universität innsbruck

Continuing Topic Explanation

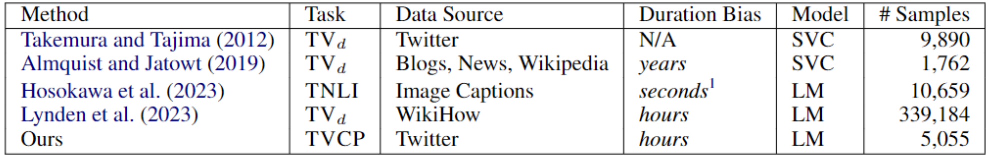
Adam Jatowt

16 Nov 2023

Topics

Topic 1: Temporal commonsense reasoning about actions

- Temporal commonsense reasoning (either task):
  1. Temporal validity change prediction
     - Dataset:
       - >5k sentence pairs with their crowdsourced labels
  2. Stationarity prediction
     - Dataset:
       - >1.5k sentences with stational, non-stational labels
  - + other datasets that can be of help:



Method	Task	Data Source	Duration Bias	Model	# Samples
Takemura and Tajima (2012)	TV_d	Twitter	N/A	SVC	9,890
Almquist and Jatowt (2019)	TV_d	Blogs, News, Wikipedia	*years*	SVC	1,762
Hosokawa et al. (2023)	TNLI	Image Captions	*seconds*[1]	LM	10,659
Lynden et al. (2023)	TV_d	WikiHow	*hours*	LM	339,184
Ours	TVCP	Twitter	*hours*	LM	5,055

Table 1: Summary of related work

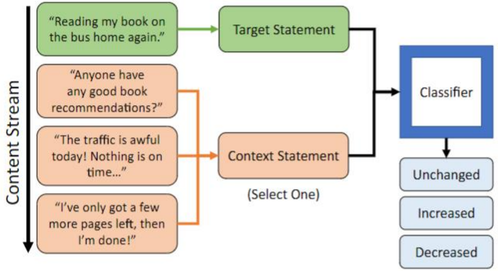

Content Stream
"Reading my book on the bus home again."
"Anyone have any good book recommendations?"
"The traffic is awful today! Nothing is on time..."
"I've only got a few more pages left, then I'm done!"
Target Statement
Context Statement
(Select One)
Classifier
Unchanged
Increased
Decreased

Figure 1: A visualization of the TVCP task

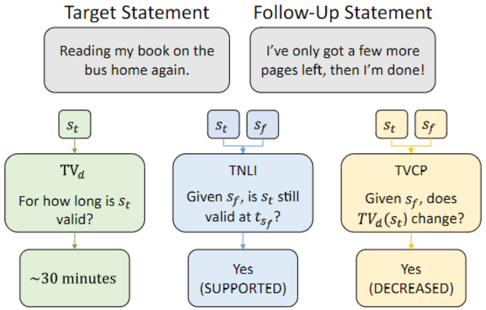
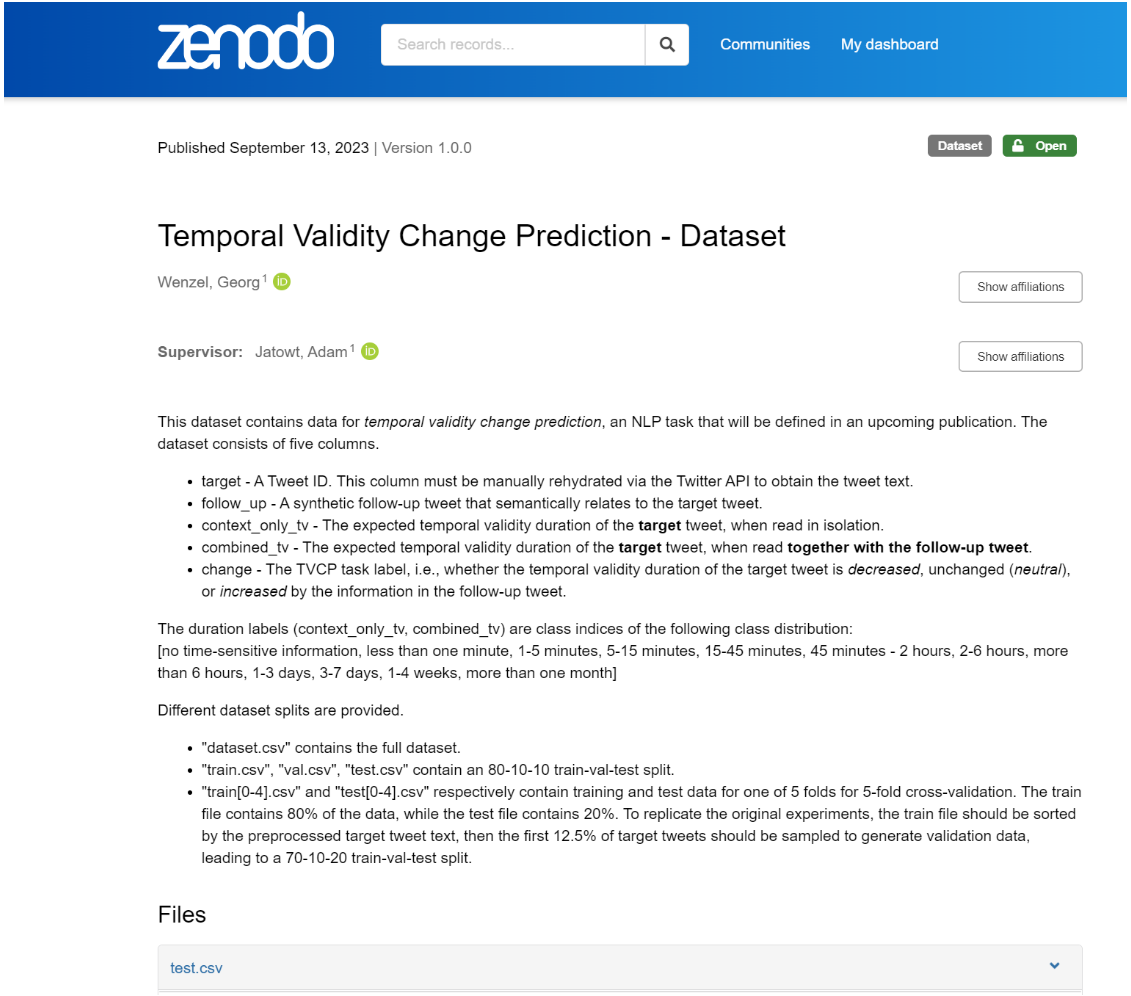

Target Statement
Follow-Up Statement
Reading my book on the bus home again.
I've only got a few more pages left, then I'm done!
s_t
s_t s_f
s_t s_f
TV_d
For how long is s_t valid?
TNLI
Given s_f, is s_t still valid at t_{s_f}?
TVCP
Given s_f, does $TV_d(s_t)$ change?
~30 minutes
Yes (SUPPORTED)
Yes (DECREASED)

Figure 3: An example of TV_d, TNLI and TVCP

G. Wenzel, A. Jatowt, *Temporal validity change prediction* (under review at ACL2024)
G. Wenzel, A. Jatowt, *An Overview Of Temporal Commonsense Reasoning and Acquisition*: https://arxiv.org/abs/2308.00002



Published September 13, 2023 | Version 1.0.0

Temporal Validity Change Prediction - Dataset

Wenzel, Georg[1]

Supervisor: Jatowt, Adam[1]

This dataset contains data for *temporal validity change prediction*, an NLP task that will be defined in an upcoming publication. The dataset consists of five columns.

- target - A Tweet ID. This column must be manually rehydrated via the Twitter API to obtain the tweet text.
- follow_up - A synthetic follow-up tweet that semantically relates to the target tweet.
- context_only_tv - The expected temporal validity duration of the **target** tweet, when read in isolation.
- combined_tv - The expected temporal validity duration of the **target** tweet, when read **together with the follow-up tweet**.
- change - The TVCP task label, i.e., whether the temporal validity duration of the target tweet is *decreased*, unchanged (*neutral*), or *increased* by the information in the follow-up tweet.

The duration labels (context_only_tv, combined_tv) are class indices of the following class distribution:
[no time-sensitive information, less than one minute, 1-5 minutes, 5-15 minutes, 15-45 minutes, 45 minutes - 2 hours, 2-6 hours, more than 6 hours, 1-3 days, 3-7 days, 1-4 weeks, more than one month]

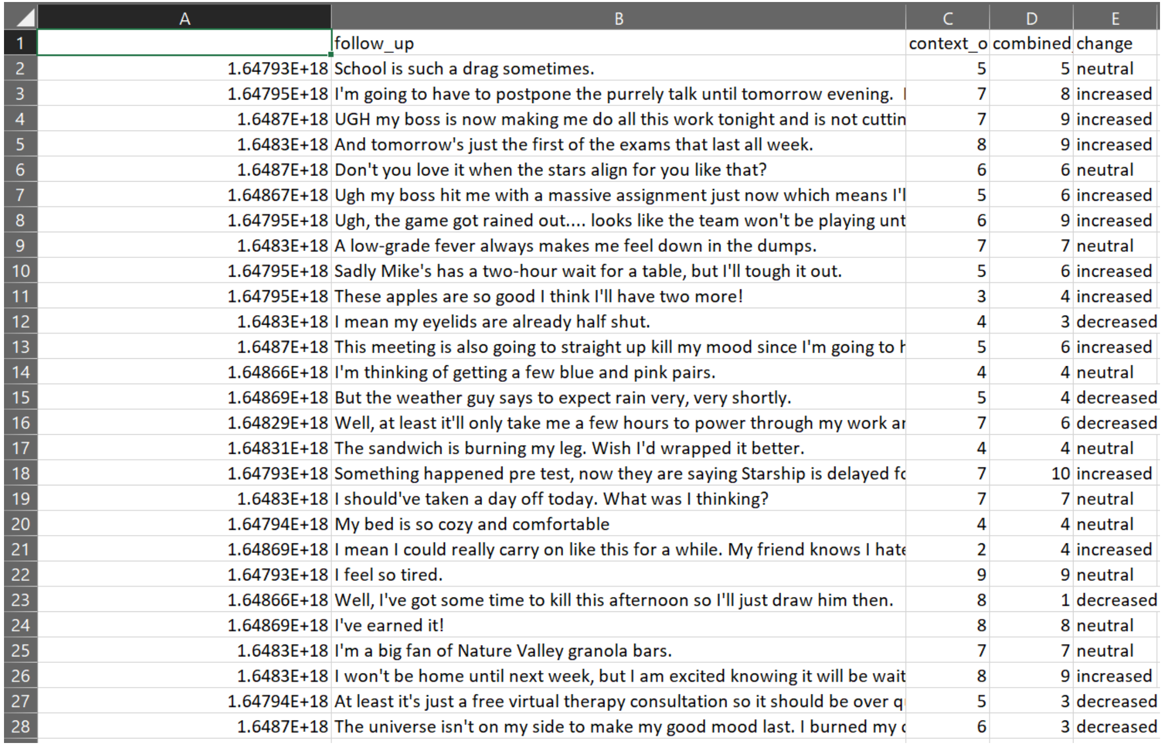
Different dataset splits are provided.

- "dataset.csv" contains the full dataset.
- "train.csv", "val.csv", "test.csv" contain an 80-10-10 train-val-test split.
- "train[0-4].csv" and "test[0-4].csv" respectively contain training and test data for one of 5 folds for 5-fold cross-validation. The train file contains 80% of the data, while the test file contains 20%. To replicate the original experiments, the train file should be sorted by the preprocessed target tweet text, then the first 12.5% of target tweets should be sampled to generate validation data, leading to a 70-10-20 train-val-test split.

Files

test.csv

Temporal Validity Change Estimation Dataset



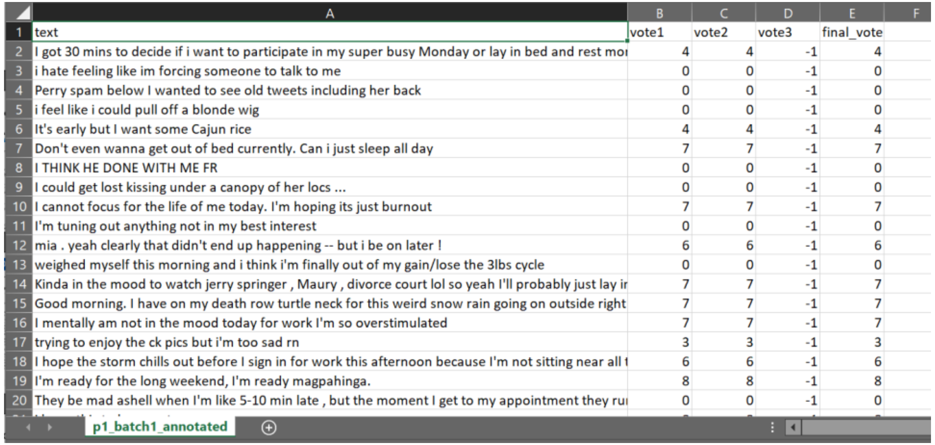
	A	B	C	D	E
1		follow_up	context_o	combined	change
2	1.64793E+18	School is such a drag sometimes.	5	5	neutral
3	1.64795E+18	I'm going to have to postpone the purrely talk until tomorrow evening. I	7	8	increased
4	1.6487E+18	UGH my boss is now making me do all this work tonight and is not cuttin	7	9	increased
5	1.6483E+18	And tomorrow's just the first of the exams that last all week.	8	9	increased
6	1.6487E+18	Don't you love it when the stars align for you like that?	6	6	neutral
7	1.64867E+18	Ugh my boss hit me with a massive assignment just now which means I'l	5	6	increased
8	1.64795E+18	Ugh, the game got rained out.... looks like the team won't be playing unt	6	9	increased
9	1.6483E+18	A low-grade fever always makes me feel down in the dumps.	7	7	neutral
10	1.64795E+18	Sadly Mike's has a two-hour wait for a table, but I'll tough it out.	5	6	increased
11	1.64795E+18	These apples are so good I think I'll have two more!	3	4	increased
12	1.6483E+18	I mean my eyelids are already half shut.	4	3	decreased
13	1.6487E+18	This meeting is also going to straight up kill my mood since I'm going to h	5	6	increased
14	1.64866E+18	I'm thinking of getting a few blue and pink pairs.	4	4	neutral
15	1.64869E+18	But the weather guy says to expect rain very, very shortly.	5	4	decreased
16	1.64829E+18	Well, at least it'll only take me a few hours to power through my work ar	7	6	decreased
17	1.64831E+18	The sandwich is burning my leg. Wish I'd wrapped it better.	4	4	neutral
18	1.64793E+18	Something happened pre test, now they are saying Starship is delayed fo	7	10	increased
19	1.6483E+18	I should've taken a day off today. What was I thinking?	7	7	neutral
20	1.64794E+18	My bed is so cozy and comfortable	4	4	neutral
21	1.64869E+18	I mean I could really carry on like this for a while. My friend knows I hate	2	4	increased
22	1.64793E+18	I feel so tired.	9	9	neutral
23	1.64866E+18	Well, I've got some time to kill this afternoon so I'll just draw him then.	8	1	decreased
24	1.64869E+18	I've earned it!	8	8	neutral
25	1.6483E+18	I'm a big fan of Nature Valley granola bars.	7	7	neutral
26	1.6483E+18	I won't be home until next week, but I am excited knowing it will be wait	8	9	increased
27	1.64794E+18	At least it's just a free virtual therapy consultation so it should be over q	5	3	decreased
28	1.6487E+18	The universe isn't on my side to make my good mood last. I burned my c	6	3	decreased

Stationarity Prediction Dataset

Stationarity estimation

- Data in a .zip file (available in OLAT)
- Data contains the annotations from the other task (temporal validity estimation) but we can reuse them here for the binary task of stationarity estimation.
- vote1 and vote2 are provided by the crowdworkers. vote3 is supplied by a student when there was no agreement between the crowdworkers (otherwise -1).
- The final_vote is the majority vote if one exists (otherwise -1).

- Class 0 denotes "no time-sensitive information" (i.e., this should usually be stationary or hypothetical information, although in some cases it may also be possible that information is contained in the past). One can nevertheless treat it as stationary data. There should be ~1500 such statements that were determined to be stationary.
- Statements of any other class would mean time-changing (time-sensitive) information.



	A	B	C	D	E	F
1	text	vote1	vote2	vote3	final_vote	
2	I got 30 mins to decide if i want to participate in my super busy Monday or lay in bed and rest mo	4	4	-1	4	
3	i hate feeling like im forcing someone to talk to me	0	0	-1	0	
4	Perry spam below I wanted to see old tweets including her back	0	0	-1	0	
5	i feel like i could pull off a blonde wig	0	0	-1	0	
6	It's early but I want some Cajun rice	4	4	-1	4	
7	Don't even wanna get out of bed currently. Can i just sleep all day	7	7	-1	7	
8	I THINK HE DONE WITH ME FR	0	0	-1	0	
9	I could get lost kissing under a canopy of her locs ...	0	0	-1	0	
10	I cannot focus for the life of me today. I'm hoping its just burnout	7	7	-1	7	
11	I'm tuning out anything not in my best interest	0	0	-1	0	
12	mia . yeah clearly that didn't end up happening -- but i be on later !	6	6	-1	6	
13	weighed myself this morning and i think i'm finally out of my gain/lose the 3lbs cycle	0	0	-1	0	
14	Kinda in the mood to watch jerry springer , Maury , divorce court lol so yeah I'll probably just lay i	7	7	-1	7	
15	Good morning. I have on my death row turtle neck for this weird snow rain going on outside right	7	7	-1	7	
16	I mentally am not in the mood today for work I'm so overstimulated	7	7	-1	7	
17	trying to enjoy the ck pics but i'm too sad rn	3	3	-1	3	
18	I hope the storm chills out before I sign in for work this afternoon because I'm not sitting near all	6	6	-1	6	
19	I'm ready for the long weekend, I'm ready magpahinga.	8	8	-1	8	
20	They be mad ashell when I'm like 5-10 min late , but the moment I get to my appointment they ru	0	0	-1	0	

Potential Ideas

- Multi-tasking setting so that training on one data enhances training on another data
  - Train classifier (e.g., RoBERTa, BERT based classifiers) for both the tasks at the same time
  - Investigate different multi-tasking combinations?

Potential Ideas when using LLMs

- N different LLMs (N={2,3,4}?) with the selection based on:
  - Size
  - Character of data used for training
- Testing:
  - 0-shot learning case
  - Few-shot learning case
  - Role-setting prompt
  - Any other prompt engineering idea?
  - Fine-tunning LLM
  - Investigate different parameter values
  - Two-stage prompt: first ask to estimate temporal validity and then to determine its change

Resources

- Hugging Face
- Langchain

- Exchange contacts between each other
- Set-up a communication channel and periodical discussion meetings