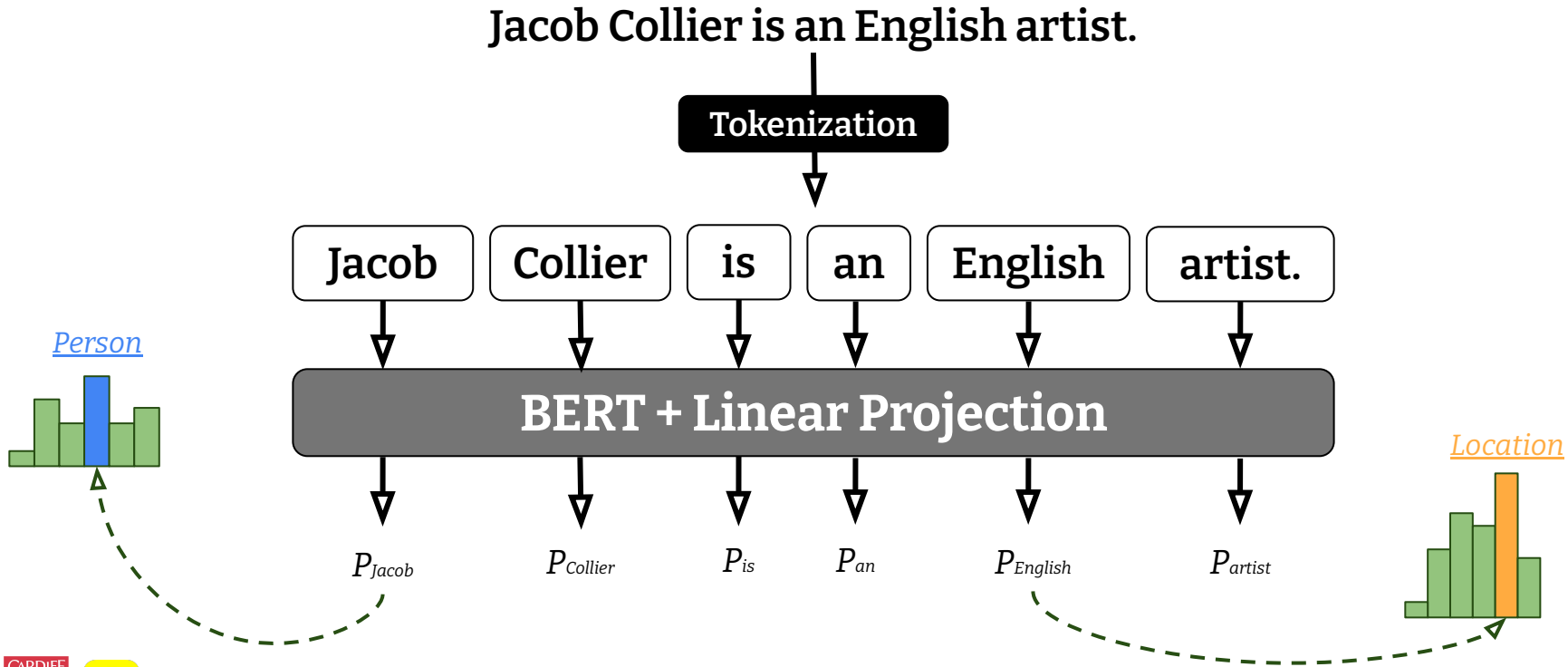
Named Entity Recognition in Twitter: A Dataset and Analysis on Short-Term Temporal Shifts

Asahi Ushio, *Leonardo Neves, Vitor Silva, Francesco Barbieri, Jose Camacho-Collados*

Computer Science & Informatics, Cardiff University, Cardiff NLP
Snap Inc.,

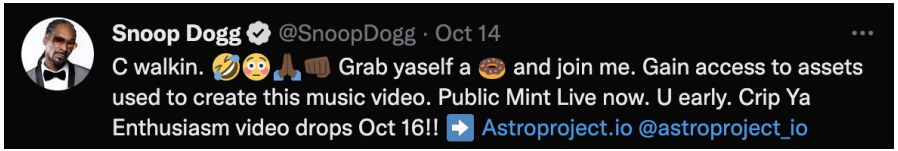
https://huggingface.co/datasets/tner/tweetner7

Named-Entity Recognition (NER)



NER on Social Media

- Noisy Text
  - Less formal.
  - Different vocabulary.
  - Emoji, urls, etc.



- **Temporal Shift**
  - The meaning of words is constantly changing or evolving over time.
  - New entities in new period.

https://huggingface.co/datasets/tner/tweetner7

TweetNER7

TweetNER7 is a NER dataset on Twitter with 7 entity types.

- Person, Location, Corporation, Creative work, Group, Product.
- Tweets are collected from Sep 2019 to Aug 2021.
  - 2020-set: Sep 2019 ~ Aug 2020 (5,768 tweets)
  - 2021-set: Sep 2020 ~ Aug 2021 (5,612 tweets)
- Temporal shift setup
  - Training/Validation: 2020-set
  - Test: 2021-set

Other Twitter NER Datasets

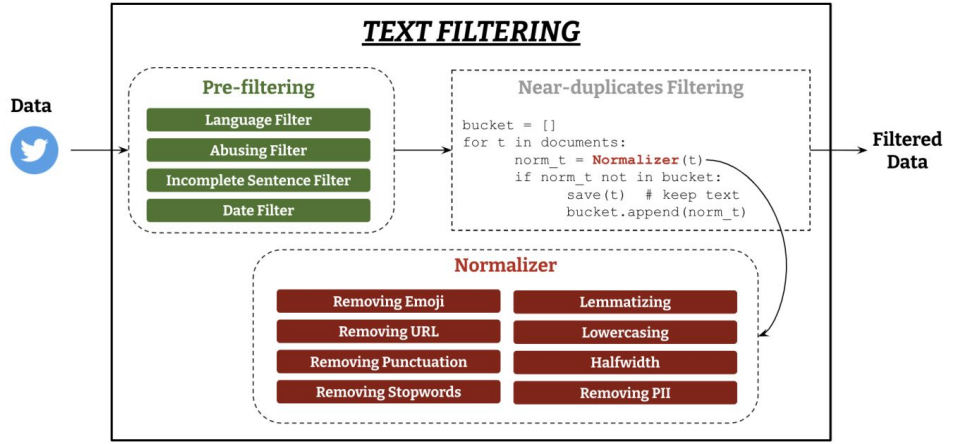
TweetNER7 is unique for

- Uniform distribution over months
  - 2,000 tweets in each month
- (Relatively) short term temporal shift
  - Over 2 years
- Large annotation (data size & entities)
  - More than 10k annotated tweets with 7 entities

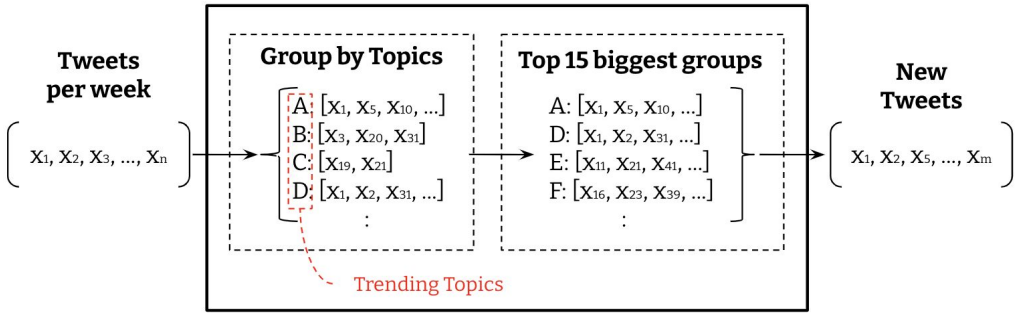
Dataset Collection

TweetTopic tweet collection:

- Query 50 tweets every two hours from September 2019 to October 2021
- Pre-filtering & Near de-duplication
- Trend-filter



Tweet collection pipeline.



Trend-filter

Preprocessing Tweets

- Special token for URLs.
- Special representation for usernames.

Raw tweet.

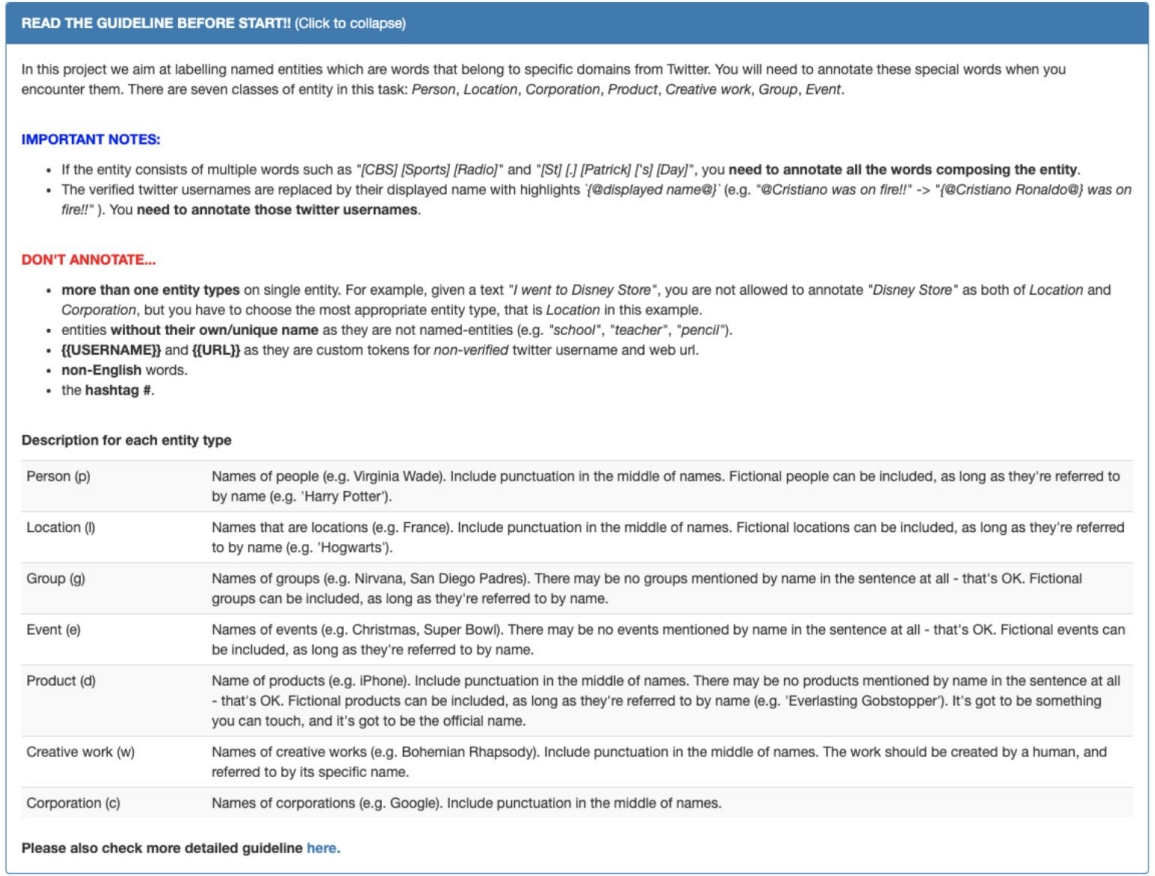
```
Get the all-analog Classic Vinyl Edition
of "Takin' Off" Album from @herbiehancock
via @bluenoterecords link below:
http://bluenote.lnk.to/AlbumOfTheWeek
```

Processed tweet.

```
Get the all-analog Classic Vinyl Edition
of "Takin' Off" Album from {@herbiehancock@}
via {@bluenoterecords@} link below: {{URL}}
```

Dataset Annotation

- Mechanical Turk:
  - 3 annotators per tweet
- Agreement
  - 1 / 3: Disregard
  - 2 / 3: Manual check



READ THE GUIDELINE BEFORE START!! (Click to collapse)

In this project we aim at labelling named entities which are words that belong to specific domains from Twitter. You will need to annotate these special words when you encounter them. There are seven classes of entity in this task: *Person*, *Location*, *Corporation*, *Product*, *Creative work*, *Group*, *Event*.

IMPORTANT NOTES:

- If the entity consists of multiple words such as "*[CBS] [Sports] [Radio]*" and "*[St] [.] [Patrick] ['s] [Day]*", you **need to annotate all the words composing the entity**.
- The verified twitter usernames are replaced by their displayed name with highlights '*{@displayed name@}*' (e.g. "*@Cristiano was on fire!!*" -> "*{@Cristiano Ronaldo@} was on fire!!*"). You **need to annotate those twitter usernames**.

DON'T ANNOTATE...

- **more than one entity types** on single entity. For example, given a text "*I went to Disney Store*", you are not allowed to annotate "*Disney Store*" as both of *Location* and *Corporation*, but you have to choose the most appropriate entity type, that is *Location* in this example.
- entities **without their own/unique name** as they are not named-entities (e.g. "*school*", "*teacher*", "*pencil*").
- **{{USERNAME}}** and **{{URL}}** as they are custom tokens for *non-verified* twitter username and web url.
- **non-English** words.
- the **hashtag #**.

Description for each entity type

Person (p)	Names of people (e.g. Virginia Wade). Include punctuation in the middle of names. Fictional people can be included, as long as they're referred to by name (e.g. 'Harry Potter').
Location (l)	Names that are locations (e.g. France). Include punctuation in the middle of names. Fictional locations can be included, as long as they're referred to by name (e.g. 'Hogwarts').
Group (g)	Names of groups (e.g. Nirvana, San Diego Padres). There may be no groups mentioned by name in the sentence at all - that's OK. Fictional groups can be included, as long as they're referred to by name.
Event (e)	Names of events (e.g. Christmas, Super Bowl). There may be no events mentioned by name in the sentence at all - that's OK. Fictional events can be included, as long as they're referred to by name.
Product (d)	Name of products (e.g. iPhone). Include punctuation in the middle of names. There may be no products mentioned by name in the sentence at all - that's OK. Fictional products can be included, as long as they're referred to by name (e.g. 'Everlasting Gobstopper'). It's got to be something you can touch, and it's got to be the official name.
Creative work (w)	Names of creative works (e.g. Bohemian Rhapsody). Include punctuation in the middle of names. The work should be created by a human, and referred to by its specific name.
Corporation (c)	Names of corporations (e.g. Google). Include punctuation in the middle of names.

Please also check more detailed guideline here.

Experiment: Temporal Shift Setup

Setup: Train & validate on 2020-set & test on 2021-set

Metric: Micro F1 / Macro F1

Result:

- RoBERTa is the best in 2021.
- BERTweet is the best in 2020.
- Performance: 2021 > 2020.

Model	Micro F1 (2021/2020)	Macro F1 (2021/2020)
BERTweet Base	64.1 / **66.4**	59.4 / 62.4
BERTweet Large	64.0 / 65.9	59.5 / **62.6**
RoBERTa Base	64.2 / 64.2	59.1 / 60.2
RoBERTa Large	**64.8** / 65.7	**60.0** / 61.9
TimeLM2019	64.3 / 65.4	59.3 / 61.1
TimeLM2020	62.9 / 64.4	58.3 / 60.3
TimeLM2021	64.2 / 65.4	59.5 / 61.1

Results Breakdown

Easy entities:

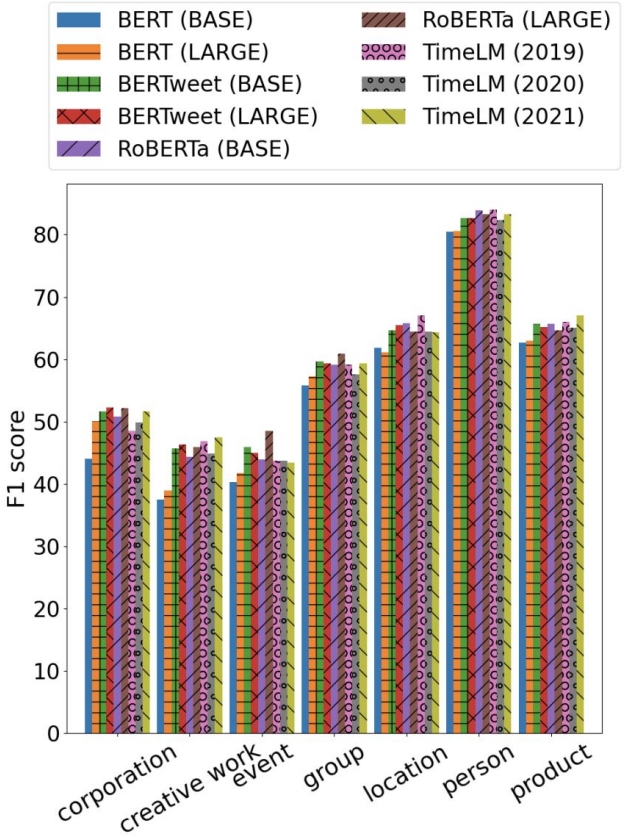
Person, product, location

- Large number of entities
- Low diversity

Challenging entities:

Creative work, event

- Small number of entities
- High diversity



Experiment: Continuous Fine-tuning

Setup: Use both of 2020 & 2021 training set by concatenation or continuous fine-tuning.

Result:

- Continuous fine-tuning achieves the best results.
- TimeLM19 is better than TimeLM20 and TimeLM21.

Model	Dataset	Micro F1 (2021/2020)	Macro F1 (2021/2020)
RoBERTa LARGE	2020-set	64.8 / 65.7	60.0 / 61.9
	Continuous	66.0 / 66.3	60.9 / 62.4
TimeLM 2019	2020-set	64.3 / 65.4	59.3 / 61.1
	Continuous	65.9 / 64.8	61.1 / 60.6
TimeLM 2020	2020-set	62.9 / 64.4	58.3 / 60.3
	Continuous	65.5 / 65.3	60.6 / 61.3
TimeLM 2021	2020-set	64.2 / 65.4	59.5 / 61.1
	Continuous	65.1 / 64.9	60.0 / 60.7

Experiment: Self-labeling

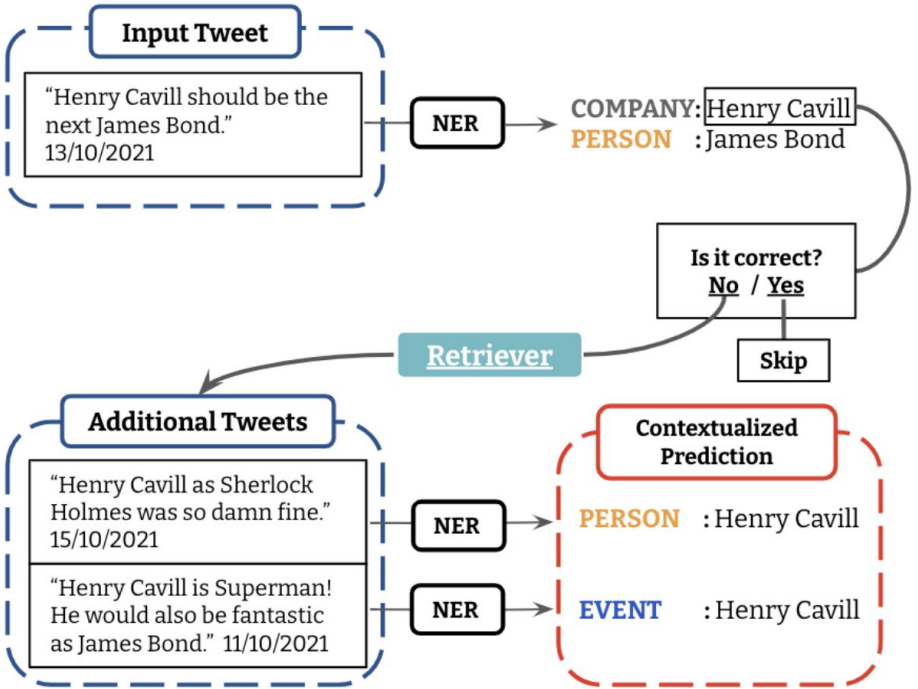
Setup: Collect additional 93,594 and 878,80 tweets from the period of 2020-set and 2021-set. Generate pseudo annotation by fine-tuned RoBERTa Large model.

Dataset	Micro F1 (2021)	Macro F1 (2021)
2020-set	**64.8**	**60**
2020-self-labeling	64.6	59.3
2021-self-labeling	64.2	59.3

Result: No significant improvements.

Are the pseudo labels not useful at all?

Setup: For incorrect prediction, retrieve the pseudo-labels for the entity within the range of 7 days, and take the most frequent prediction as *the contextualized prediction*.

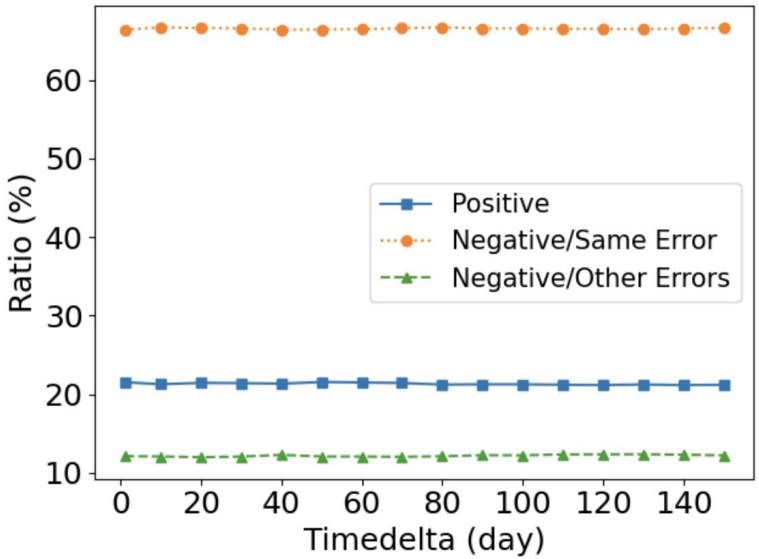


Result of Contextualized Prediction

Contextualized prediction usually matches the original one, meaning is wrong 😩

But, the second most frequent prediction is correct in average 🤔

There is a signal at least, but not easy way to utilize it.



Summary

- ***TweetNER7***, a NER dataset on Twitter with temporal-shift and 7 entity types
- Report LM fine-tuning results
  - Temporal split vs Random split
  - Continuous fine-tuning
- Self-labeling
  - Contextual prediction

Thank you!!