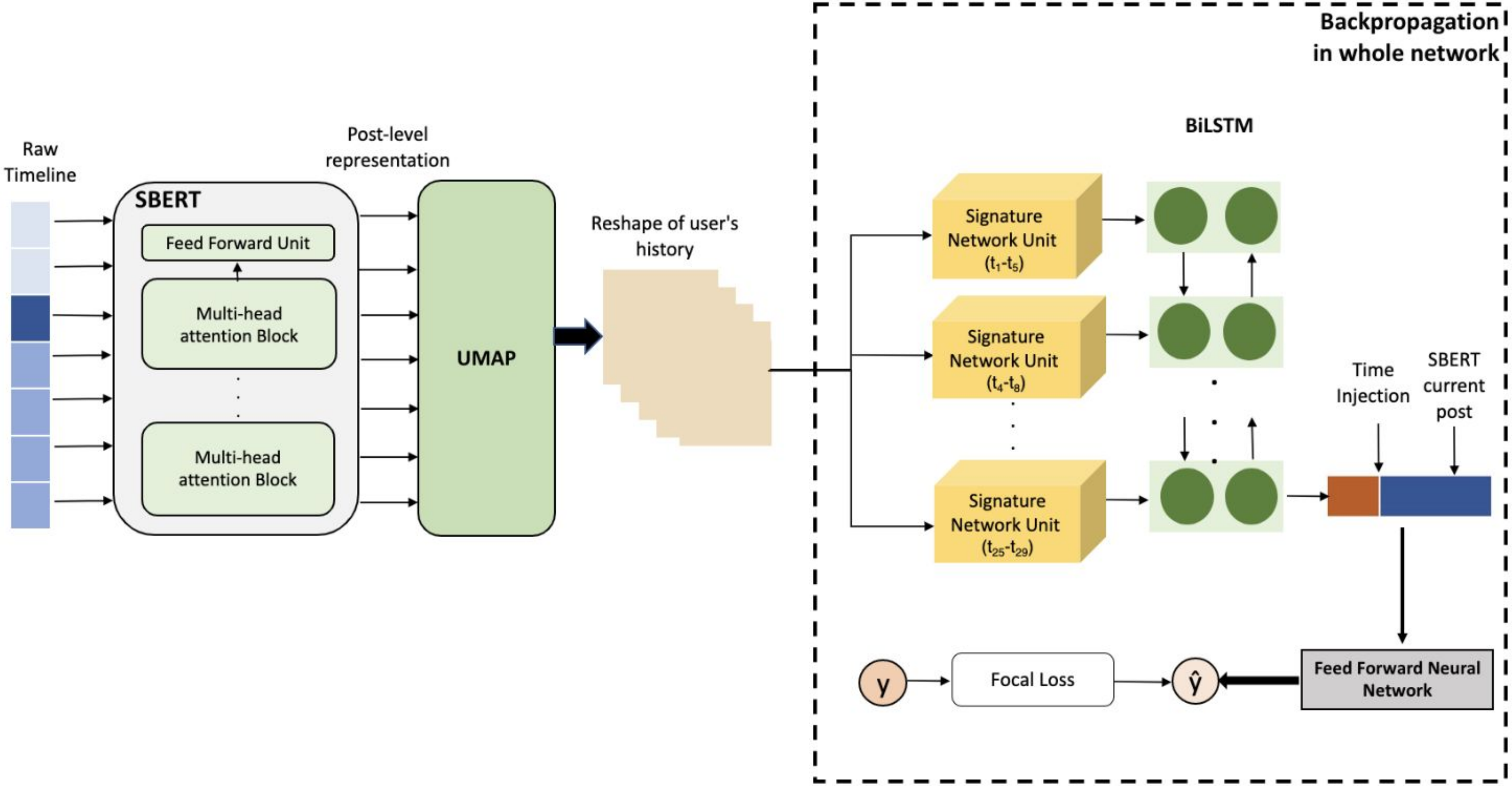
Dimensionality Reduction

Path Signatures

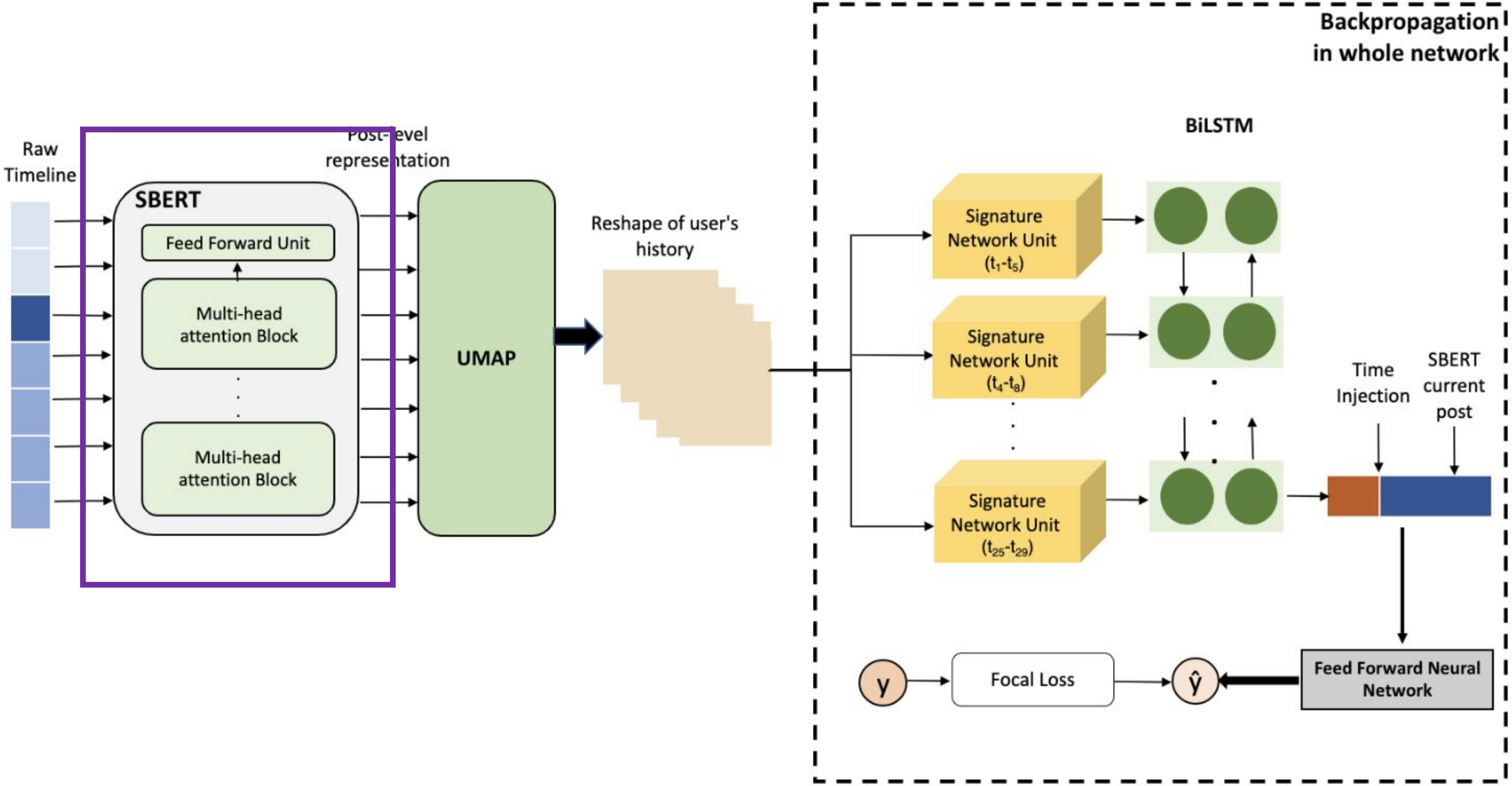
Advantages:

- Efficient and compressed encoding of **sequential data**.
- Sequential **pooling** operator in Neural Models.
- Enhances **short-term dependencies** in linguistic timelines.
- Account for **time irregularities**.

Model

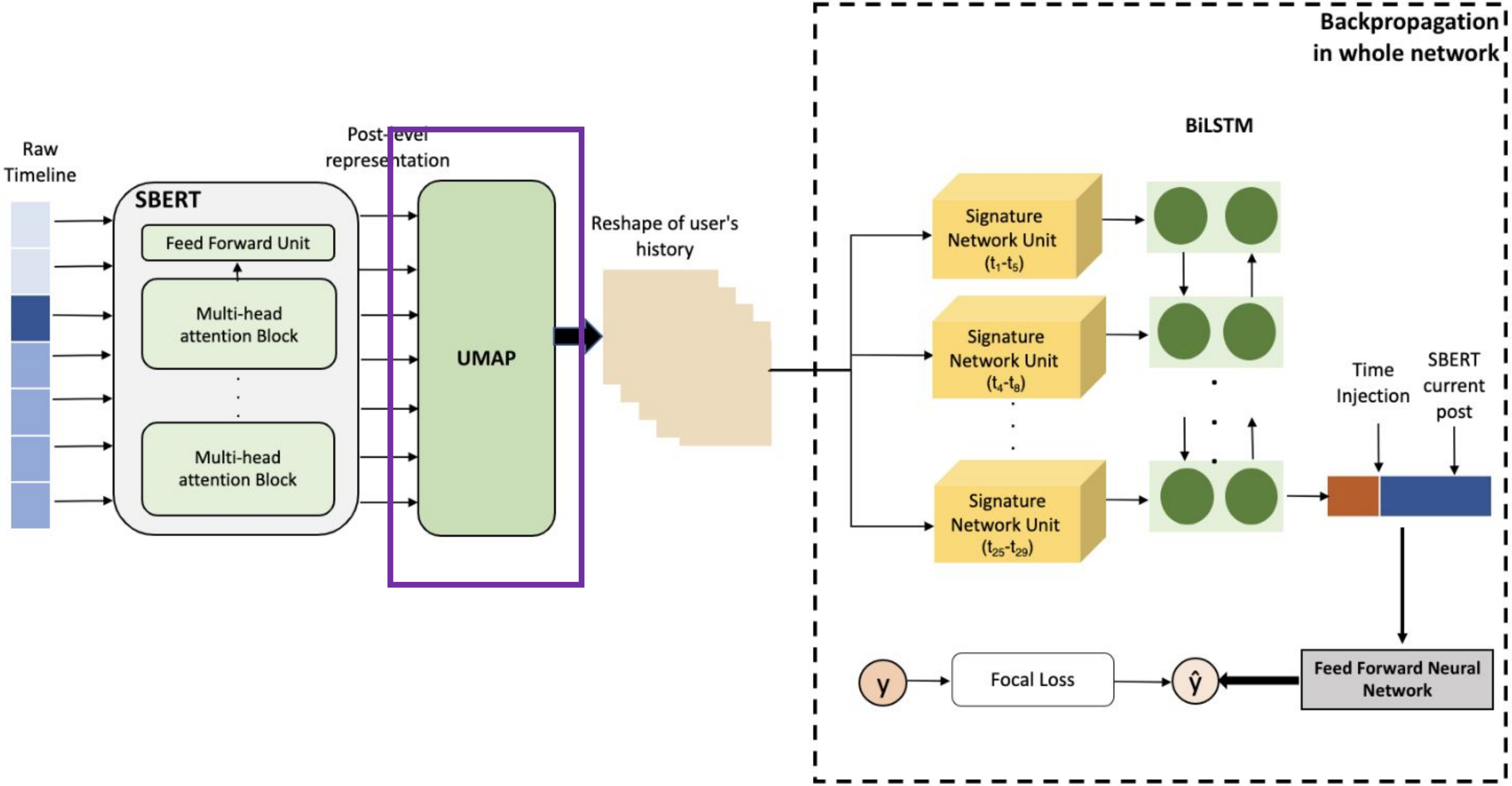

Backpropagation in whole network
Raw Timeline
SBERT
Feed Forward Unit
Multi-head attention Block
Multi-head attention Block
Post-level representation
UMAP
Reshape of user's history
BiLSTM
Signature Network Unit (t_1-t_5)
Signature Network Unit (t_4-t_8)
Signature Network Unit (t_{25}-t_{29})
Time Injection
SBERT current post
Feed Forward Neural Network
y
Focal Loss
ŷ

Model



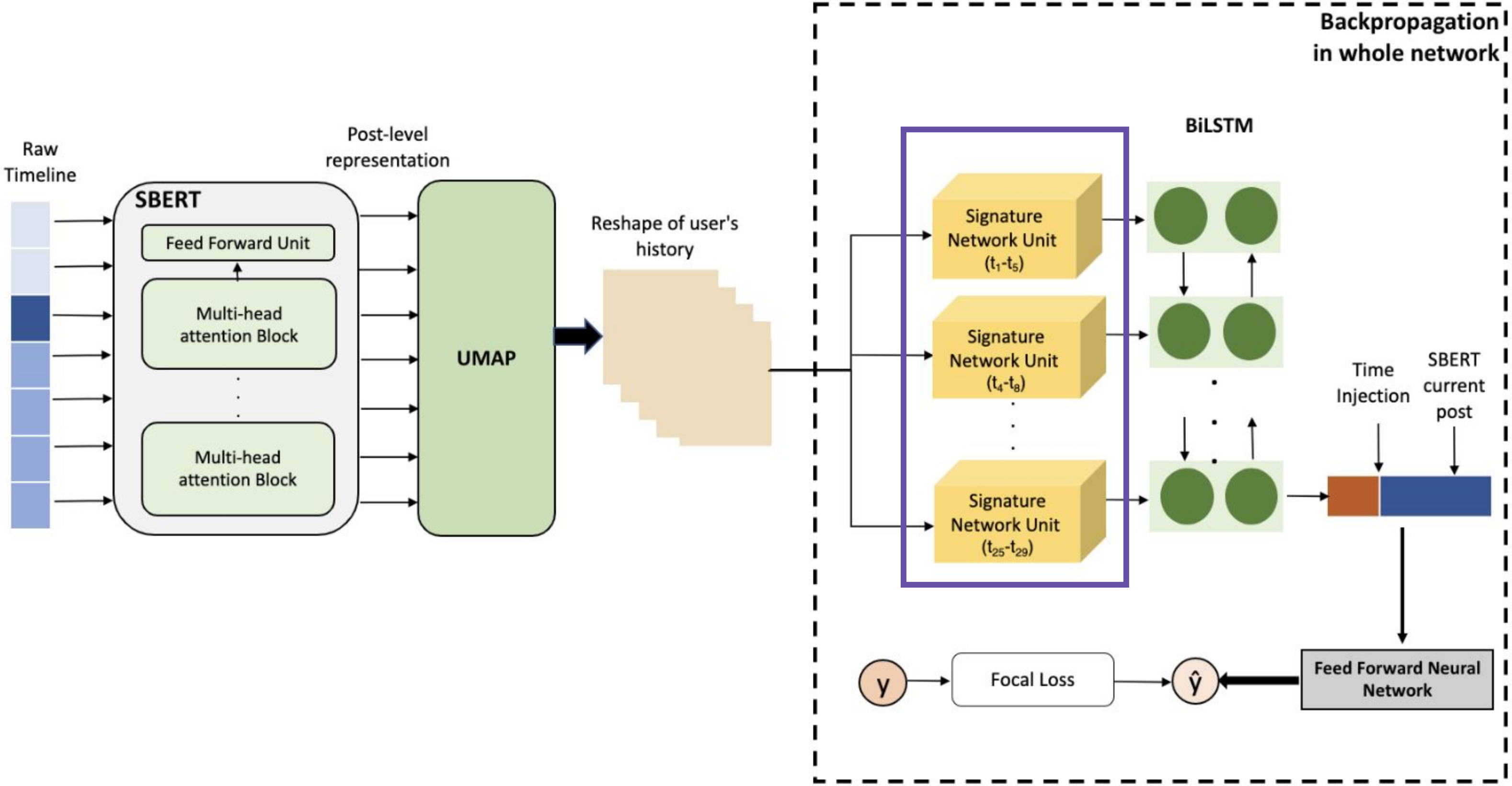
Raw Timeline
SBERT
Feed Forward Unit
Multi-head attention Block
Multi-head attention Block
Post-level representation
UMAP
Reshape of user's history
Backpropagation in whole network
BiLSTM
Signature Network Unit (t_1-t_5)
Signature Network Unit (t_4-t_8)
Signature Network Unit (t_{25}-t_{29})
Time Injection
SBERT current post
Feed Forward Neural Network
y
Focal Loss
ŷ

Model



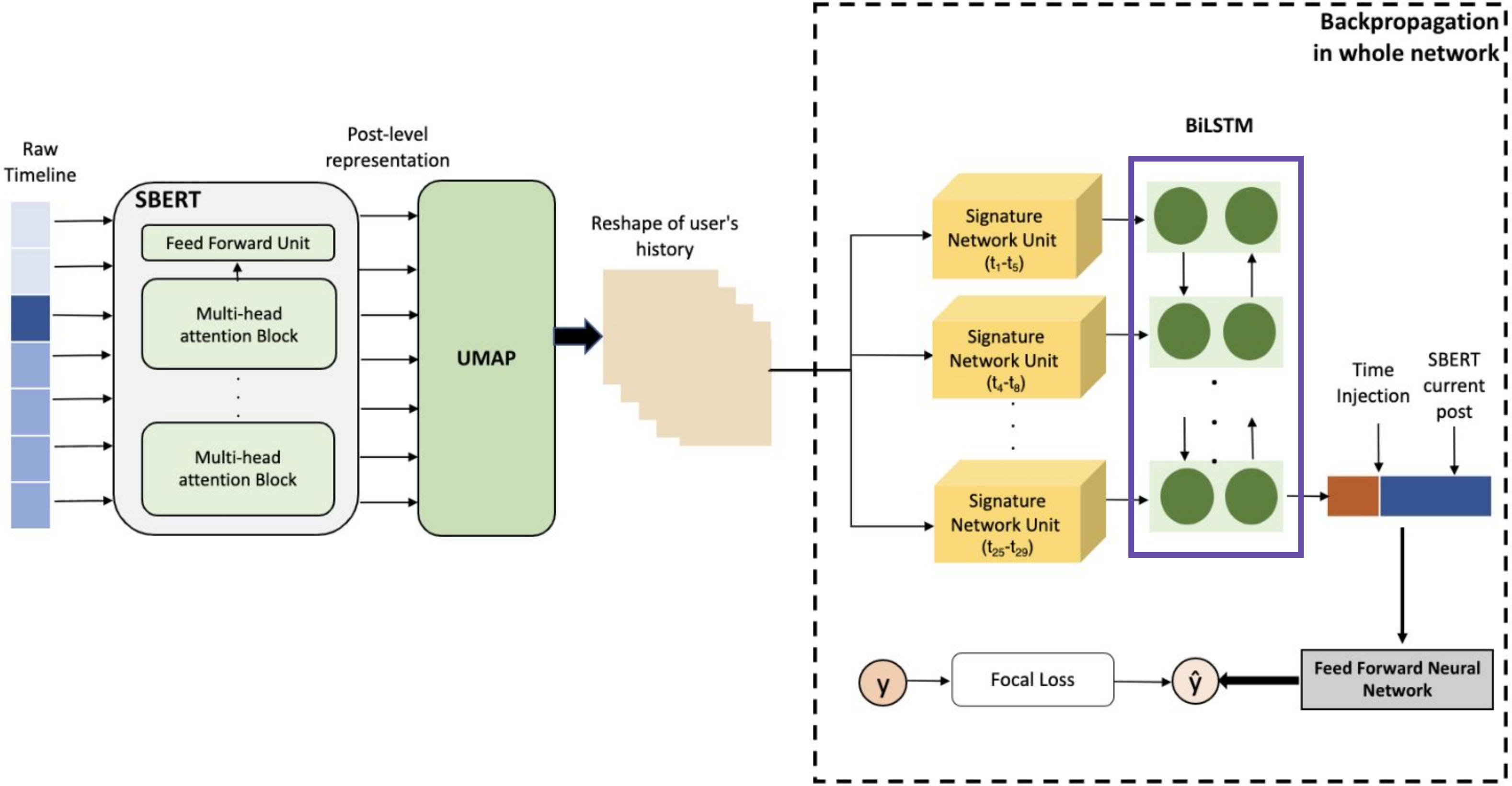
Raw Timeline
SBERT
Feed Forward Unit
Multi-head attention Block
Multi-head attention Block
Post-level representation
UMAP
Reshape of user's history
Backpropagation in whole network
BiLSTM
Signature Network Unit (t_1-t_5)
Signature Network Unit (t_4-t_8)
Signature Network Unit (t_{25}-t_{29})
Time Injection
SBERT current post
Feed Forward Neural Network
y
Focal Loss
ŷ

Model



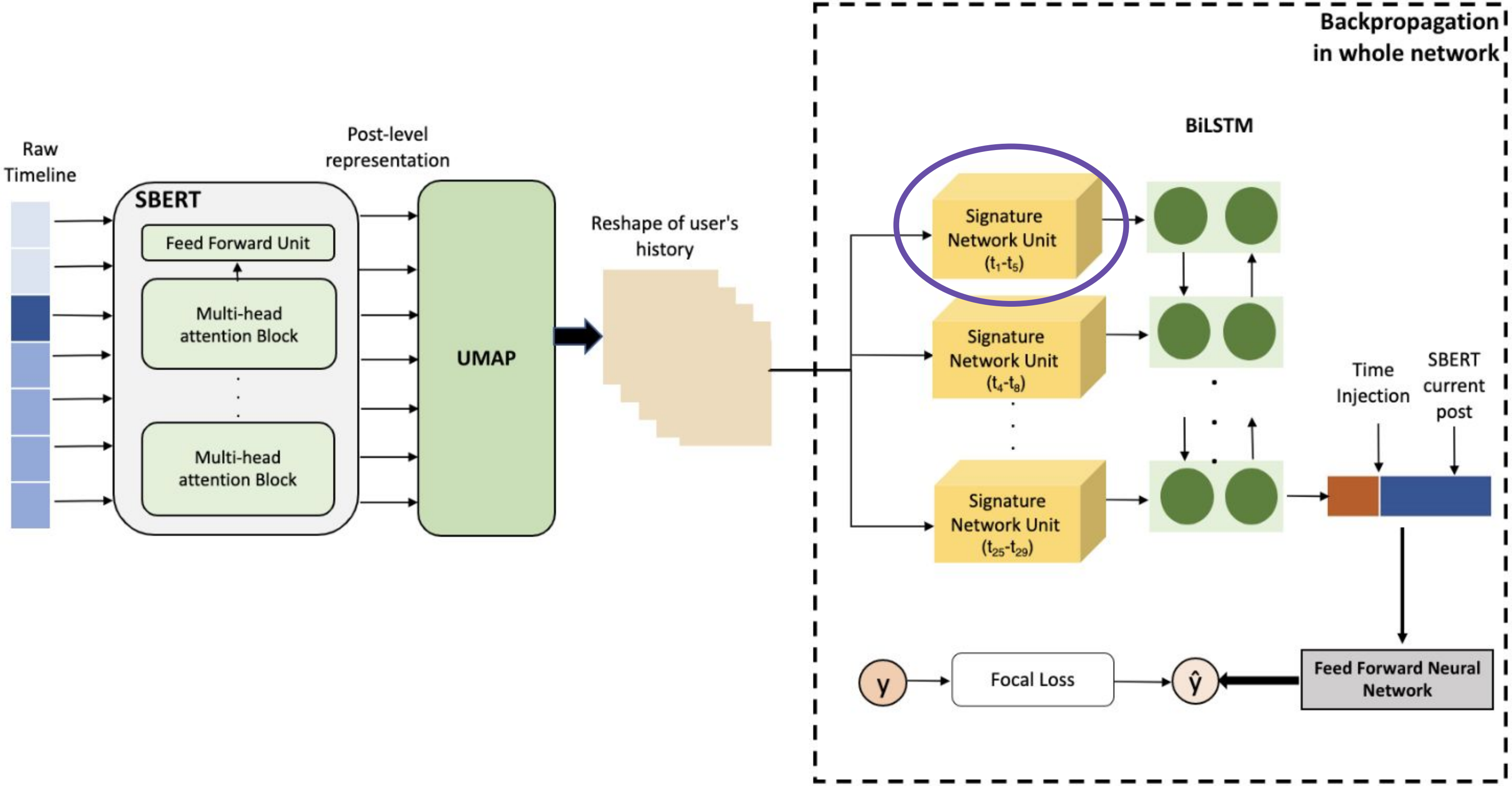
Raw Timeline
SBERT
Feed Forward Unit
Multi-head attention Block
Multi-head attention Block
Post-level representation
UMAP
Reshape of user's history
Backpropagation in whole network
Signature Network Unit (t_1-t_5)
Signature Network Unit (t_4-t_8)
Signature Network Unit (t_{25}-t_{29})
BiLSTM
Time Injection
SBERT current post
Feed Forward Neural Network
y
Focal Loss
ŷ

Model



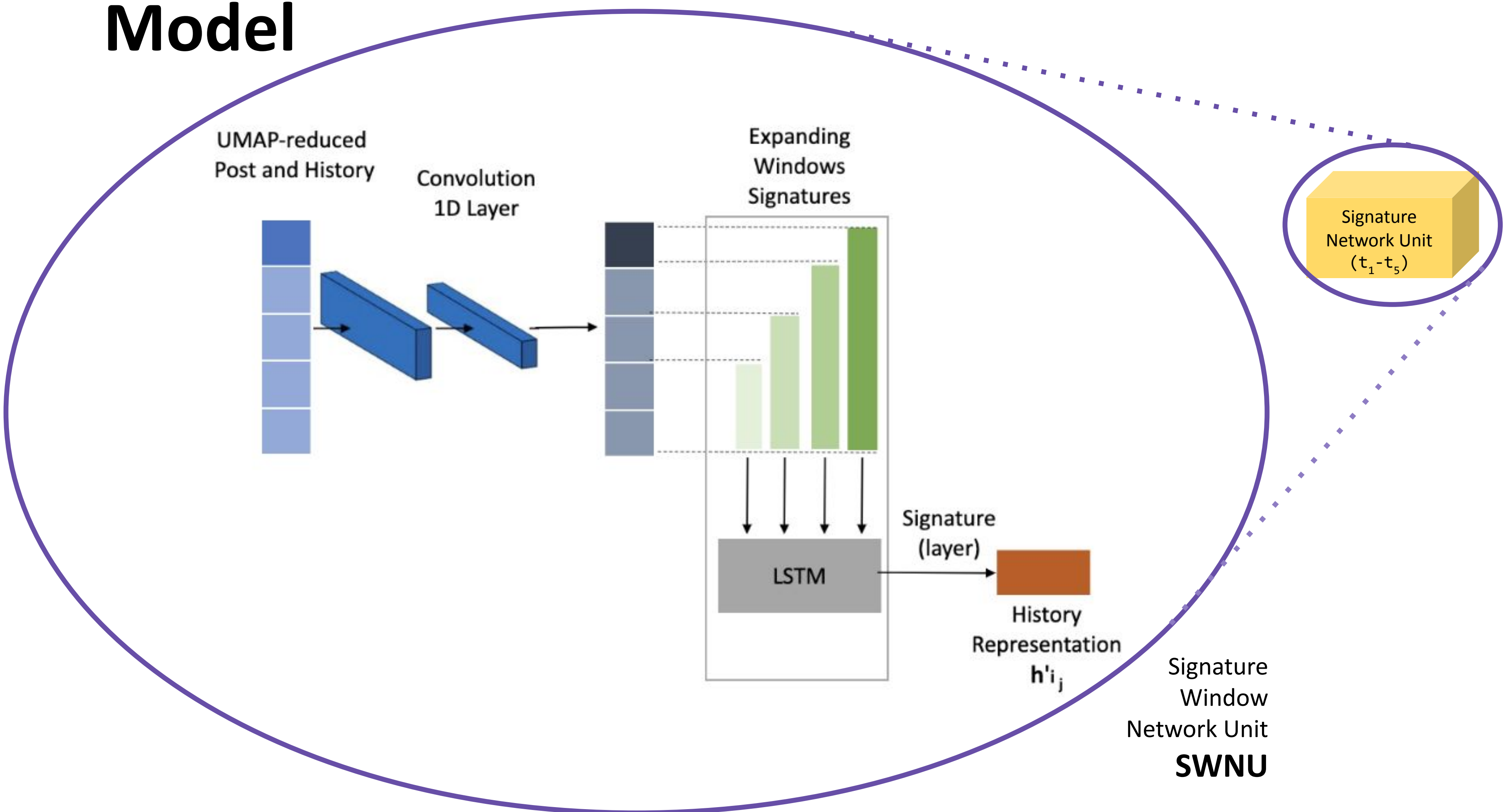
Backpropagation in whole network
Raw Timeline
Post-level representation
SBERT
Feed Forward Unit
Multi-head attention Block
Multi-head attention Block
UMAP
Reshape of user's history
BiLSTM
Signature Network Unit (t_1-t_5)
Signature Network Unit (t_4-t_8)
Signature Network Unit (t_{25}-t_{29})
Time Injection
SBERT current post
Feed Forward Neural Network
y
Focal Loss
ŷ

Model



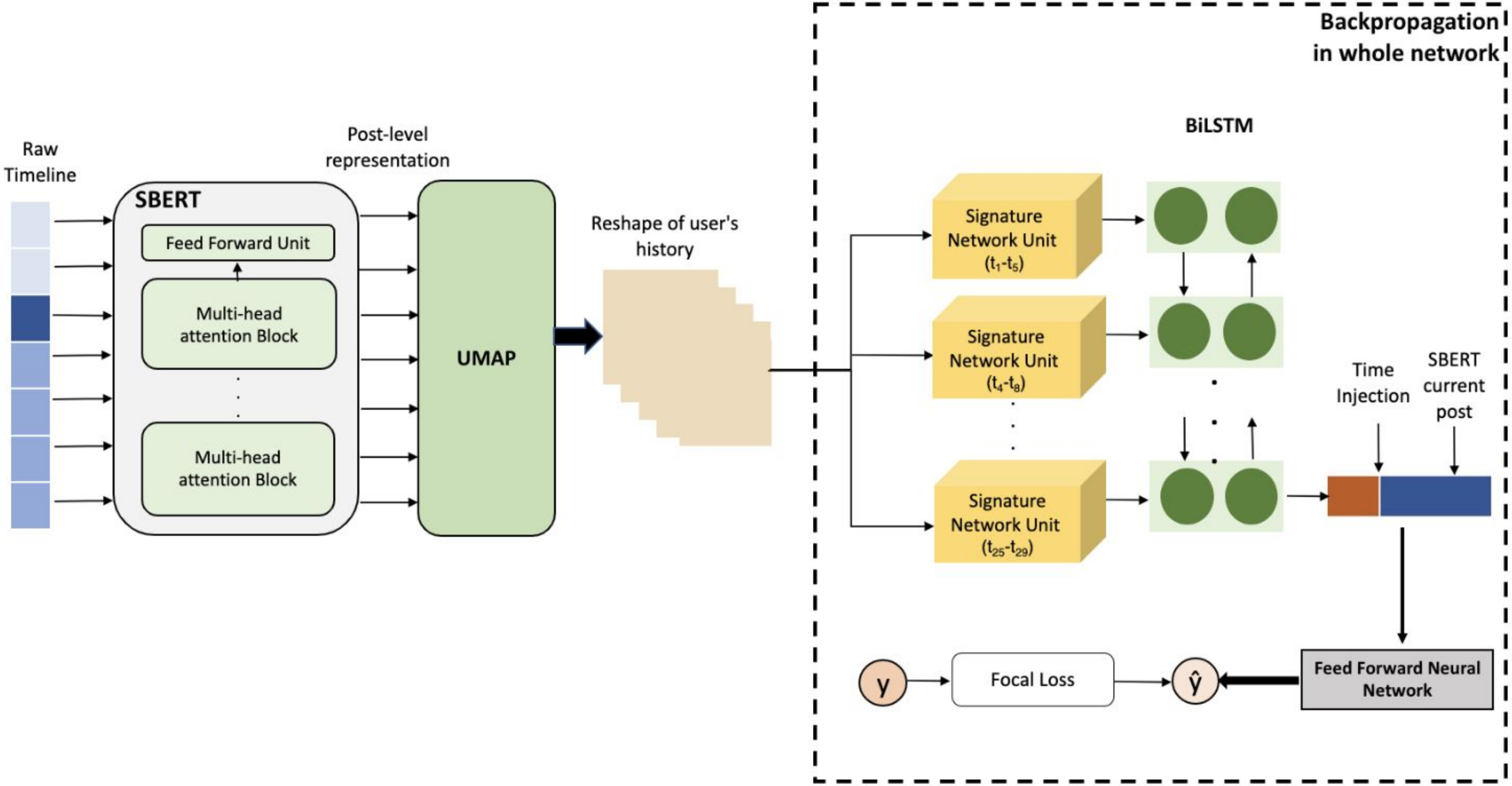
Backpropagation in whole network
Raw Timeline
SBERT
Feed Forward Unit
Multi-head attention Block
Multi-head attention Block
Post-level representation
UMAP
Reshape of user's history
BiLSTM
Signature Network Unit (t_1-t_5)
Signature Network Unit (t_4-t_8)
Signature Network Unit (t_{25}-t_{29})
Time Injection
SBERT current post
Feed Forward Neural Network
y
Focal Loss
ŷ

Model



UMAP-reduced Post and History
Convolution 1D Layer
Expanding Windows Signatures
LSTM
Signature (layer)
History Representation h'i_j
Signature Network Unit (t_1-t_5)
Signature Window Network Unit
SWNU

Model



Backpropagation in whole network
Raw Timeline
SBERT
Feed Forward Unit
Multi-head attention Block
Multi-head attention Block
Post-level representation
UMAP
Reshape of user's history
BiLSTM
Signature Network Unit (t_1-t_5)
Signature Network Unit (t_4-t_8)
Signature Network Unit (t_{25}-t_{29})
Time Injection
SBERT current post
Feed Forward Neural Network
y
Focal Loss
ŷ

Results

ACL 2023 (findings) Sequential Path Signature Networks for Personalised Longitudinal Language Modeling

			IS			IE			O			macro-avg			Model Type	
			P	R	F1	P	R	F1	P	R	F1	P	R	F1	Emotion	Future
TalkLife	Naïve	Majority	–	–	–	–	–	–	.845	1	.916	.282	.333	.305		
		Random	.047	.047	.047	.108	.108	.108	.845	.845	.845	.333	.333	.333		
	Post-level	BERT(f) (Tsakalidis et al., 2022b)	.260	**.321**	.287	.401	.478	.436	898	.864	.881	.520	.554	.534		
	Timeline-level	EM-DM	**.553**	.118	.193	.479	.351	.405	.880	**.948**	.913	**.631**	.472	.504	✓	✓
	(Tsakalidis et al., 2022b)	BiLSTM-bert	.397	.264	**.316**	**.568**	.461	**.508**	.898	.936	**.917**	.621	.553	**.580**		✓
	Timeline-level	SBERT(avg hist)	.283	.244	.262	.424	.486	.452	.896	.885	.890	.534	.539	.535		
	(-signature)	BiLSTM-sbert(hist)	.258	.272	.264	.442	.506	.468	.901	.879	.890	.534	.553	.541		
		BiLSTM-bert(hist)	.405	.241	.302	.536	.415	.468	.892	.938	.914	.611	.531	.561		
	Timeline-level (+signature)	Seq-Sig-Net (our work)	.331	.290	.309	.435	**.555**	.487	**.907**	.881	.894	.558	**.576**	.563		
Reddit	Naïve	Majority	–	.000	.000	–	.000	.000	.724	1.000	.840	–	.333	.280		
		Random	.066	.066	.066	.158	.158	.158	.776	.776	.776	.333	.333	.333		
		IIITH (Boinepelli et al., 2022)	.206	**.524**	.296	.402	.630	.491	**.954**	.647	.771	.520	.600	.519		
	Timeline-level	LAMA (AlHamed et al., 2022)	.166	.354	.226	.609	.389	.475	.882	.861	.871	.552	.535	.524		
	(CLPsych)	WResearch (Bayram and Benhiba, 2022)	.362	.256	.300	.646	.553	.596	.868	.929	.897	.625	.579	.598	✓	
		UoS (Azim et al., 2022)	**.490**	.305	.376	**.697**	.630	**.662**	.881	**.940**	**.909**	**.689**	.625	.649	✓	✓
	Timeline-level	SBERT(avg hist)	.340	.329	.330	.605	.563	.582	.893	.912	.902	.613	.601	.605		
	(-signature)	BiLSTM-sbert(hist)	.463	.407	**.430**	.629	**.637**	.630	.895	.901	.898	.663	**.648**	**.653**		
	Timeline-level (+signature)	Seq-Sig-Net (our work)	.454	.405	.425	.643	.607	.624	.896	.919	.908	.664	.644	.652		

Results

real-time application

			IS			IE			O			macro-avg			Model Type	
			P	R	F1	P	R	F1	P	R	F1	P	R	F1	Emotion	Future
TalkLife	Naïve	Majority	–	–	–	–	–	–	.845	1	.916	.282	.333	.305		
		Random	.047	.047	.047	.108	.108	.108	.845	.845	.845	.333	.333	.333		
	Post-level	BERT(f) (Tsakalidis et al., 2022b)	.260	**.321**	.287	.401	.478	.436	898	.864	.881	.520	.554	.534		
	Timeline-level	EM-DM	**.553**	.118	.193	.479	.351	.405	.880	**.948**	.913	**.631**	.472	.504	✓	✓
	(Tsakalidis et al., 2022b)	BiLSTM-bert	.397	.264	**.316**	**.568**	.461	**.508**	.898	.936	**.917**	.621	.553	**.580**		✓
	Timeline-level	SBERT(avg hist)	.283	.244	.262	.424	.486	.452	.896	.885	.890	.534	.539	.535		
	(-signature)	BiLSTM-sbert(hist)	.258	.272	.264	.442	.506	.468	.901	.879	.890	.534	.553	.541		
		BiLSTM-bert(hist)	.405	.241	.302	.536	.415	.468	.892	.938	.914	.611	.531	.561		
	Timeline-level (+signature)	Seq-Sig-Net (our work)	.331	.290	.309	.435	**.555**	.487	**.907**	.881	.894	.558	**.576**	.563		
Reddit	Naïve	Majority	–	.000	.000	–	.000	.000	.724	1.000	.840	–	.333	.280		
		Random	.066	.066	.066	.158	.158	.158	.776	.776	.776	.333	.333	.333		
		IIITH (Boinepelli et al., 2022)	.206	**.524**	.296	.402	.630	.491	**.954**	.647	.771	.520	.600	.519		
	Timeline-level	LAMA (AlHamed et al., 2022)	.166	.354	.226	.609	.389	.475	.882	.861	.871	.552	.535	.524		
	(CLPsych)	WResearch (Bayram and Benhiba, 2022)	.362	.256	.300	.646	.553	.596	.868	.929	.897	.625	.579	.598	✓	
		UoS (Azim et al., 2022)	**.490**	.305	.376	**.697**	.630	**.662**	.881	**.940**	**.909**	**.689**	.625	.649	✓	✓
	Timeline-level	SBERT(avg hist)	.340	.329	.330	.605	.563	.582	.893	.912	.902	.613	.601	.605		
	(-signature)	BiLSTM-sbert(hist)	.463	.407	**.430**	.629	**.637**	.630	.895	.901	.898	.663	**.648**	**.653**		
	Timeline-level (+signature)	Seq-Sig-Net (our work)	.454	.405	.425	.643	.607	.624	.896	.919	.908	.664	.644	.652		

Results

real-time application

Dataset			IS			IE			O			macro-avg			Model Type	
			P	R	F1	P	R	F1	P	R	F1	P	R	F1	Emotion	Future
TalkLife	Naïve	Majority	–	–	–	–	–	–	.845	1	.916	.282	.333	.305		
		Random	.047	.047	.047	.108	.108	.108	.845	.845	.845	.333	.333	.333		
	Post-level	BERT(f) (Tsakalidis et al., 2022b)	.260	**.321**	.287	.401	.478	.436	898	.864	.881	.520	.554	.534		
	Timeline-level (Tsakalidis et al., 2022b)	EM-DM	**.553**	.118	.193	.479	.351	.405	.880	**.948**	.913	**.631**	.472	.504	✓	✓
		BiLSTM-bert	.397	.264	**.316**	**.568**	.461	**.508**	.898	.936	**.917**	.621	.553	**.580**		✓
	Timeline-level (-signature)	SBERT(avg hist)	.283	.244	.262	.424	.486	.452	.896	.885	.890	.534	.539	.535		
		BiLSTM-sbert(hist)	.258	.272	.264	.442	.506	.468	.901	.879	.890	.534	.553	.541		
		BiLSTM-bert(hist)	.405	.241	.302	.536	.415	.468	.892	.938	.914	.611	.531	.561		
	Timeline-level (+signature)	Seq-Sig-Net (our work)	.331	.290	.309	.435	**.555**	.487	**.907**	.881	.894	.558	**.576**	.563		
Reddit	Naïve	Majority	–	.000	.000	–	.000	.000	.724	1.000	.840	–	.333	.280		
		Random	.066	.066	.066	.158	.158	.158	.776	.776	.776	.333	.333	.333		
	Timeline-level (CLPsych)	IIITH (Boinepelli et al., 2022)	.206	**.524**	.296	.402	.630	.491	**.954**	.647	.771	.520	.600	.519		
		LAMA (AlHamed et al., 2022)	.166	.354	.226	.609	.389	.475	.882	.861	.871	.552	.535	.524		
		WResearch (Bayram and Benhiba, 2022)	.362	.256	.300	.646	.553	.596	.868	.929	.897	.625	.579	.598	✓	
		UoS (Azim et al., 2022)	**.490**	.305	.376	**.697**	.630	**.662**	.881	**.940**	**.909**	**.689**	.625	.649	✓	✓
	Timeline-level (-signature)	SBERT(avg hist)	.340	.329	.330	.605	.563	.582	.893	.912	.902	.613	.601	.605		
		BiLSTM-sbert(hist)	.463	.407	**.430**	.629	**.637**	.630	.895	.901	.898	.663	**.648**	**.653**		
	Timeline-level (+signature)	Seq-Sig-Net (our work)	.454	.405	.425	.643	.607	.624	.896	.919	.908	.664	.644	.652		

Results

real-time application

generalisable

Dataset		Model	IS P	IS R	IS F1	IE P	IE R	IE F1	O P	O R	O F1	macro-avg P	macro-avg R	macro-avg F1	Model Type: Emotion	Model Type: Future
TalkLife	Naïve	Majority	–	–	–	–	–	–	.845	1	.916	.282	.333	.305		
TalkLife		Random	.047	.047	.047	.108	.108	.108	.845	.845	.845	.333	.333	.333		
TalkLife	Post-level	BERT(f) (Tsakalidis et al., 2022b)	.260	**.321**	.287	.401	.478	.436	898	.864	.881	.520	.554	.534		
TalkLife	Timeline-level	EM-DM	**.553**	.118	.193	.479	.351	.405	.880	**.948**	.913	**.631**	.472	.504	✓	✓
TalkLife	(Tsakalidis et al., 2022b)	BiLSTM-bert	.397	.264	**.316**	**.568**	.461	**.508**	.898	.936	**.917**	.621	.553	**.580**		✓
TalkLife	Timeline-level	SBERT(avg hist)	.283	.244	.262	.424	.486	.452	.896	.885	.890	.534	.539	.535		
TalkLife	(-signature)	BiLSTM-sbert(hist)	.258	.272	.264	.442	.506	.468	.901	.879	.890	.534	.553	.541		
TalkLife		BiLSTM-bert(hist)	.405	.241	.302	.536	.415	.468	.892	.938	.914	.611	.531	.561		
TalkLife	Timeline-level (+signature)	Seq-Sig-Net (our work)	.331	.290	.309	.435	**.555**	.487	**.907**	.881	.894	.558	**.576**	.563		
Reddit	Naïve	Majority	–	.000	.000	–	.000	.000	.724	1.000	.840	–	.333	.280		
Reddit		Random	.066	.066	.066	.158	.158	.158	.776	.776	.776	.333	.333	.333		
Reddit	Timeline-level	IIITH (Boinepelli et al., 2022)	.206	**.524**	.296	.402	.630	.491	**.954**	.647	.771	.520	.600	.519		
Reddit	(CLPsych)	LAMA (AlHamed et al., 2022)	.166	.354	.226	.609	.389	.475	.882	.861	.871	.552	.535	.524		
Reddit		WResearch (Bayram and Benhiba, 2022)	.362	.256	.300	.646	.553	.596	.868	.929	.897	.625	.579	.598	✓	
Reddit		UoS (Azim et al., 2022)	**.490**	.305	.376	**.697**	.630	**.662**	.881	**.940**	**.909**	**.689**	.625	.649	✓	✓
Reddit	Timeline-level	SBERT(avg hist)	.340	.329	.330	.605	.563	.582	.893	.912	.902	.613	.601	.605		
Reddit	(-signature)	BiLSTM-sbert(hist)	.463	.407	**.430**	.629	**.637**	.630	.895	.901	.898	.663	**.648**	**.653**		
Reddit	Timeline-level (+signature)	Seq-Sig-Net (our work)	.454	.405	.425	.643	.607	.624	.896	.919	.908	.664	.644	.652		

Computational Resources

Seq-Sig-Net is much greener

Model name	Memory (MB)	Parameters (million)	Avg Training time (minutes)
BiLSTM-bert(hist)	18.9	2.5	36.7
Seq-Sig-Net	12.9	1.7	33.9

Ablation Study

		TalkLife				Reddit			
Model name	Explanation of ablation	IS	IE	O	avg	IS	IE	O	avg
SBERT post	(*)	.281	.431	.887	.533	.200	.541	.909	.550
SBERT(avg hist)	(*) + mean hist. + t	.262	.452	.890	.535	.330	.582	.902	.605
SWNU Network	(*) + 1 SWNU + t	.296	.477	**.894**	.556	.308	.623	**.911**	.614
Seq-Sig-Net	(*) + BiLSTM on SWNU + t	**.309**	**.487**	**.894**	**.563**	**.425**	**.624**	.908	**.652**

Ablation Study

efficient time windows

		TalkLife				Reddit			
Model name	Explanation of ablation	IS	IE	O	avg	IS	IE	O	avg
SBERT post	(*)	.281	.431	.887	.533	.200	.541	.909	.550
SBERT(avg hist)	(*) + mean hist. + t	.262	.452	.890	.535	.330	.582	.902	.605
SWNU Network	(*) + 1 SWNU + t	.296	.477	**.894**	.556	.308	.623	**.911**	.614
Seq-Sig-Net	(*) + BiLSTM on SWNU + t	**.309**	**.487**	**.894**	**.563**	**.425**	**.624**	.908	**.652**

Ablation Study

efficient time windows

memorising local parts of timeline

		TalkLife				Reddit			
Model name	Explanation of ablation	IS	IE	O	avg	IS	IE	O	avg
SBERT post	(*)	.281	.431	.887	.533	.200	.541	.909	.550
SBERT(avg hist)	(*) + mean hist. + t	.262	.452	.890	.535	.330	.582	.902	.605
SWNU Network	(*) + 1 SWNU + t	.296	.477	**.894**	.556	.308	.623	**.911**	.614
Seq-Sig-Net	(*) + BiLSTM on SWNU + t	**.309**	**.487**	**.894**	**.563**	**.425**	**.624**	.908	**.652**

Analysis

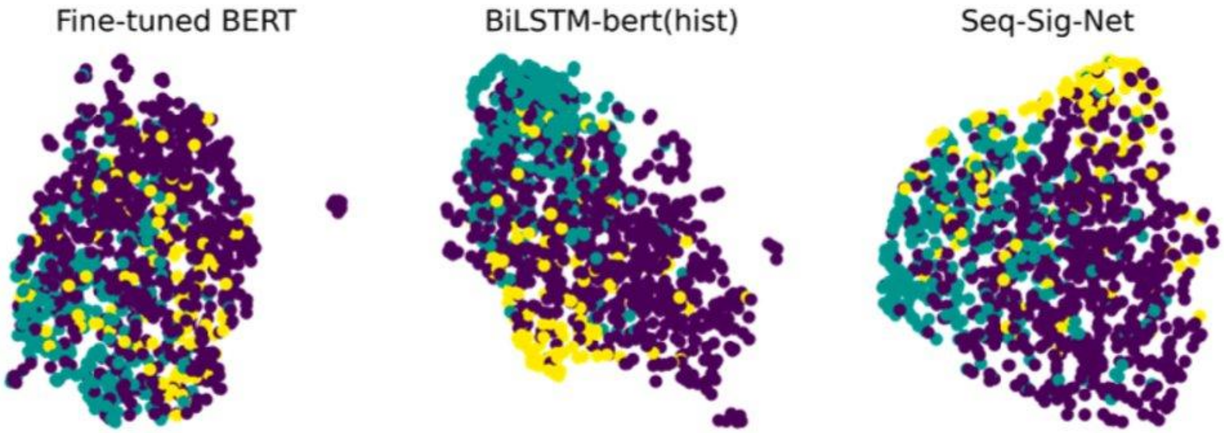
Clustering ability of representations?

Analysis

Clustering ability of representations?

Sequentiality - better

	Silhouette ($\sim$1)	Calinski Harabasz $\uparrow$	Davies Bouldin $\downarrow$
BERT fine-tuned	-0.091	134.01	3.15
BiLSTM-bert(hist)	-0.050	275.51	2.59
Seq-Sig-Net	**-0.014**	**294.66**	**2.45**



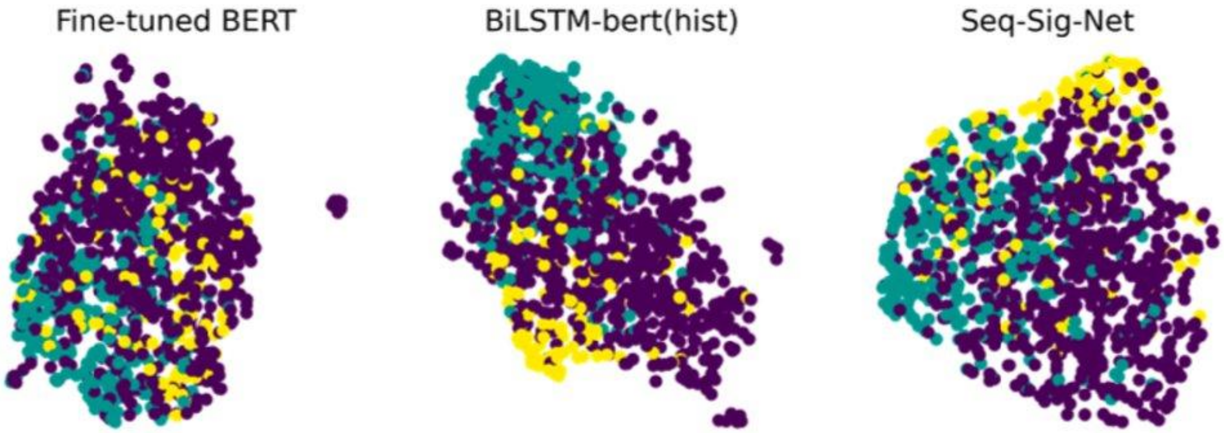
Analysis

Clustering ability of representations?

	Silhouette (~1)	Calinski Harabasz ↑	Davies Bouldin ↓
BERT fine-tuned	-0.091	134.01	3.15
BiLSTM-bert(hist)	-0.050	275.51	2.59
Seq-Sig-Net	**-0.014**	**294.66**	**2.45**

Sequentiality - better

Signatures - even better



Contributions

Signature Transforms in Neural Networks for Language Modeling.

Generalisable to sequential **real-time** applications.

SOTA performance against historical user data baselines.

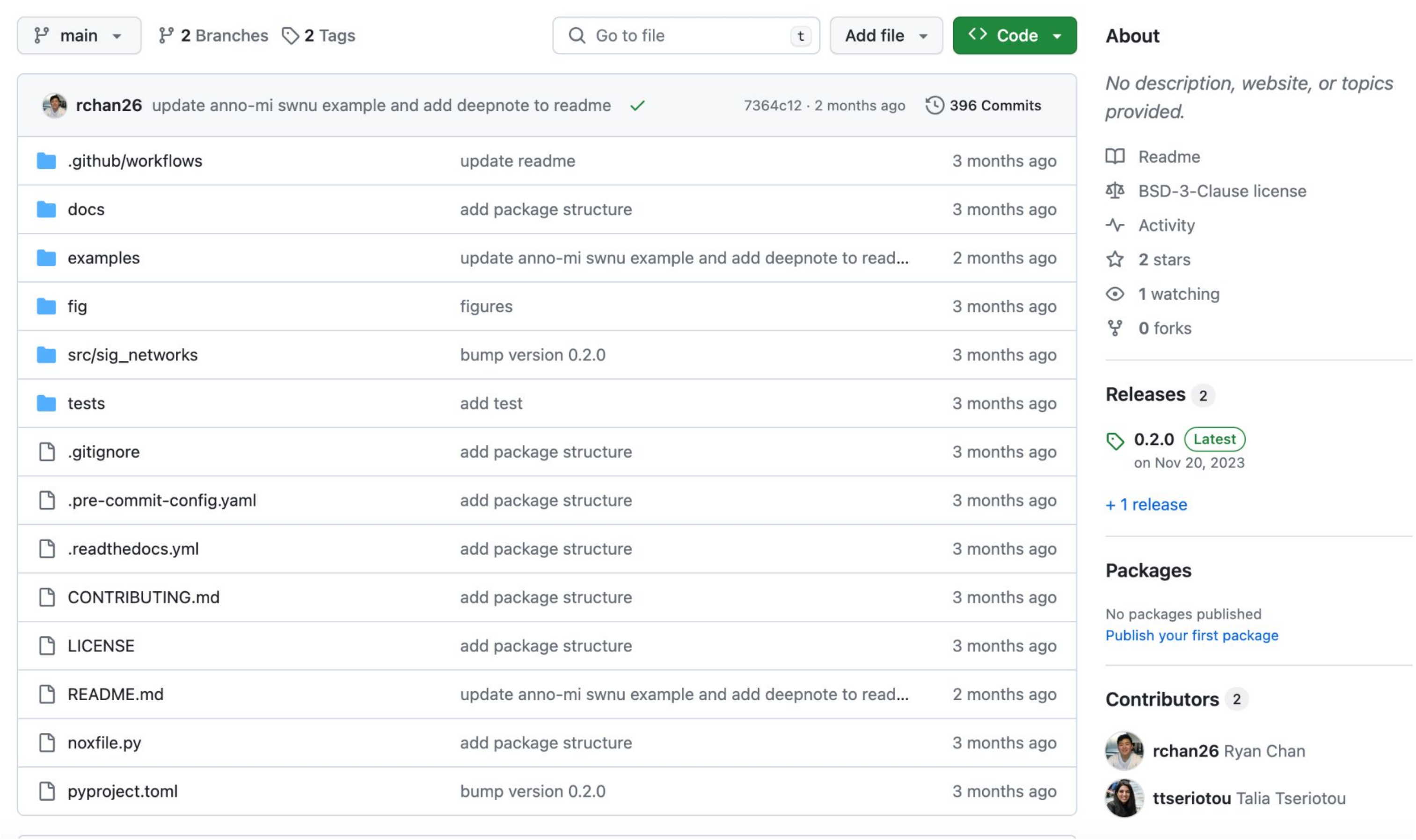
Sig-Networks Toolkit: Signature Networks for Longitudinal Language Modelling

First authors: Talia Tseriotou, Ryan Chan

Tseriotou, T., Chan, R., Tsakalidis, A., Bilal, I.M., Kochkina, E., Lyons, T. and Liakata, M., 2024, March. Sig-Networks Toolkit: Signature Networks for Longitudinal Language Modelling. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 223-237).

Sig-Networks Toolkit

pip installable PyTorch package for longitudinal NLP modelling.



https://github.com/ttseriotou/sig-networks

Contributions

1. **sig-networks** open-source pip installable toolkit for longitudinal NLP tasks.

Contributions

1. **sig-networks** open-source pip installable toolkit for longitudinal NLP tasks.
2. **nlpsig** pip installable library for data preprocessing.

Contributions

1. **sig-networks** open-source pip installable toolkit for longitudinal NLP tasks.
2. **nlpsig** pip installable library for data preprocessing.
3. **SOTA** performance on three tasks.

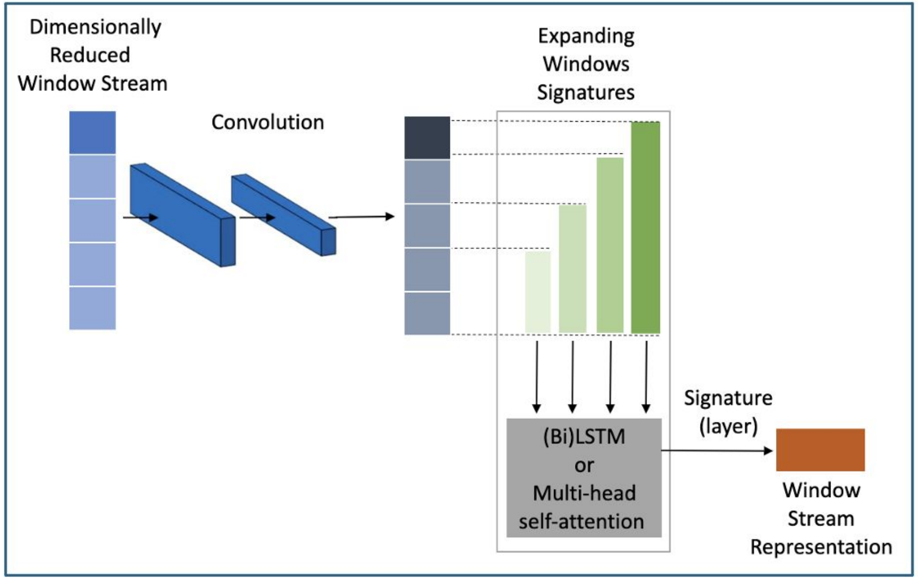
Contributions

1. **sig-networks** open-source pip installable toolkit for longitudinal NLP tasks.
2. **nlpsig** pip installable library for data preprocessing.
3. **SOTA** performance on three tasks.
4. Flexible dataset adaptation, model building blocks, benchmarking, feature and parameter selection.

Signature Network Models

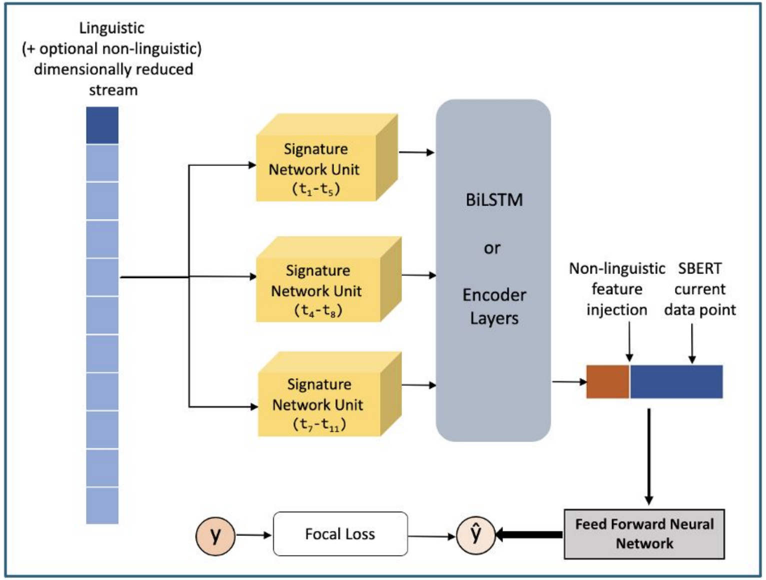
Window-based

Signatures over short expanding windows fed in BiLSTM/MHA

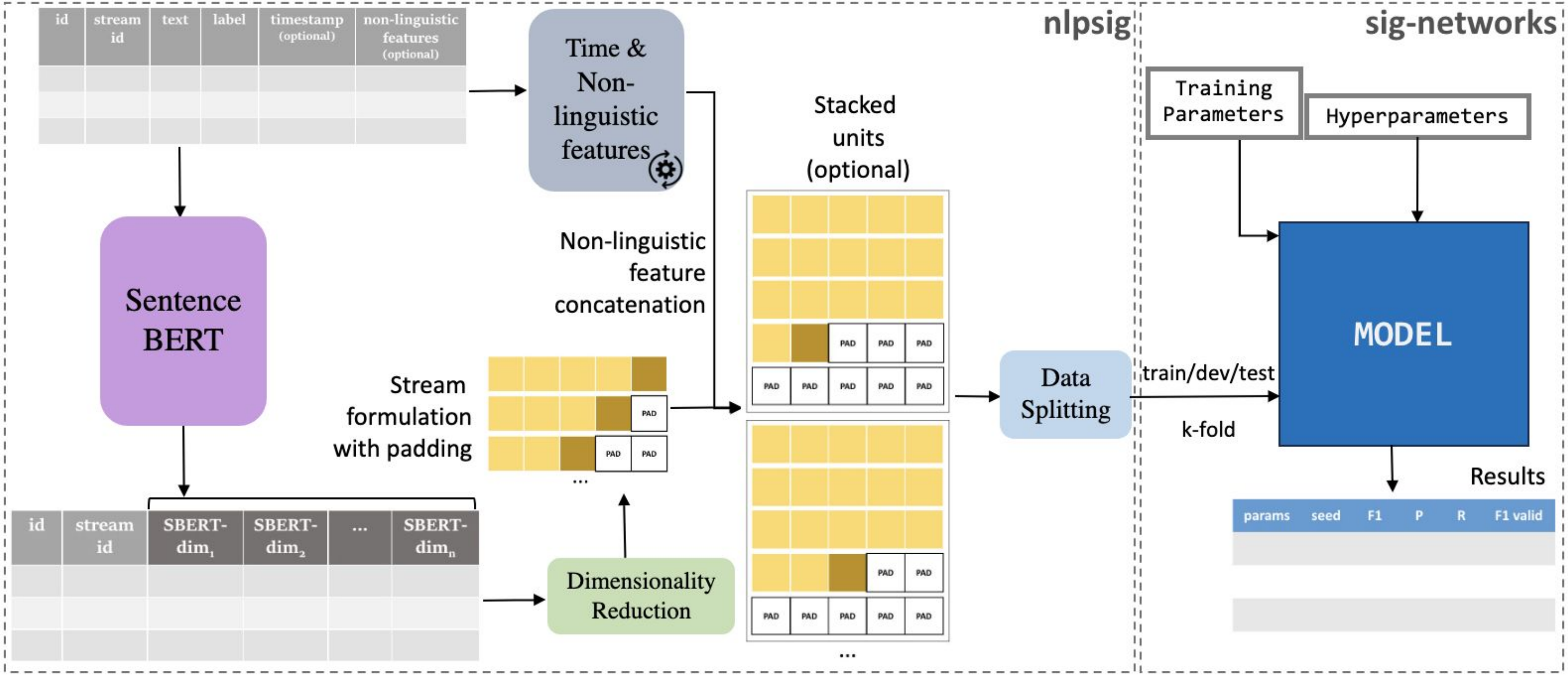


Unit-based

Sequential modeling of window units through a BiLSTM/Encoder



System Overview



System Components

Data Preparation and Training Modules

- nlpsig.encode_text → embedding generation
- nlpsig.DimReduce →dimensionality reduction
- nlpsig.PrepareData →padded embedding streams, time features calculation
- nlpsig.classification_utils → k-fold cross validation, stratified splitting, user-defined folds, data points exclusion.
- User selections: loss function, validation metric, patience, random seeds, grid search.

System Components

Data Preparation and Training Modules

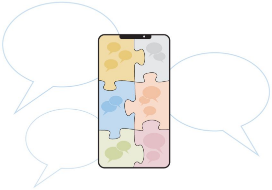
- nlpsig.encode_text → embedding generation
- nlpsig.DimReduce →dimensionality reduction
- nlpsig.PrepareData →padded embedding streams, time features calculation
- nlpsig.classification_utils → k-fold cross validation, stratified splitting, user-defined folds, data points exclusion.
- User selections: loss function, validation metric, patience, random seeds, grid search, feature concatenation in/out of path

Model Modules

- PyTorch classes as building blocks of our models to allow new systems development.
- Baselines: BERT, FFN (with/out history stream), BiLSTM
- Signature Network Models: range of options i.e. no. encoder layers.

Data

	AnnoMI	Longitudinal Rumour Stance	TalkLife
Description	Counselling Dialogues	Twitter conversations discussing rumours	Posts from peer-to-peer support network
Timelines	133	325	500
Data points	4,817 (client utterances)	5,568	18,604
Predicting	Client response type	Switch in the aggregate stance towards claim	Change in user’s mood



Results

To appear in EACL 2024 Sig-Networks Toolkit: Signature Networks for Longitudinal Language Modelling

Model	**Anno-MI (3-class)**			**LRS (2-class)**			**TalkLife (3-class)**		
BERT (focal)	.519			.589			.531		
BERT (ce)	.501			.596			.521		
FFN	.512			.581			.534		
FFN History	.520			.625			.537		
BiLSTM ($w = 5$)	.517			.637			.544		
SWNU ($w = 5$)	.522			.670			**.563**		
SW-Attn ($w = 5$)	.515			.667			.556		
History Length	**11**	**20**	**35**	**11**	**20**	**35**	**11**	**20**	**35**
#units (w**=5,** k**=3)**	**3**	**6**	**11**	**3**	**6**	**11**	**3**	**6**	**11**
BiLSTM	.518	.507	.510	.657	.648	.648	.539	.533	.525
SWNU	.522	.512	.493	.671	.654	.673	.550	.537	.539
SW-Attn	.517	.508	.508	.659	.665	.661	.547	.541	.539
Seq-Sig-Net	**.525**	.523	.517	.672	**.678**	.654	**.563**	.561	.559
SW-Attn+BiLSTM	.511	.514	.515	.663	.657	.660	.554	.557	.550
SW-Attn+Encoder	.498	.506	.505	.664	.657	.662	.552	.552	.545

Results

- Seq-Sig-Net achieves SOTA or on-par with SWNU across all tasks

Model	Anno-MI (3-class)			LRS (2-class)			TalkLife (3-class)		
BERT (focal)	.519			.589			.531		
BERT (ce)	.501			.596			.521		
FFN	.512			.581			.534		
FFN History	.520			.625			.537		
BiLSTM ($w = 5$)	.517			.637			.544		
SWNU ($w = 5$)	.522			.670			**.563**		
SW-Attn ($w = 5$)	.515			.667			.556		
History Length **#units (w=5, k=3)**	**11** **3**	**20** **6**	**35** **11**	**11** **3**	**20** **6**	**35** **11**	**11** **3**	**20** **6**	**35** **11**
BiLSTM	.518	.507	.510	.657	.648	.648	.539	.533	.525
SWNU	.522	.512	.493	.671	.654	<u>.673</u>	.550	.537	.539
SW-Attn	.517	.508	.508	.659	.665	.661	.547	.541	.539
Seq-Sig-Net	**.525**	<u>.523</u>	.517	.672	**.678**	.654	**.563**	<u>.561</u>	.559
SW-Attn+BiLSTM	.511	.514	.515	.663	.657	.660	.554	.557	.550
SW-Attn+Encoder	.498	.506	.505	.664	.657	.662	.552	.552	.545

Results

- In LRS and TalkLife SigNetworks outperforms all baselines, for each history length
- Anno-MI is the least longitudinal - small performance gains of sequential models

Model	**Anno-MI (3-class)**			**LRS (2-class)**			**TalkLife (3-class)**		
BERT (focal)	.519			.589			.531		
BERT (ce)	.501			.596			.521		
FFN	.512			.581			.534		
FFN History	.520			.625			.537		
BiLSTM ($w = 5$)	.517			.637			.544		
SWNU ($w = 5$)	.522			.670			**.563**		
SW-Attn ($w = 5$)	.515			.667			.556		
History Length	**11**	**20**	**35**	**11**	**20**	**35**	**11**	**20**	**35**
#units (w=5, k=3)	**3**	**6**	**11**	**3**	**6**	**11**	**3**	**6**	**11**
BiLSTM	.518	.507	.510	.657	.648	.648	.539	.533	.525
SWNU	.522	.512	.493	.671	.654	<u>.673</u>	.550	.537	.539
SW-Attn	.517	.508	.508	.659	.665	.661	.547	.541	.539
Seq-Sig-Net	**.525**	<u>.523</u>	.517	.672	**.678**	.654	**.563**	<u>.561</u>	.559
SW-Attn+BiLSTM	.511	.514	.515	.663	.657	.660	.554	.557	.550
SW-Attn+Encoder	.498	.506	.505	.664	.657	.662	.552	.552	.545

Time-Scale Analysis

- Different degree of temporal granularity

Dataset	Anno-MI		Longitudinal Rumour Stance	TalkLife MoC	
	Change	Sustain	Switch	Switch	Escalation
Mean Point Time Diff.	5sec		1hr 26min 40sec	6hr 51min 11sec	
Median Point Time Diff.	3sec		1min 39sec	59min 38sec	
Mean consecutive events	2.21	1.68	8.52	1.58	4.12
Median consecutive events	1	1	4	1	3
Mean no. of events in stream	8.86	4.05	6.45	1.77	4.03
Median no. of events in stream	5	3	0	1	1

TempoFormer: A Transformer for Temporally-aware Representations in Change Detection

First author: Talia Tseriotou

Tseriotou, T., Tsakalidis, A., & Liakata, M. (2024). TempoFormer: A Transformer for Temporally-aware Representations in Change Detection. Under Review.

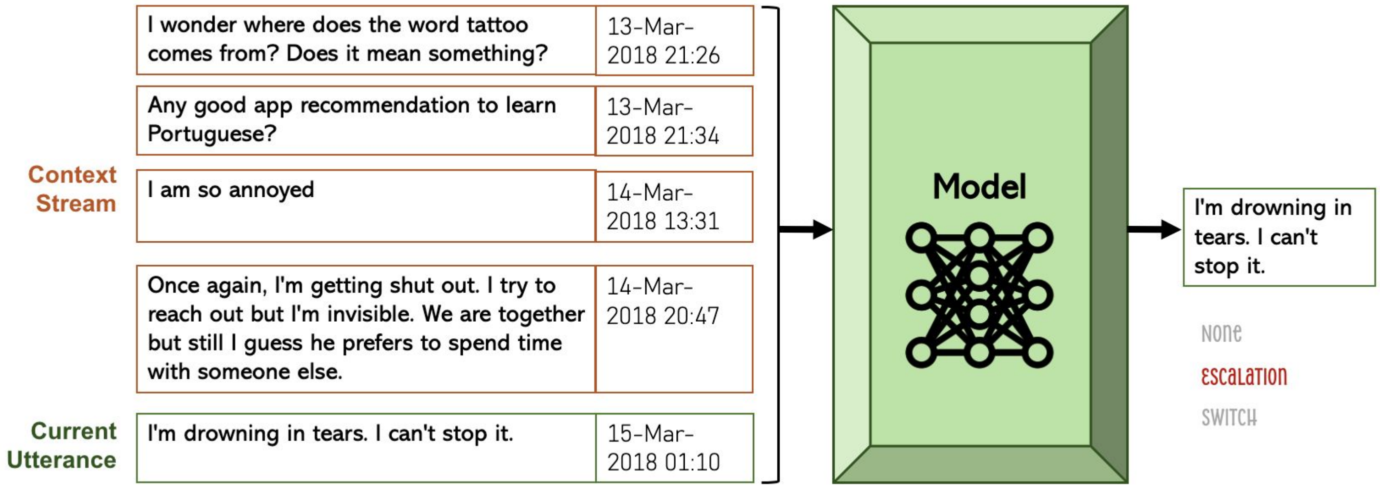
Limitations of previous work

- Uses recurrent models which are slow and prone to overfitting.
- Relying on pre-trained representations.

Can we strike even better balance between representation learning and task dynamics?

Change is everywhere

Example: Identifying Mood Changes in User Posts



Context is crucial

Time provides important signal

•

Time in LMs

Longitudinal Modeling:

- Limited research that accounts for time intervals between events.
- Contextualized representations concatenated with timestamps are not informed by temporal dynamics.

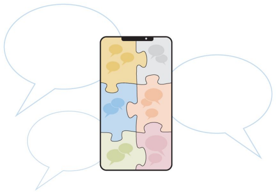
Language Models:

- Transformer-based models lack temporal sensitivity.
- LLMs have poor temporal reasoning capabilities.

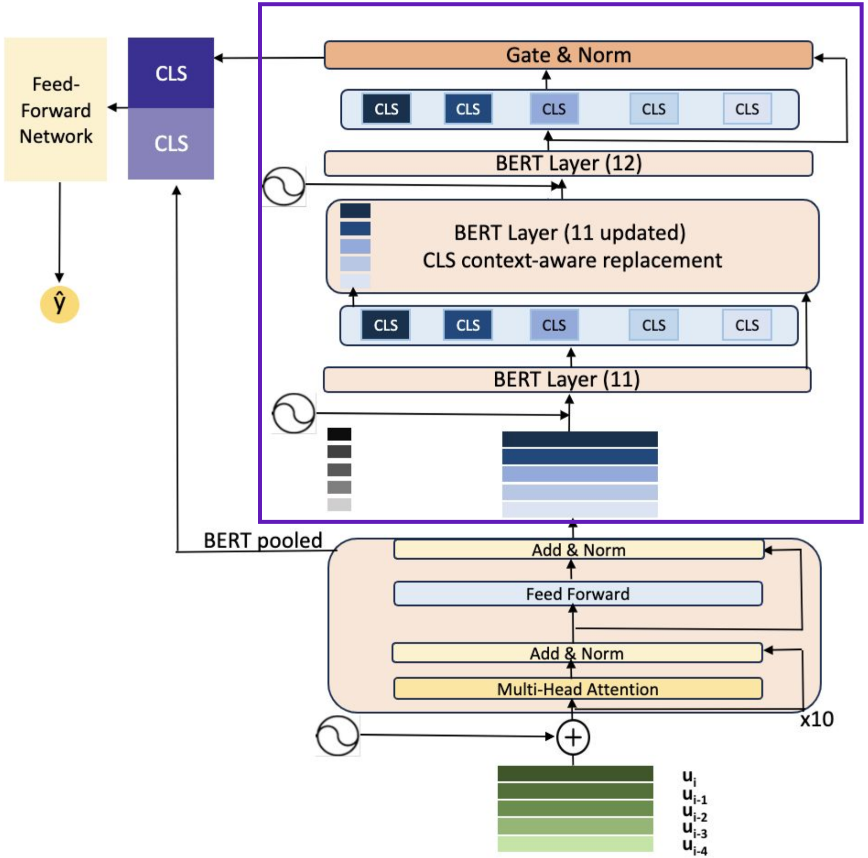
We introduce TempoFormer - a temporally aware transformer model.

Data

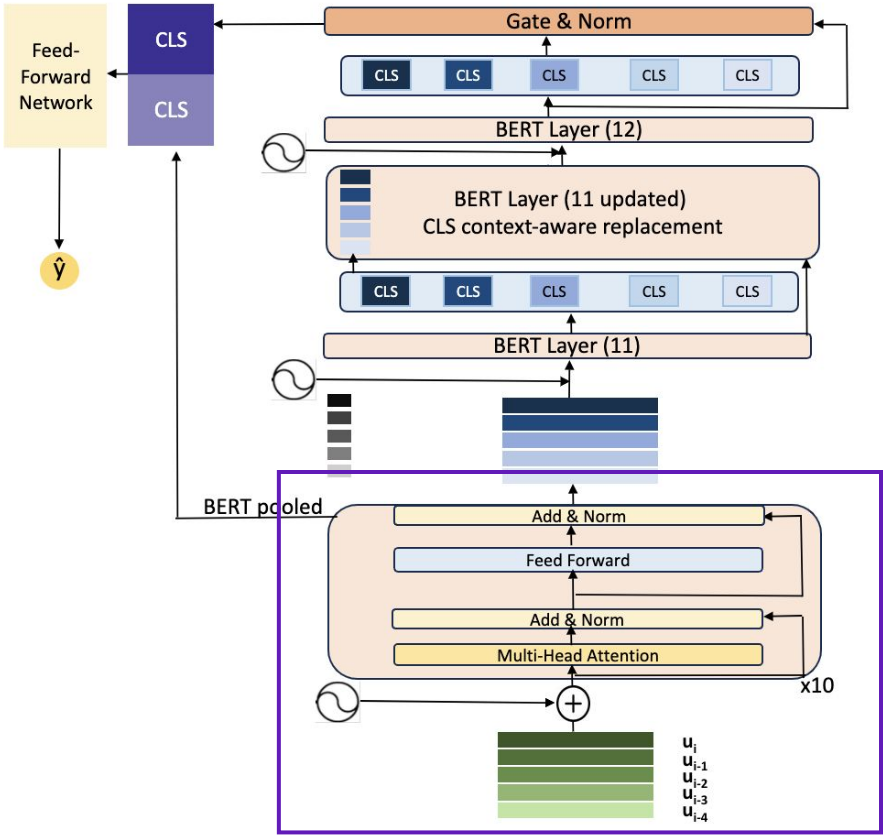
	Topic Shift	Longitudinal Rumour Stance	TalkLife
Description	Open-domain human to human conversation	Twitter conversations discussing rumours	Posts from peer-to-peer support network
Timelines	74	325	500
Data points	12,536	5,568	18,604
Predicting	Major topic derailment	Switch in the aggregate stance towards claim	Change in user's mood



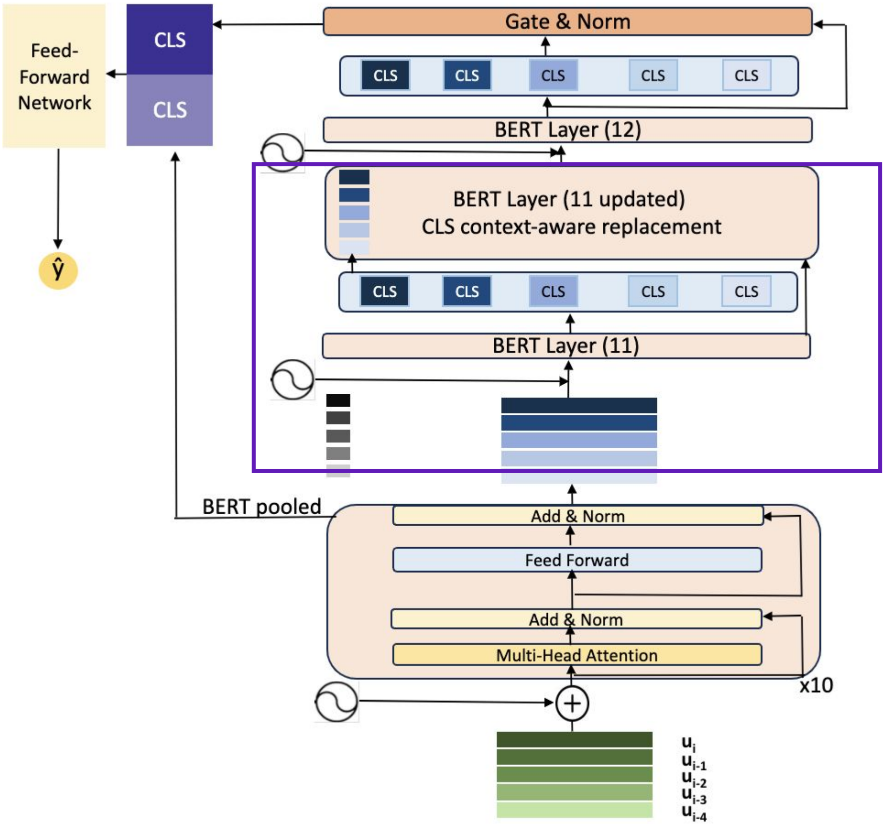
TempoFormer



TempoFormer

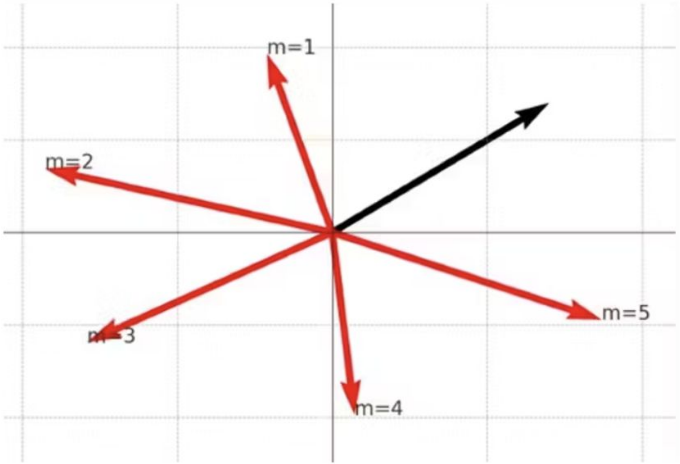


TempoFormer

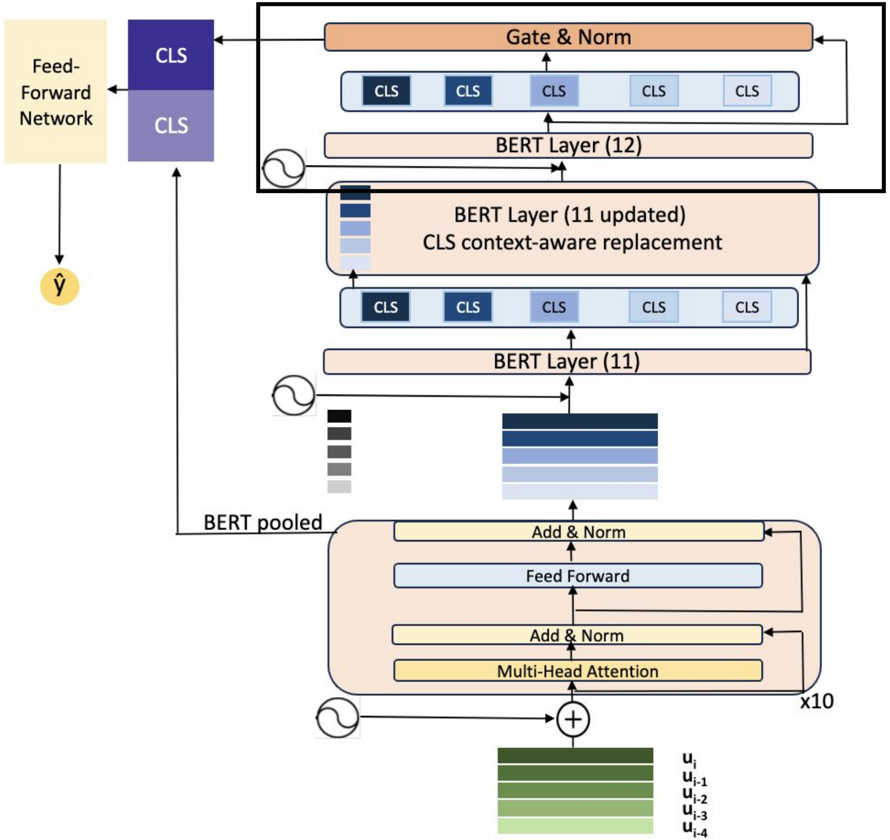


Temporal Rotary Positional Embeddings (RoPE)

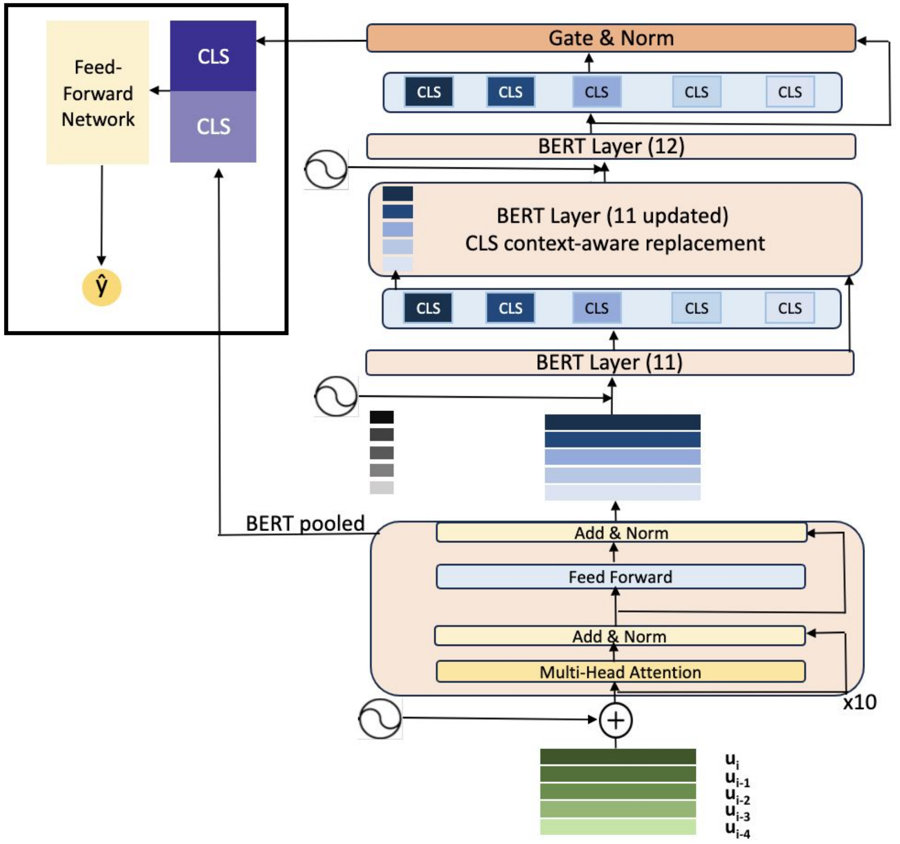
- **RoPE**: encodes the absolute position m of tokens through a rotation matrix **Rθ, m**
- Relative position (m-n) in self-attention.
- **Main idea:** embed token position by rotating q (query) and k (key) with different rotation at each position.
- q,k are rotated based on their ~~sequence position.~~ temporal position
- **Transformer**: rotation is applied to q,k before applying the attention mechanism.



TempoFormer



TempoFormer



TempoFormer

- **SOTA** performance of TempoFormer on all 3 datasets
- Dynamics **integrated** in the LM architecture

	Model	LRS N-Sw	LRS Sw	LRS macro-avg	TalkLife IE	TalkLife IS	TalkLife O	TalkLife macro-avg	Topic Shift MI M	Topic Shift MI R	Topic Shift MI macro-avg
Post-level	Random	61.4	37.5	49.5	11.2	4.5	84.4	33.4	35.9	63.9	49.9
	Llama2-7B-U (5-shot)	22.4	50.6	36.5	10.1	7.5	31.9	16.5	46.6	45.4	46.0
	MistralInst2-7B-U (5-shot)	71.4	28.0	49.7	23.3	4.1	67.8	31.7	46.4	44.6	45.5
	Llama2-7B-U (10-shot)	8.8	52.5	30.7	12.8	6.2	31.3	16.7	48.5	39.5	44.0
	MistralInst2-7B-U (10-shot)	71.2	30.5	50.8	27.6	3.5	72.1	34.4	42.6	55.7	49.1
	BERT	69.0	45.3	57.1	43.9	28.1	86.8	52.9	36.0	70.0	53.0
	RoBERTa	68.2	46.4	57.3	46.3	30.4	86.6	54.4	34.5	70.2	52.4
Stream-level	FFN History	71.6	52.8	62.2	45.4	27.1	88.0	53.5	39.4	70.1	54.8
	SWNU	75.5	55.5	65.5	48.0	29.3	**89.5**	55.6	38.7	66.0	52.3
	Seq-Sig-Net	74.7	58.9	66.8	48.4	30.2	**89.5**	56.0	37.4	66.7	52.1
	BiLSTM	75.0	60.7	67.8	46.1	27.0	89.2	54.1	37.8	**73.8**	55.8
	Llama2-7B-S (5-shot)	2.2	50.2	26.2	15.5	7.6	24.2	15.7	**52.6**	1.3	27.0
	MistralInst2-7B-S (5-shot)	58.3	50.2	54.3	22.0	4.6	70.0	32.2	42.3	57.3	49.8
	MistralInst2-7B-S (10-shot)	54.4	51.8	53.1	23.4	3.5	74.9	33.9	37.8	63.7	50.8
	TempoFormer (ours)	**75.9**	**62.0**	**68.9**	**50.0**	**32.4**	88.8	**57.1**	41.6	70.7	**56.1**

TempoFormer

- **Recurrent** models based on pretrained SBERT representations are now the **second best**.

	Model	LRS			TalkLife				Topic Shift MI		
		N-Sw	Sw	macro-avg	IE	IS	O	macro-avg	M	R	macro-avg
Post-level	Random	61.4	37.5	49.5	11.2	4.5	84.4	33.4	35.9	63.9	49.9
	Llama2-7B-U (5-shot)	22.4	50.6	36.5	10.1	7.5	31.9	16.5	46.6	45.4	46.0
	MistralInst2-7B-U (5-shot)	71.4	28.0	49.7	23.3	4.1	67.8	31.7	46.4	44.6	45.5
	Llama2-7B-U (10-shot)	8.8	52.5	30.7	12.8	6.2	31.3	16.7	48.5	39.5	44.0
	MistralInst2-7B-U (10-shot)	71.2	30.5	50.8	27.6	3.5	72.1	34.4	42.6	55.7	49.1
	BERT	69.0	45.3	57.1	43.9	28.1	86.8	52.9	36.0	70.0	53.0
	RoBERTa	68.2	46.4	57.3	46.3	30.4	86.6	54.4	34.5	70.2	52.4
Stream-level	FFN History	71.6	52.8	62.2	45.4	27.1	88.0	53.5	39.4	70.1	54.8
	SWNU	75.5	55.5	65.5	48.0	29.3	**89.5**	55.6	38.7	66.0	52.3
	Seq-Sig-Net	74.7	58.9	66.8	48.4	30.2	**89.5**	56.0	37.4	66.7	52.1
	BiLSTM	75.0	60.7	67.8	46.1	27.0	89.2	54.1	37.8	**73.8**	55.8
	Llama2-7B-S (5-shot)	2.2	50.2	26.2	15.5	7.6	24.2	15.7	**52.6**	1.3	27.0
	MistralInst2-7B-S (5-shot)	58.3	50.2	54.3	22.0	4.6	70.0	32.2	42.3	57.3	49.8
	MistralInst2-7B-S (10-shot)	54.4	51.8	53.1	23.4	3.5	74.9	33.9	37.8	63.7	50.8
	TempoFormer (ours)	**75.9**	**62.0**	**68.9**	**50.0**	**32.4**	88.8	**57.1**	41.6	70.7	**56.1**

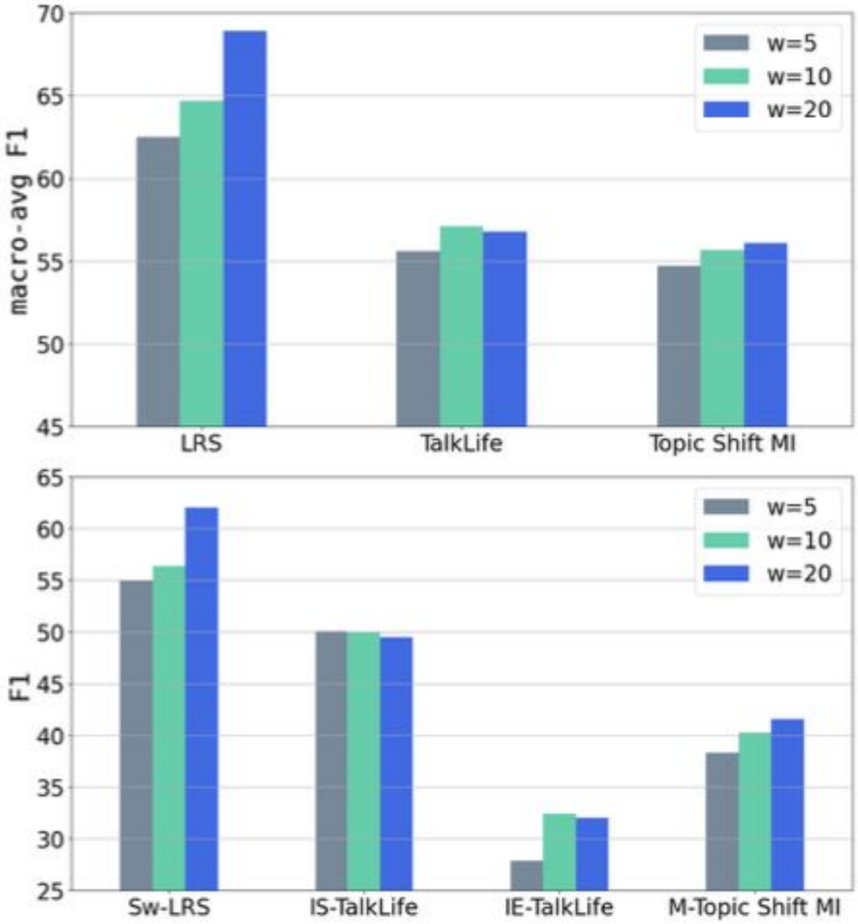
TempoFormer

- LLMs fall behind by a large margin

	Model	LRS			TalkLife				Topic Shift MI		
		N-Sw	Sw	macro-avg	IE	IS	O	macro-avg	M	R	macro-avg
Post-level	Random	61.4	37.5	49.5	11.2	4.5	84.4	33.4	35.9	63.9	49.9
	Llama2-7B-U (5-shot)	22.4	50.6	36.5	10.1	7.5	31.9	16.5	46.6	45.4	46.0
	MistralInst2-7B-U (5-shot)	71.4	28.0	49.7	23.3	4.1	67.8	31.7	46.4	44.6	45.5
	Llama2-7B-U (10-shot)	8.8	52.5	30.7	12.8	6.2	31.3	16.7	48.5	39.5	44.0
	MistralInst2-7B-U (10-shot)	71.2	30.5	50.8	27.6	3.5	72.1	34.4	42.6	55.7	49.1
	BERT	69.0	45.3	57.1	43.9	28.1	86.8	52.9	36.0	70.0	53.0
	RoBERTa	68.2	46.4	57.3	46.3	30.4	86.6	54.4	34.5	70.2	52.4
Stream-level	FFN History	71.6	52.8	62.2	45.4	27.1	88.0	53.5	39.4	70.1	54.8
	SWNU	75.5	55.5	65.5	48.0	29.3	**89.5**	55.6	38.7	66.0	52.3
	Seq-Sig-Net	74.7	58.9	66.8	48.4	30.2	**89.5**	56.0	37.4	66.7	52.1
	BiLSTM	75.0	60.7	67.8	46.1	27.0	89.2	54.1	37.8	**73.8**	55.8
	Llama2-7B-S (5-shot)	2.2	50.2	26.2	15.5	7.6	24.2	15.7	**52.6**	1.3	27.0
	MistralInst2-7B-S (5-shot)	58.3	50.2	54.3	22.0	4.6	70.0	32.2	42.3	57.3	49.8
	MistralInst2-7B-S (10-shot)	54.4	51.8	53.1	23.4	3.5	74.9	33.9	37.8	63.7	50.8
	TempoFormer (ours)	**75.9**	**62.0**	**68.9**	**50.0**	**32.4**	88.8	**57.1**	41.6	70.7	**56.1**

Window Selection

Determined by model performance and dataset characteristics



Ablation Study

Temporal RoPE: advantage of modeling temporal dynamics

RoPE MHA: enables MHA integration without FFN

Stream embeddings: propagation of position information

Gate&Norm: fusion of stream and word dynamics

	LRS			TalkLife				Topic Shift MI		
Models	N-Sw	Sw	macro-avg	IE	IS	O	macro-avg	M	R	macro-avg
TempoFormer	**75.9**	**62.0**	**68.9**	**50.0**	**32.4**	88.8	**57.1**	**41.6**	70.7	**56.1**
¬Temporal RoPE	75.5	**62.0**	68.7	49.3	31.7	88.7	56.6	-	-	-
¬RoPE MHA	74.1	57.9	66.0	48.0	31.5	88.2	55.9	39.6	**71.4**	55.5
¬Stream embed.	75.4	59.0	67.2	49.4	31.7	**89.2**	56.8	38.9	70.5	54.7
¬Gate&Norm	74.5	61.3	67.9	49.8	31.1	88.7	56.6	40.7	69.6	55.2

The curious case of recurrence

- Adapting TempoFormer for **recurrence** → over each post's CLS tokens.
- RoTempoFormer consistently **outperforms RoBERT.**
- Right balance between context aware post representations and recurrence in stream dynamics.
- **TempoFormer is a foundation model for temporal representation learning.**

	LRS			TalkLife				Topic Shift MI		
model	N-Sw	Sw	macro-avg	IE	IS	O	macro-avg	M	R	macro-avg
TempoFormer	75.9	62.0	68.9	**50.0**	**32.4**	**88.8**	**57.1**	**41.6**	70.7	**56.1**
RoBERT	75.8	62.3	69.0	36.7	3.3	88.4	42.8	33.3	**75.7**	54.5
RoTempoFormer	**76.2**	**63.6**	**69.9**	47.1	27.5	88.3	54.3	36.6	73.2	54.9

Model Adaptability

- Examining if the RoBERTa gain in TalkLife transfers to the TempoFormer.
- TempoFormer RoBERTa achieves a **new SOTA of 58.8%** F1 driven by IS and IE.
- **TempoFormer is adaptable.**

model	IE	IS	O	macro-avg
BERT	43.9	28.1	86.8	52.9
RoBERTa	46.3	30.4	86.6	54.4
TempoFormer (BERT)	50.0	32.4	**88.8**	57.1
TempoFormer (RoBERTa)	**52.4**	**36.9**	87.3	**58.8**

Lecture IV: Timeline extraction, Timeline summarisation and demo on identifying longitudinal changes.

Combining Hierarchical VAEs with LLMs for clinically meaningful timeline summarisation in social media

First authors: Jiayu Song, Jenny Chim

Song, J., Chim, J., Tsakalidis, A., Ive, J., Atzil-Slonim, D. and Liakata, M., 2024. Combining Hierachical VAEs with LLMs for clinically meaningful timeline summarisation in social media. (To appear in Proceedings of ACL 2024, Findings)

Motivation

User-generated content (UGC) contains behavioural cues

- Enable studying the evolution of individual’s mental health over time

- Augment clinician capacity, beyond patient self-reports.

Motivation

Timeline : sequences of chronologically ordered posts by a user

how to get rid of hunger pains without eating
I hate it that everyday I have to lie to my mom. I'm feeling guilty about it but if I tell her I'm in trouble
home alone what to do in order to keep positive and not fall through the cracks
any tips on what exercise to do in order to get a thigh gap
Well I guess my one month of no cutting is fone. I knew it the whole time I'm a failure at everything
last week snow this week rain and its April what the he'll is going on this is making me depressed
My mom is making me want to see blood flowing from my skin. I already cut today I don't want more. what to do..
I can't hold down any food so my mom says take a anti acid but it doesn't help and now my throat is killing me any one know what I should do
My homeschooling begins tomorrow I'm not ready for school yet
time to get another test about 50 needles into my skin
Why me I just feel like crying. when I get home i think the blade will be my best friend. my mom needs to quit tricking me and dragging me places without telling me. I'm done with life. I'm not worth it. I will never be. I'm fat, ugly, worthless, freak and invisible. why did I have to be born……….
I want to ruin my 5 clean so back to my old self
thanks mom now i want to cut
well other goes my 6 days of being clean well I may as well die
Well now I have the stuff to commit suicide if I want to so my mom better stop fucken pressuring me

⋮

Clinically meaningful timeline summary:

capture fluctuations in individuals' state-of-mind
(can assist in monitoring, prevention and early detection of mental health issues.)

organize information about a person's diagnosis, behavior, emotions and cognition
(can help clinicians understand important aspects of the individual)

Clinical Concepts

Diagnosis/mental states	• Presenting issues. • Mental health symptoms, level of functioning, well being. • Physical symptoms. • Risk assessment. • Motivation to change. • Lifestyle. • Agency, coping mechanisms, strengths and resources. • Meaning/goals/direction in life • Behaviour. • Important events
Intrapersonal and Interpersonal patterns	• Main need/wish/desire. • Interpersonal relationships. • Self perception, self esteem
Moments of change	• Sort of emotion: sad, happy etc; • Arousal level: high/low; • Emotion regulation strategies • *Switches*: drastic change of one's mood. • *Escalations*: intensification in one's mood. • Self understanding.

Challenges

- How do we create clinically meaningful summaries from long rambling social media timelines?
- How do we identify and synthesise useful information from the timelines?
- How do we evaluate this type of summaries?
- Difference to timeline summarisation in news (few distinct topics, explosion of emotions, recurring themes)
- Difference to non-temporal mental health summaries

Objectives

O1: Develop an unsupervised timeline summarization method that captures important information from social media timelines

O2: Provide a method for translating timeline summaries to a clinically meaningful format

O3: Develop an evaluation strategy for assessing these types of summary

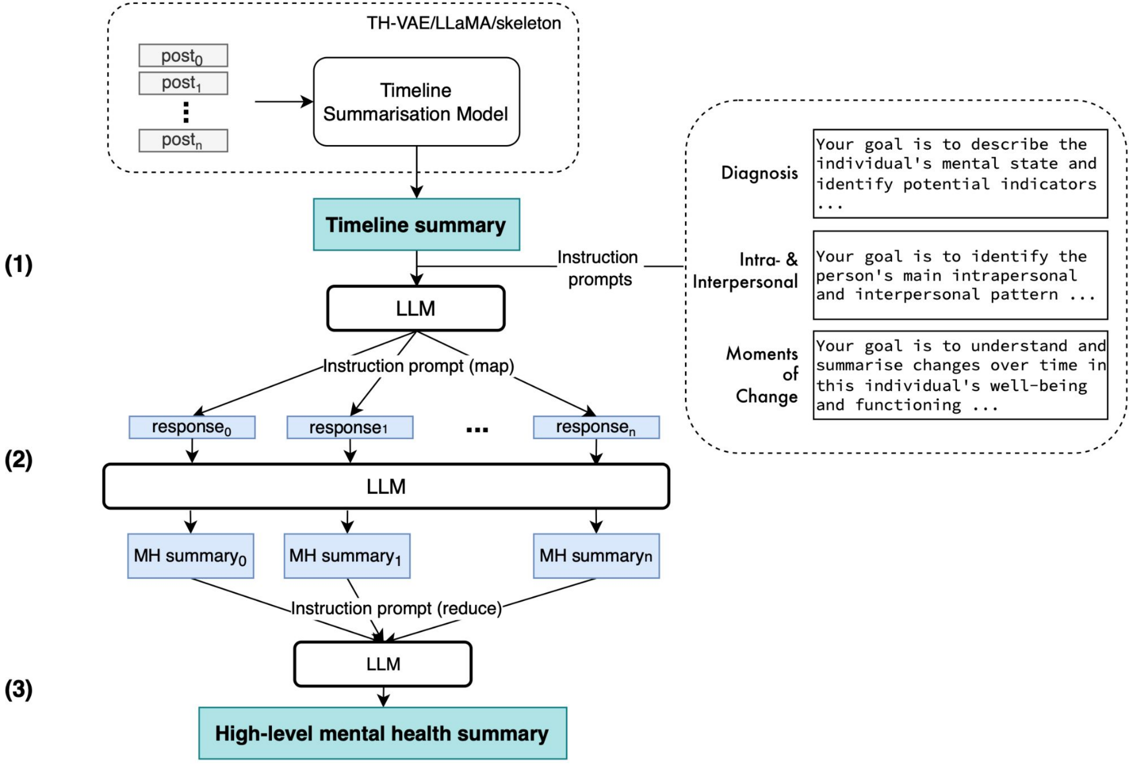
Solution: Summary with two different narratives

Evidence summary (first person) : abstractive summary of a user's timeline summary focussing on important points.

High-level summary (third person): high-level information useful for a clinician

- individuals' diagnosis,
- intrapersonal and interpersonal patterns
- the extent to which their mental state changes over time

Model



1. **Timeline/Evidence summary:** Generate an abstractive Timeline/Evidence summary with unsupervised method

2. **High-level summary:** feed the generated timeline summary into LLM to generate high-level summary corresponding to clinical content covering aspects such as diagnosis, inter- and intra- personal relationships and fluctuations in mood.

Timeline Example

"Gonna be offline for a bit, having lunch atm"

Timeline

O

t_n

Timeline Example

“Gonna be offline for a bit, having lunch atm”

“OK, I am back. Kinda bored.”

Timeline

O	O
t_n	t_{n+1}

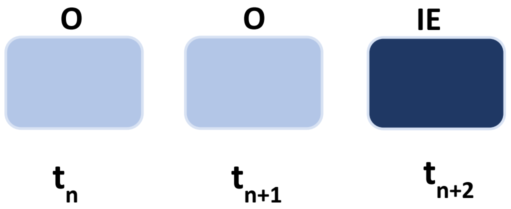
Timeline Example

“Gonna be offline for a bit, having lunch atm”

“OK, I am back. Kinda bored.”

“Everything is just so wrong in my life”

Timeline



Timeline Example

“Gonna be offline for a bit, having lunch atm”

“OK, I am back. Kinda bored.”

“Everything is just so wrong in my life”

“Omg can’t stop crying, everything is ruined”

Timeline

O	O	IE	IE
t_n	t_{n+1}	t_{n+2}	t_{n+3}

Timeline Example

“Gonna be offline for a bit, having lunch atm”

“OK, I am back. Kinda bored.”

“Everything is just so wrong in my life”

“Omg can’t stop crying, everything is ruined”

“Need someone before I do something stupid !PLEASE HELP!”

Timeline

O	O	IE	IE	IE Peak
t_n	t_{n+1}	t_{n+2}	t_{n+3}	t_{n+4}

Timeline Example

“Gonna be offline for a bit, having lunch atm”

“OK, I am back. Kinda bored.”

“Everything is just so wrong in my life”

“Omg can’t stop crying, everything is ruined”

“Need someone before I do something stupid !PLEASE HELP!”

“Wish things were differently.. Miss my gf..”

Timeline

O	**O**	**IE**	**IE**	**IE Peak**	**IE**
t_n	t_{n+1}	t_{n+2}	t_{n+3}	t_{n+4}	t_{n+5}

Timeline Example

“Gonna be offline for a bit, having lunch atm”

“OK, I am back. Kinda bored.”

“Everything is just so wrong in my life”

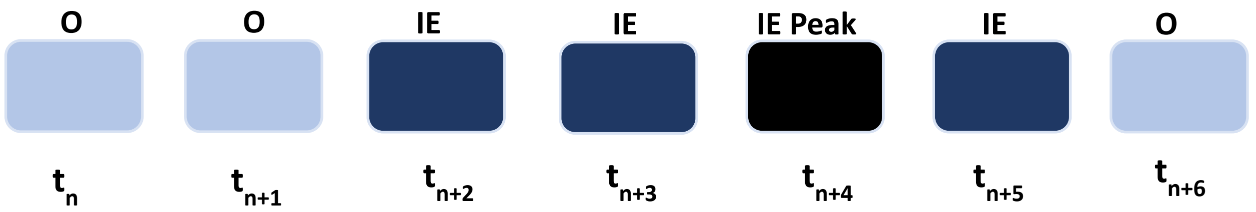
“Omg can’t stop crying, everything is ruined”

“Need someone before I do something stupid !PLEASE HELP!”

“Wish things were differently.. Miss my gf..”

“Having an exam in one week, hope to be able to do well”

Timeline



Timeline Example

“Gonna be offline for a bit, having lunch atm”

“OK, I am back. Kinda bored.”

“Everything is just so wrong in my life”

“Omg can’t stop crying, everything is ruined”

“Need someone before I do something stupid !PLEASE HELP!”

“Wish things were differently.. Miss my gf..”

“Having an exam in one week, hope to be able to do well”

“Has anyone watched the last episode of Game of Thrones? Is it any good?”

Timeline

O	O	IE	IE	IE Peak	IE	O	O
t_n	t_{n+1}	t_{n+2}	t_{n+3}	t_{n+4}	t_{n+5}	t_{n+6}	t_{n+7}

Timeline Example

“Gonna be offline for a bit, having lunch atm”

“OK, I am back. Kinda bored.”

“Everything is just so wrong in my life”

“Omg can’t stop crying, everything is ruined”

“Need someone before I do something stupid !PLEASE HELP!”

“Wish things were differently.. Miss my gf..”

“Having an exam in one week, hope to be able to do well”

“Has anyone watched the last episode of Game of Thrones? Is it any good?”

“My partner called, told me she misses me and wants to meet! Best day E-V-E-R!!”

Timeline

O	O	IE	IE	IE Peak	IE	O	O	IS
t_n	t_{n+1}	t_{n+2}	t_{n+3}	t_{n+4}	t_{n+5}	t_{n+6}	t_{n+7}	t_{n+8}

Timeline Example

“Gonna be offline for a bit, having lunch atm”

“OK, I am back. Kinda bored.”

“Everything is just so wrong in my life”

“Omg can’t stop crying, everything is ruined”

“Need someone before I do something stupid !PLEASE HELP!”

“Wish things were differently.. Miss my gf..”

“Having an exam in one week, hope to be able to do well”

“Has anyone watched the last episode of Game of Thrones? Is it any good?”

“My partner called, told me she misses me and wants to meet! Best day E-V-E-R!!”

“Can anyone recommend any decent crime book to read”?

Timeline

O	O	IE	IE	IE Peak	IE	O	O	IS	O
t_n	t_{n+1}	t_{n+2}	t_{n+3}	t_{n+4}	t_{n+5}	t_{n+6}	t_{n+7}	t_{n+8}	t_{n+9}

Social media timelines

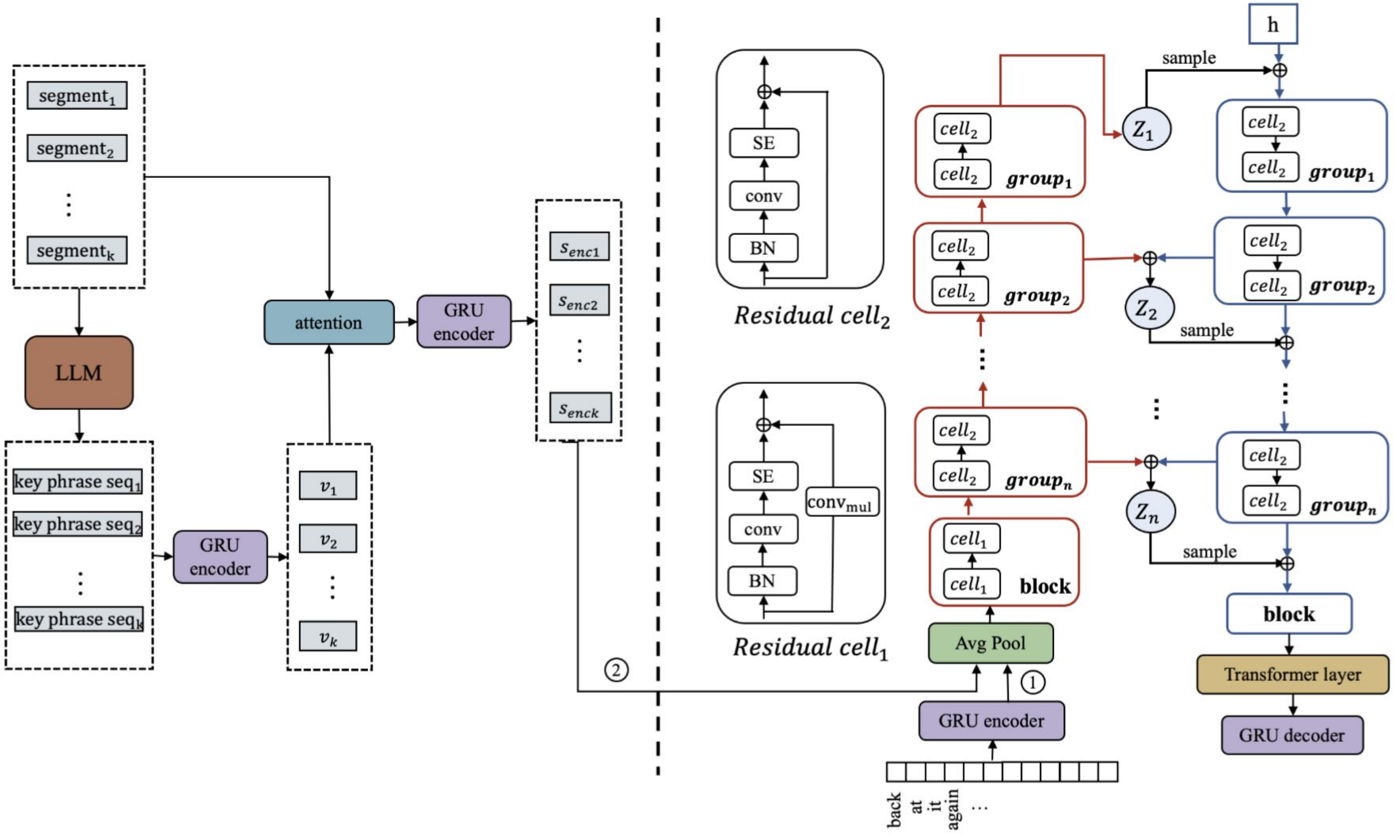
segment_1	Back at it again with using sex as self harm …	IE
	I'm so suicidal right now …	IEP
	……	…
segment_i	Can't believe I'm finally starting to find reasons…you'll always have the first…	ISB
	Finally gonna get my sleeping back on track (at least I hope) thanks to staying up all night …	IS
	……	…

Dataset collected by Tsakalidis et al.(2022b) consisting of 500 users' timelines from Talklife, anonymised and annotated with MoC (Moments of Change).

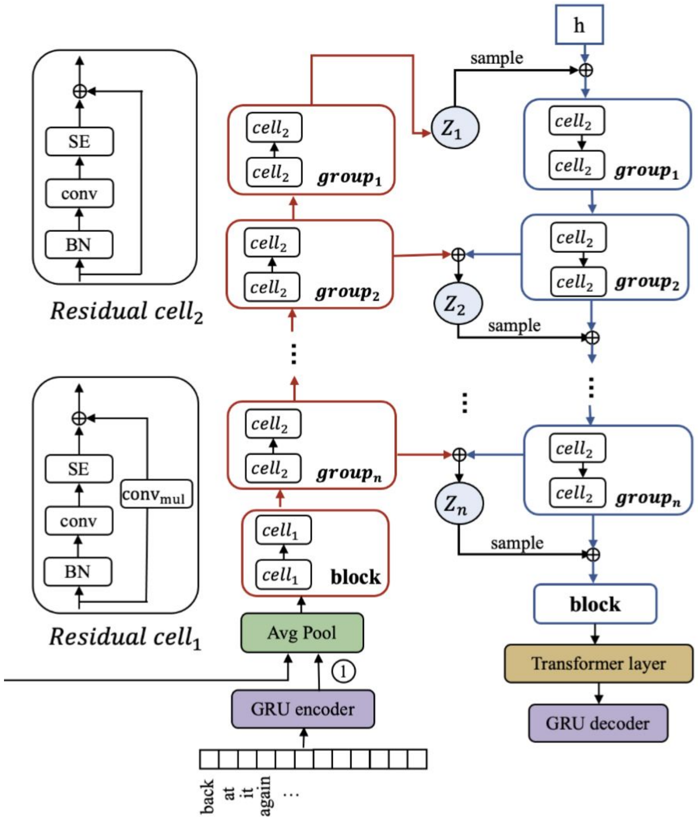
Switch: sudden mood shift
IS (*In Switch*) and ISB (*In Switch Beginning*)
Escalation: gradual mood progression
IE (*In Escalation*) and IEP (*In Escalation Peak*)

Split the timeline into several segments (sub-timeline)
Based on 'Moc'

TH-VAE


segment$_1$
segment$_2$
segment$_k$
LLM
key phrase seq$_1$
key phrase seq$_2$
key phrase seq$_k$
GRU encoder
v_1
v_2
v_k
attention
GRU encoder
s_{enc1}
s_{enc2}
s_{enck}
②
SE
conv
BN
Residual cell$_2$
SE
conv
BN
conv$_{mul}$
Residual cell$_1$
cell$_2$
cell$_2$
group$_1$
cell$_2$
cell$_2$
group$_2$
cell$_2$
cell$_2$
group$_n$
cell$_1$
cell$_1$
block
Avg Pool
①
GRU encoder
back
at
it
again
…
Z_1
Z_2
Z_n
sample
sample
sample
h
cell$_2$
cell$_2$
group$_1$
cell$_2$
cell$_2$
group$_2$
cell$_2$
cell$_2$
group$_n$
block
Transformer layer
GRU decoder

TH-VAE



$$\boldsymbol{z}=\{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_l\}$$

$$q_\phi(\boldsymbol{z}|\boldsymbol{x})$$

$$q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \prod_l q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{<l}, \boldsymbol{x})$$

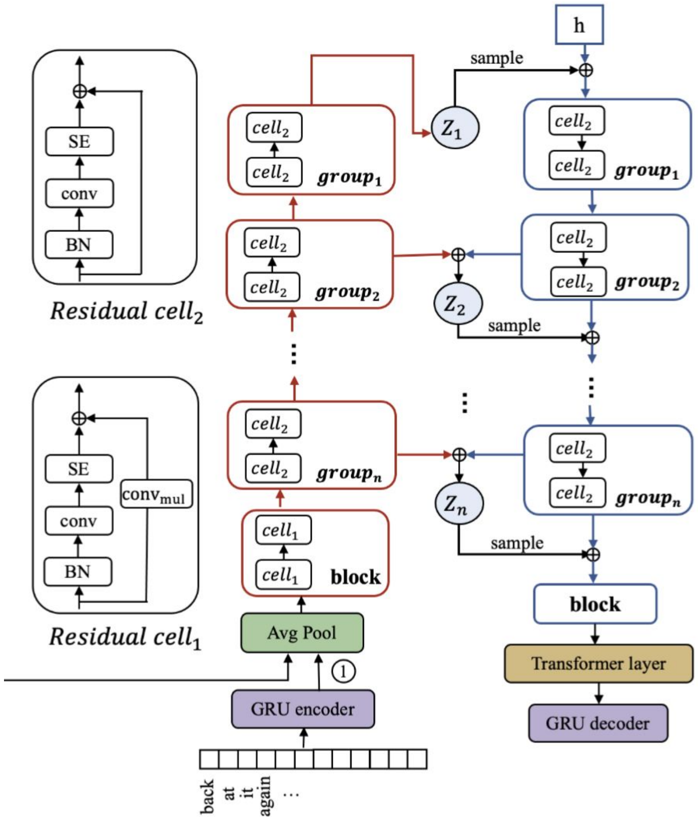
$$p(\boldsymbol{z})$$

$$p(\boldsymbol{z}) = \prod_l p(\boldsymbol{z}_l|\boldsymbol{z}_{<l})$$

Maximise Lower bound:

$$L(\theta;\boldsymbol{x}) = -KL(q_\phi(\boldsymbol{z}_1|\boldsymbol{x})||p(\boldsymbol{z}_1)) \sum_{l=2}^{L} \mathbb{E}_{q_\phi(\boldsymbol{z}<l|\boldsymbol{x})}[-KL(q_\phi(\boldsymbol{z}_l|\boldsymbol{x},\boldsymbol{z}<l)||p(\boldsymbol{z}_l|\boldsymbol{z}<l))] + \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})].$$

TH-VAE



BN: batch normalization with swish activation

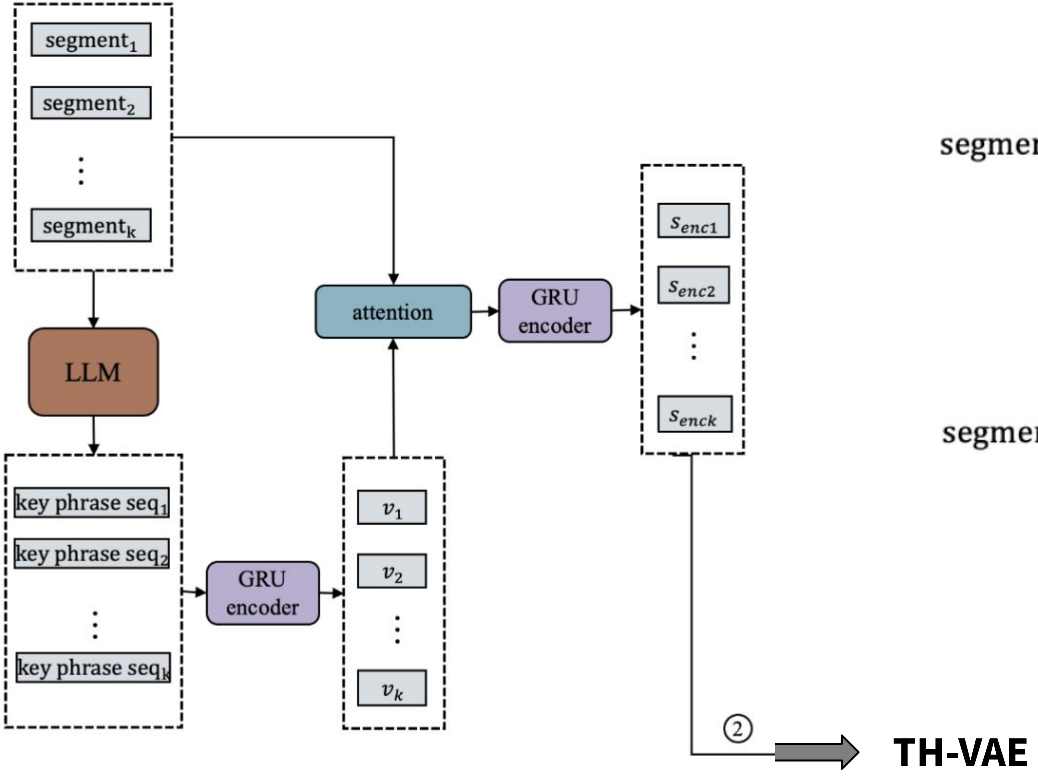
SE: squeeze and excitation

Series BN, conv (CNN), SE ➡ Residual Cell

Block: residual cell1 (capture different features)
Group: residual cell2 (establish long-range relationships), build the hierarchy by layered groups

Transformer layer: load parameter from BART-base

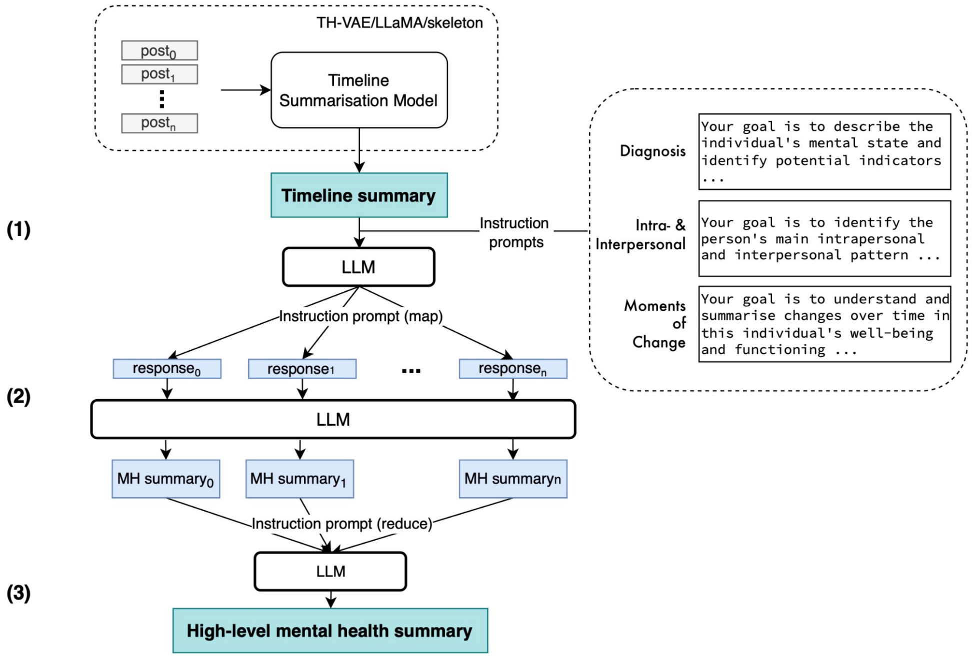
Summary representation for generation



segment$_1$	Back at it again with using sex as self harm …	IE
	I'm so suicidal right now …	IEP
	……	…
segment$_i$	Can't believe I'm finally starting to find reasons…you'll always have the first…	ISB
	Finally gonna get my sleeping back on track (at least I hope) thanks to staying up all night …	IS
	……	…

timeline / key phrases pairs

Methods: High-level Summary

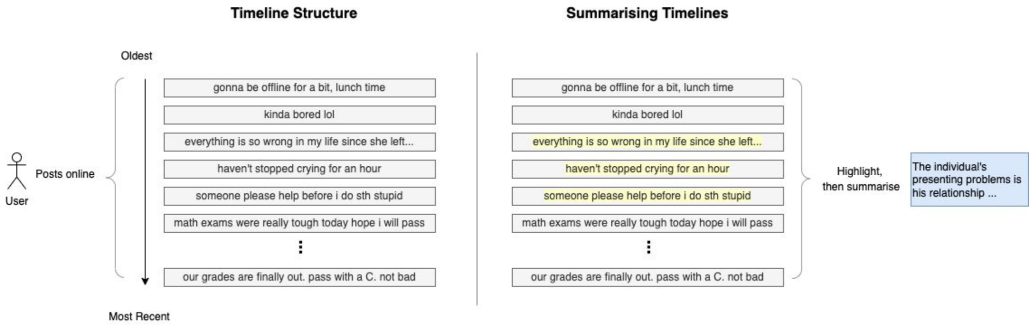


- Input: timeline summary
- Prompting framework
  - *Map*: Iteratively prompt instruction-tuned LLM to extract information around various clinical concepts
  - *Reduce*: Combine extracted information into prose, then distill into single coherent document
- Output: High-level mental health summary capturing overall insights & user changes over time

Methods: Gold standard data creation

Test set with gold reference
mental health summaries

- We sampled 30 timelines from the TalkLife MoC dataset
- Annotators trained in psychology
    - Three annotators
    - Non-native fluent English speakers
    - Postgraduate students in clinical psychology
    - Completed three training sessions under the supervision of a senior clinical expert



Annotators highlighted clinically relevant information in user timelines, then wrote mental health summaries (high-level observations with relevant supporting evidence).

Evaluation

What makes a good summary in this scenario?

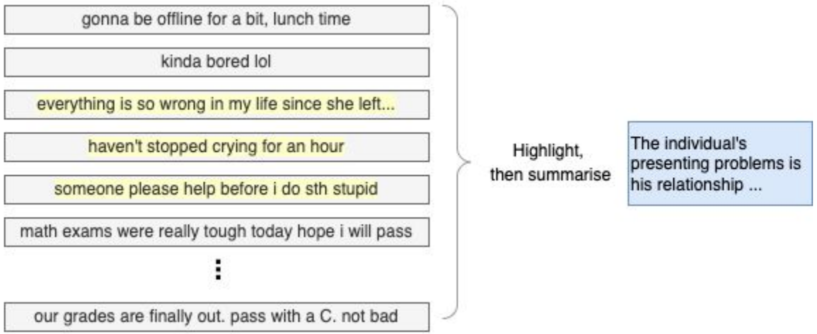
- Salient meaning preservation: does it capture (clinically) important information?
- Factual consistency
  - Timeline summaries consistent with original timeline data?
  - High-level observations consistent with gold mental health summaries?
    - In terms of (1) overall observations and (2) moments of change
- Evidence appropriateness
  - High-level summaries consistent with respective timeline evidence?
    - In terms of (1) overall observations and (2) moments of change
- Coherence: do summaries present a logically connected, easy to follow body of information?
- Fluency: are texts in the summary well-formed and natural?

Evaluation

Salient meaning preservation

Srivastava et al., 2022 proposed MHIC for psychotherapy conversations. We adapt the metric to our task.

1) Use semantic intersection of gold highlighted evidence spans
2) Compute semantic similarity, instead of ROUGE

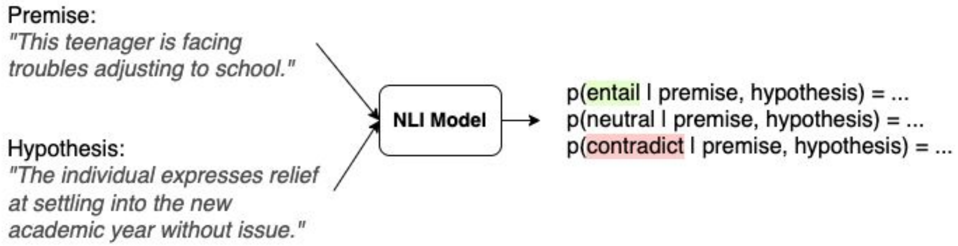


$$\mathrm{MHIC}_{sem} = \frac{1}{|E|} \sum_{e \in E} \max_{t \in T} R_{\mathrm{BERT}}(e, t)$$

Evaluation

Factual Consistency

- Prior work in summary evaluation has used NLI to measure factuality and summary consistency (Maynez et al., 2020; Laban et al., 2022; inter alia).
- We use a RoBERTa model fine-tuned on NLI, ANLI, FEVER (Nie et al., 2020).
- Since our summaries are two-layered:
  - Consistency with timeline
  - Consistency with gold summaries (main summary body, MoC)



Evaluation

Factual Consistency

- Timeline summary with source timeline
  - Faithfulness score (Maynez et al., 2020)
  - Modify chunking procedure
- High-level summary with gold summaries
  - Separately compute consistency score for:
    - m: main summary body
    - c: segment focusing on changes

$$\mathrm{FC}_{\text{Timeline}} = \frac{1}{|T|} \sum_{t \in T} \max_{d \in D} \mathrm{NLI}(\mathrm{Entail}|d, t)$$

$$C(A, B) = \frac{1}{|A| \cdot |B|} \sum_{a \in A} \sum_{b \in B} (1 - \mathrm{NLI}(\mathrm{Contradict}|a, b))$$

Evaluation

Evidence Appropriateness

- Consistency of high-level summaries' (1) main body and (2) section focused on changes with their respective summarised timeline evidence timeline summary

$$\mathrm{EA_c} = \frac{1}{|S|}\sum_{s \in S} \mathrm{NLI}(\mathrm{Entail}|T', s).$$

Coherence

- How easy is it to follow the summary? - BARTScore (Yuan et al., 2021)
- Intra-summary logical consistency - IntraNLI

Fluency

- PPL - Perplexity under GPT-2 (Radford et al., 2019)

Example Summary

Timeline Summary	High-level Summary
I hate school. im so scared.i love the cat so much. i'm gonna go be counter for a while. unknown good night all! i hate my nose. fuck the sat. i got a commitment ring 3 i've got counterbodied under my eyes. i feel alone i feel like everyone hates me. i hate saying this because it sounds bad, but i really want someone to pay attention to me. solid my mom only cares about my brothers learning disabilities and her new boyfriend expressive. my boyfriend goes throug of ptsd like depression symptoms, so he gets distant sometimes levels. i literally have no friends argue. i've always been quiet but sometimes i just want someone never pay attention to me. i argue to know only when't be a body. i have to know. i'll have to be happy.	The individual presents with multiple mental health concerns, including low self-esteem, anxiety, depression, relationship difficulties. They experience feelings of isolation and express fear and sadness, with specific triggers such as their mother's favoritism and their boyfriend's PTSD symptoms. Negative self-talk and body dissatisfaction, particularly regarding their nose, contribute to their low self-esteem. To address these challenges, the individual requires comprehensive and compassionate support, including attention, validation, and recognition of their feelings and experiences. The individual's mood remains consistently negative throughout the texts, with expressions of sadness, loneliness, and frustration. Specific changes include increased self-reported stress, feelings of isolation, and decreased self-esteem. However, there are also positive changes, such as excitement about receiving a commitment ring and a desire for human connection. No evidence of intensification of moods is present in the provided texts.

Table 1: Example TH-VAE timeline summary and its high-level summary. Examples for all systems in Appendix A.

Results

Aspect	Metric	LLaMA	TH-VAE	Skeleton	Naive
SMP	MHIC_{sem}	.65	**.66**	.57	–
FC	$\text{FC}_{\text{Timeline}}$	**.63**	**.63**	.21	–
	$\text{FC}_{\text{Expert}}$	.95	**.96**	.95	.93
EA	EA	**.97**	**.97**	.95	–
Coherence	IntraNLI	.95	<u>**.96**</u>	.95	.93
	BARTScore	**-2.96**	-3.10	-3.09	–
Fluency	$\text{PPL}_{\text{Timeline}}$ (↓)	<u>**13.80**</u>	56.33	31.82	–
	$\text{PPL}_{\text{High-level}}$ (↓)	9.32	**9.30**	9.45	11.38

Table 2: Automatic evaluation for salient meaning preservation (SMP), factual consistency (FC), evidence appropriateness (EA), coherence, and fluency. Higher is better, except for PPL. BARTScore uses log likelihood, hence higher (less negative) is better. Best in **bold**, significant improvement over second-best <u>underlined</u>.

Aspect	LLaMA	TH-VAE	Naive
Factual Consistency	3.08	**3.35**	3.28
Usefulness (General)	**3.38**	3.28	2.55
(Diagnosis)	**3.40**	3.25	2.93
(Inter-& Intrapersonal)	**3.48**	3.33	2.23
(MoC)	3.30	**3.35**	1.18

Table 3: Human evaluation results based on 5-point Likert scales (1 is worst, 5 is best). Best in **bold**.

For human evaluation and other details, please refer to the paper.

Results

Aspect	Metric	LLaMA	TH-VAE	Skeleton	Naive
SMP	MHIC_{sem}	.65	**.66**	.57	–
FC	$\text{FC}_{\text{Timeline}}$	**.63**	**.63**	.21	–
	$\text{FC}_{\text{Expert}}$	.95	**.96**	.95	.93
EA	EA	**.97**	**.97**	.95	–
Coherence	IntraNLI	.95	<u>**.96**</u>	.95	.93
	BARTScore	**-2.96**	-3.10	-3.09	–
Fluency	$\text{PPL}_{\text{Timeline}}$ ($\downarrow$)	<u>**13.80**</u>	56.33	31.82	–
	$\text{PPL}_{\text{High-level}}$ ($\downarrow$)	9.32	**9.30**	9.45	11.38

Table 2: Automatic evaluation for salient meaning preservation (SMP), factual consistency (FC), evidence appropriateness (EA), coherence, and fluency. Higher is better, except for PPL. BARTScore uses log likelihood, hence higher (less negative) is better. Best in **bold**, significant improvement over second-best <u>underlined</u>.

Main observations and takeaways:

1. TH-VAE generates **factually consistent**, **coherent**, and **mental health information preserving** summaries of user timelines.
2. LLaMA-based summaries excel in **fluency** and are comparable to TH-VAE in capturing mental health information, but falls short in **logical coherence**.
3. Human judges preferred LLaMA summaries for usefulness, but preferred TH-VAE in terms of **factual consistency** and usefulness in **capturing changes over time.**
4. Prompting enables generation of fluent mental health insights grounded in summarised timeline evidence.
    - Importance of high-quality expert examples in few-shot prompting for keyphrases

For human evaluation and other details, please refer to the paper.

Conclusions

- New task: mental health summaries from users' social media timelines.
  - Two-layer summaries combining (1) high-level information and (2) summarised timeline content.
  - Evaluation methods addressing unique requirements for this task.
  - Gold standard mental health summaries of user timelines.
- Novel timeline summarisation method: TH-VAE
  - We proposed a timeline summarisation system based on a hierarchical VAE for long texts.
  - Method captures long dependencies between sub-timelines.
  - Summaries most logically coherent, factually consistent, and useful for capturing changes.
- Prompting framework for generating high-level mental health summaries.

Ongoing work

- Injection of temporality and context into PLMs
- Personalisation of PLMs and particularly longitudinal models
- Work on multi-modal longitudinal content
- Summarisation of longitudinal sequences
- Explanations of longitudinal models
- Synthetic generation and evaluation of user timelines
- Evaluation of generation focussing on privacy and factuality

Research Team

Longitudinal models & Dynamic Representations


A. Tsakalidis


F.Nanni

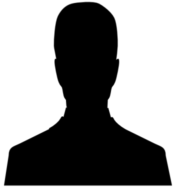

J. Chim


A. Hills


T. Tseriotou

Rumour Verification

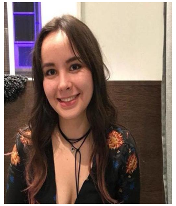

E. Kochkina

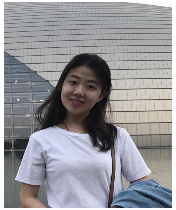

J. Dougrez-Lewis

Synthetic Language Data Generation


M. Liakata

Opinion Summarisation


I.Bilal


J.Song


D. Gkoumas

Multi-modal data for mental health diagnoses

Cross-Domain Semantics & QA


J.Ravenscroft


M. Maufe

Thank you!

Addressing sociotechnical limitations of LLMs in medical and social contexts

https://adsolve.github.io/ – Hiring PhDs and postdocs!

Queen Mary University of London · University of Nottingham UK | CHINA | MALAYSIA · University of Sheffield · WARWICK THE UNIVERSITY OF WARWICK · RAi