UMassAmherst | Manning College of Information & Computer Sciences

Cross-Dialect Social Media Dependency Parsing for Social Scientific Entity Attribute Analysis

Chloe Eggleston, Brendan O’Connor

Eighth Workshop on Noisy User-generated Text (W-NUT 2022)

Paper available online at github.com/slanglab/TweetIE_WNUT2022

Motivation

- Social media dependency parsing performance has improved significantly with in-domain transformers and corpora.

Motivation

- Social media dependency parsing performance has improved significantly with in-domain transformers and corpora.
- This allows for relevant applications to computational social science, especially when taking full advantage of social elements like user-declared metadata.

Motivation

- Social media dependency parsing performance has improved significantly with in-domain transformers and corpora.
- This allows for relevant applications to computational social science, especially when taking full advantage of social elements like user-declared metadata.
- Our work seeks to synthesize a key application of dependency parsing, information extraction, with this computational social scientific perspective.

Motivation

- Social media dependency parsing performance has improved significantly with in-domain transformers and corpora.
- This allows for relevant applications to computational social science, especially when taking full advantage of social elements like user-declared metadata.
- Our work seeks to synthesize a key application of dependency parsing, information extraction, with this computational social scientific perspective.
- We hope to enable information extraction applications focusing not on fact, but beliefs and opinions held by communities and individuals online.

Contributions

1. State-of-the-art social media syntactic dependency parser

   We achieve SotA performance using a pretrained transformer pretrained on Twitter and biaffine attention.

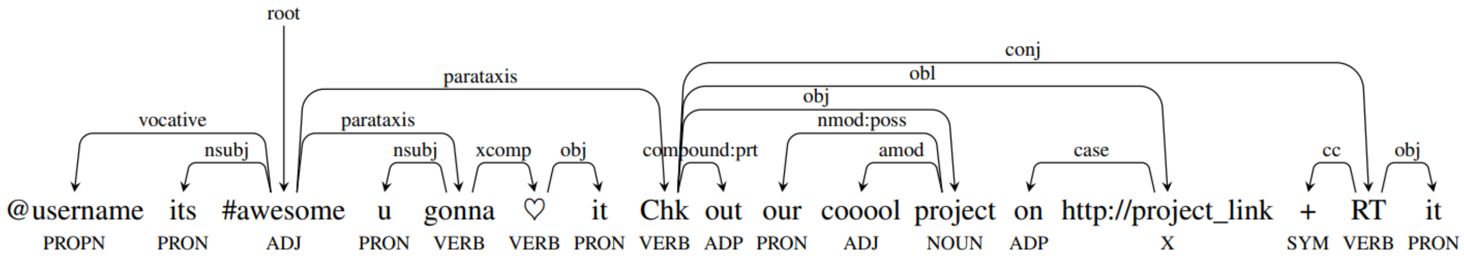


Figure 1: A contrived example of social media syntactic dependencies. Figure 2 from (Liu et al 2018).

Contributions

1. State-of-the-art social media syntactic dependency parser
2. Analysis of its performance with regards to dialectical disparity

   Using Mainstream American English (MAE) and African American English (AAE) segments, we show that our SotA model is not only more performant, but also has less relative error between MAE and AAE than other pretrained models.

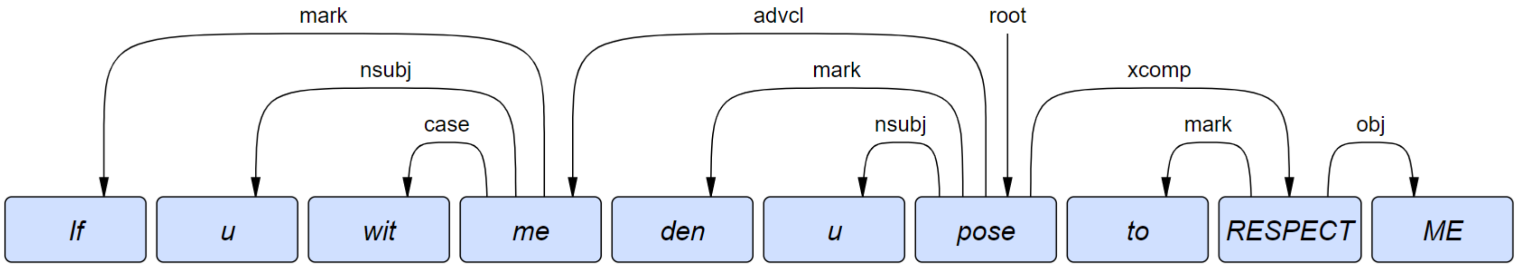


Figure 2: An AAE universal dependencies sample from the TwitterAAE dependencies corpus (Blodgett et al 2018). It features two null copulas and is interpreted as “If you (are) with me, then you (are) supposed to respect me”.

Contributions

1. State-of-the-art social media syntactic dependency parser
2. Analysis of its performance with regards to dialectical disparity
3. Its application – belief extraction with our system TweetIE

   We design a system for extracting beliefs from our dependency parses and compare its application to alternatives using open information extraction.

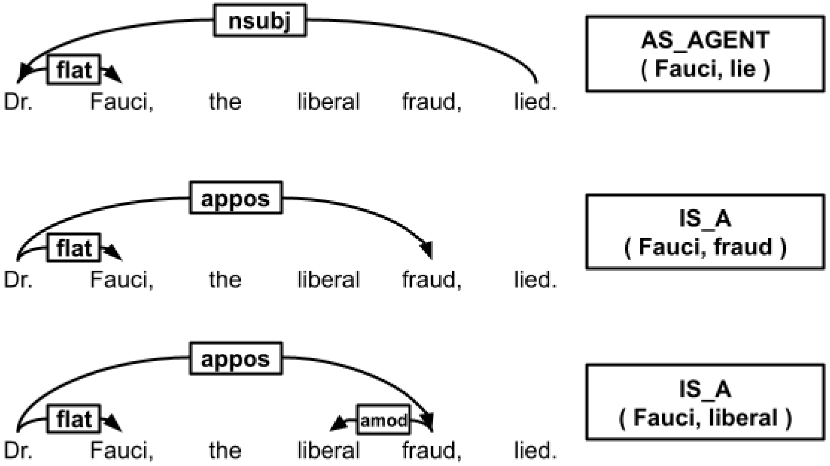


Figure 3: Examples of some of the rules for TweetIE: figure 1 from our paper.

Contributions

1. State-of-the-art social media syntactic dependency parser
2. Analysis of its performance with regards to dialectical disparity
3. Its application – belief extraction with our system TweetIE
4. Case study on COVID-19 political polarization

   A case study using the system above to observe and study beliefs concerning Dr. Anthony Fauci. Replicated studies on public opinion towards him found in sociological literature.

Figure 4: Contrived examples of political discourse towards Dr. Anthony Fauci, inspired from results in the case study.

Dependency Parsing

- We finetune transformer-based dependency parsers on a Twitter dependency corpus (Tweebank v2) using BERTweet, a transformer pretrained entirely on English Twitter.

Dependency Parsing

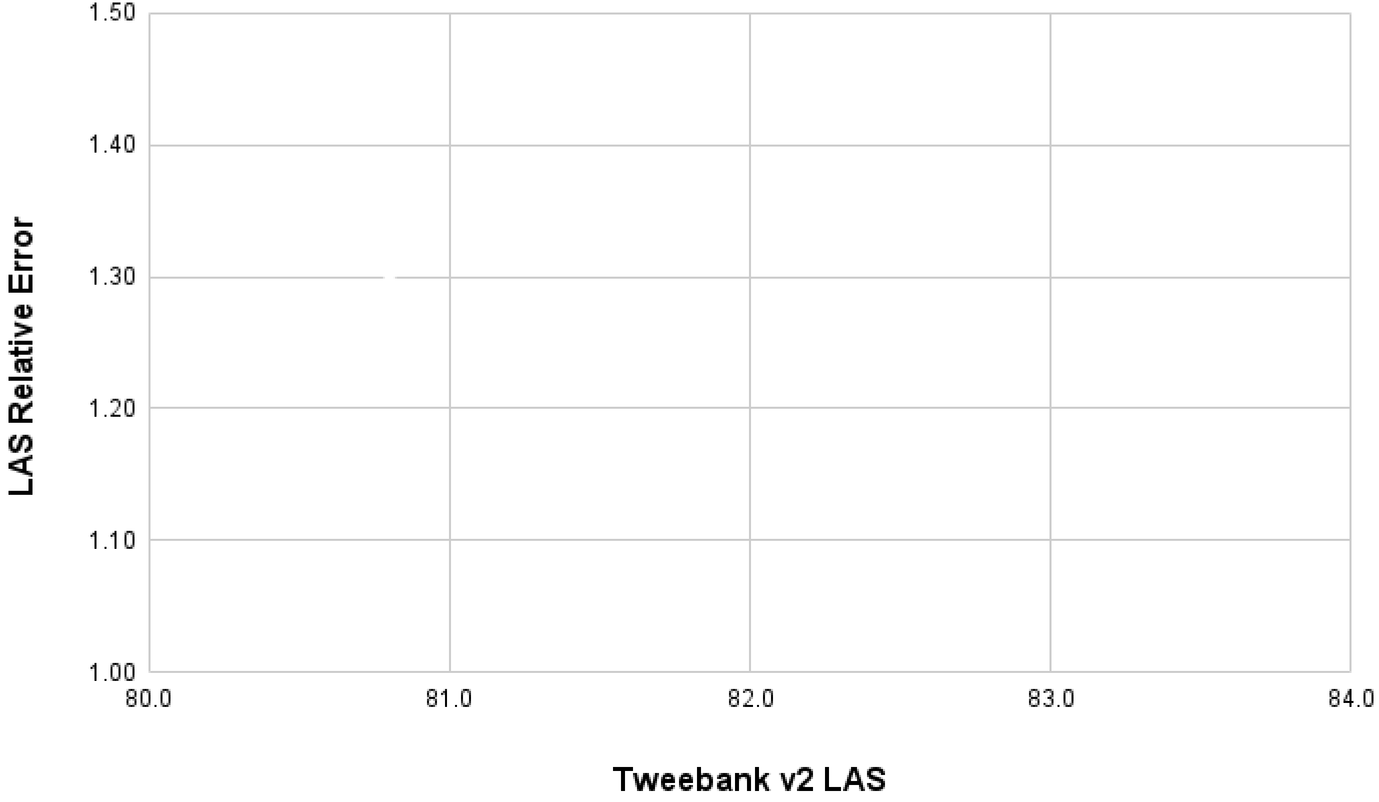
- We finetune transformer-based dependency parsers on a Twitter dependency corpus (Tweebank v2) using BERTweet, a transformer pretrained entirely on English Twitter.
- We achieve state-of-the-art results by 3.4 UAS and 4.0 LAS.

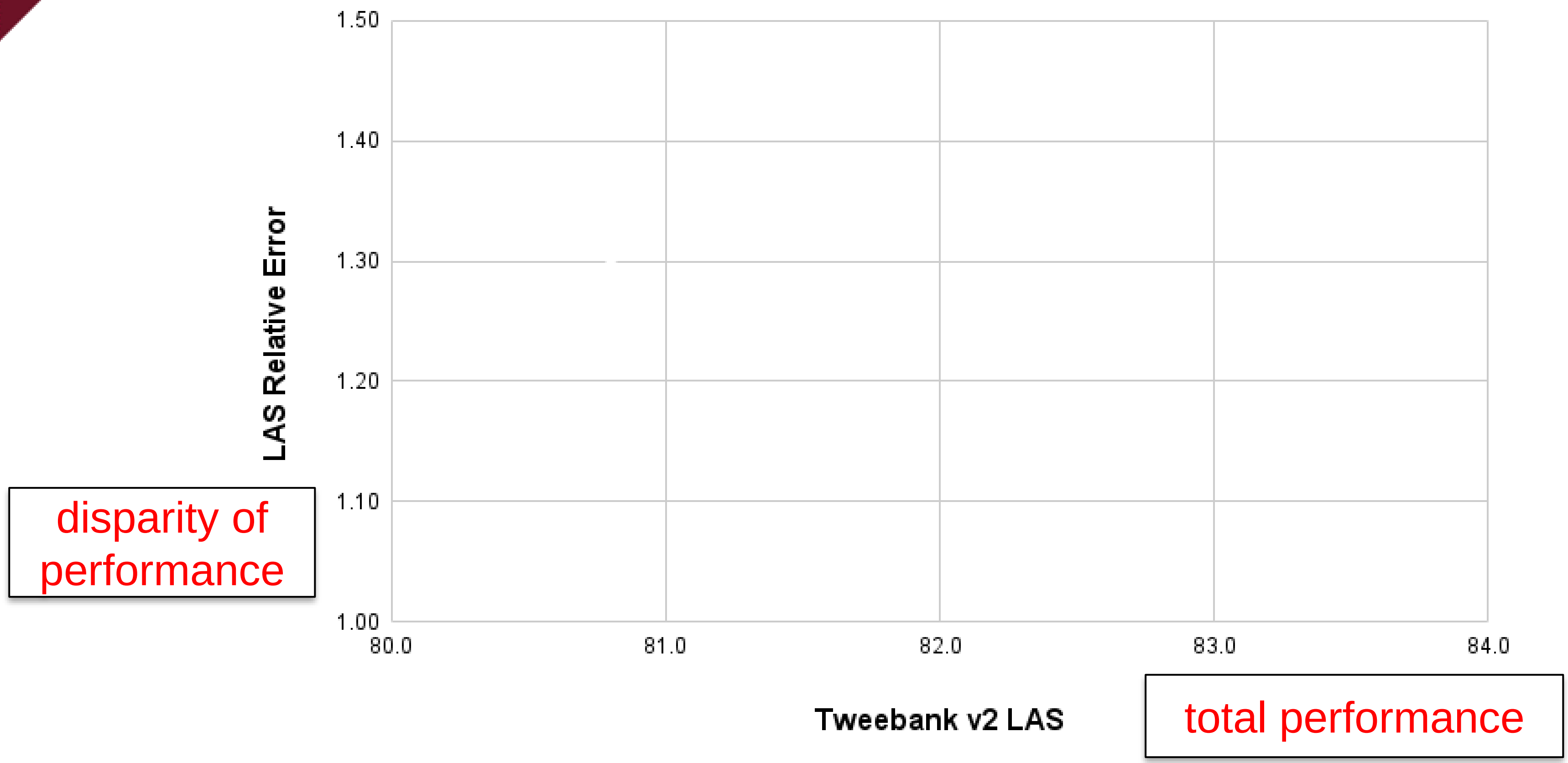
Dependency Parsing

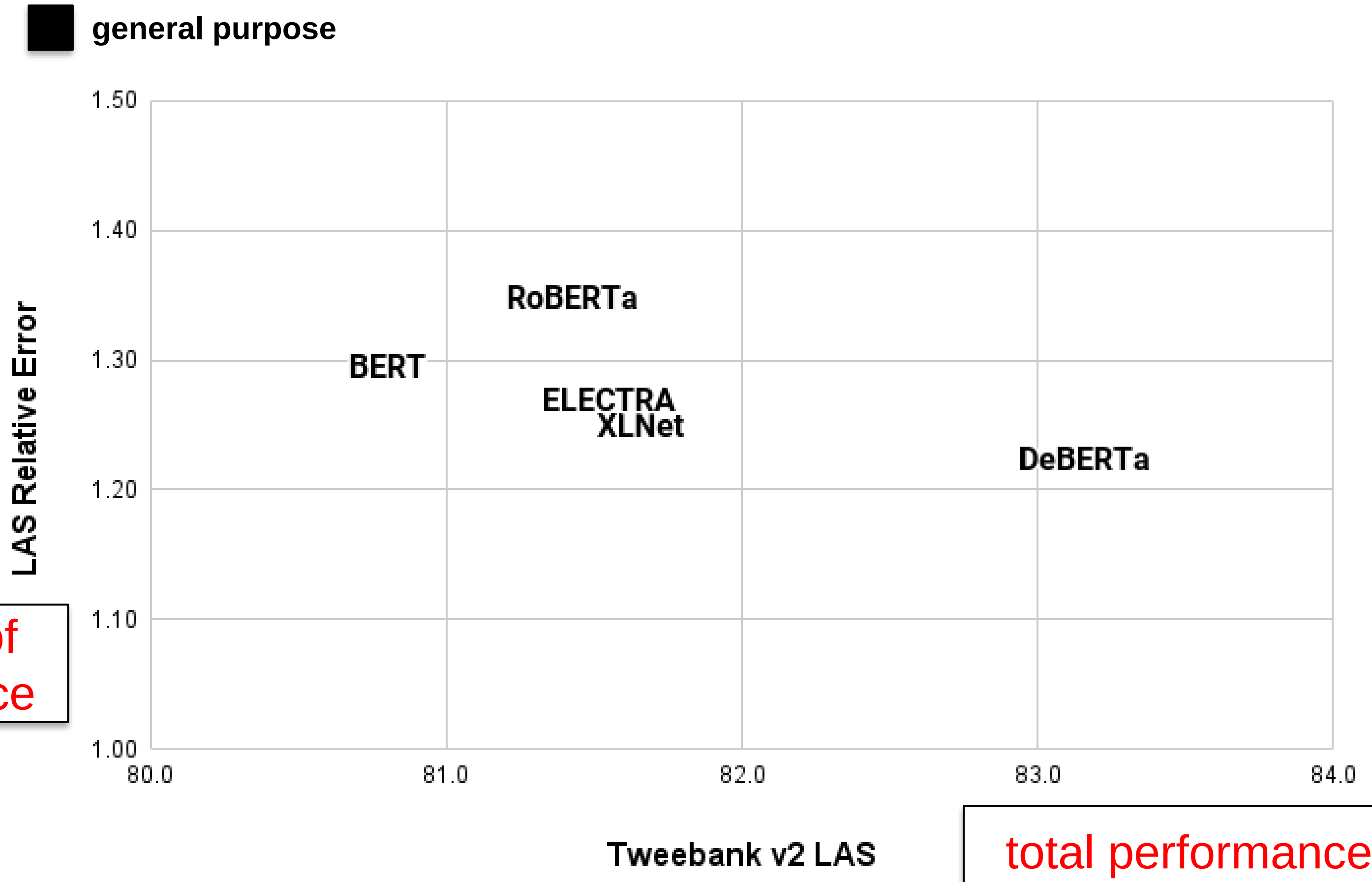
- We finetune transformer-based dependency parsers on a Twitter dependency corpus (Tweebank v2) using BERTweet, a transformer pretrained entirely on English Twitter.
- We achieve state-of-the-art results by 3.4 UAS and 4.0 LAS.
- Interestingly, we observe cross-dialect performance from this model compared to alternatives; this is through subsampling our test set for mainstream English and African American English subsets with a demographic-language model (Blodgett 2016) and comparing the performance across these sets.

Dependency Parsing

- We finetune transformer-based dependency parsers on a Twitter dependency corpus (Tweebank v2) using BERTweet, a transformer pretrained entirely on English Twitter.
- We achieve state-of-the-art results by 3.4 UAS and 4.0 LAS.
- Interestingly, we observe cross-dialect performance from this model compared to alternatives; this is through subsampling our test set for mainstream English and African American English subsets with a demographic-language model (Blodgett 2016) and comparing the performance across these sets.
- This is shown through the *relative error* of the MAE/AAE sets.

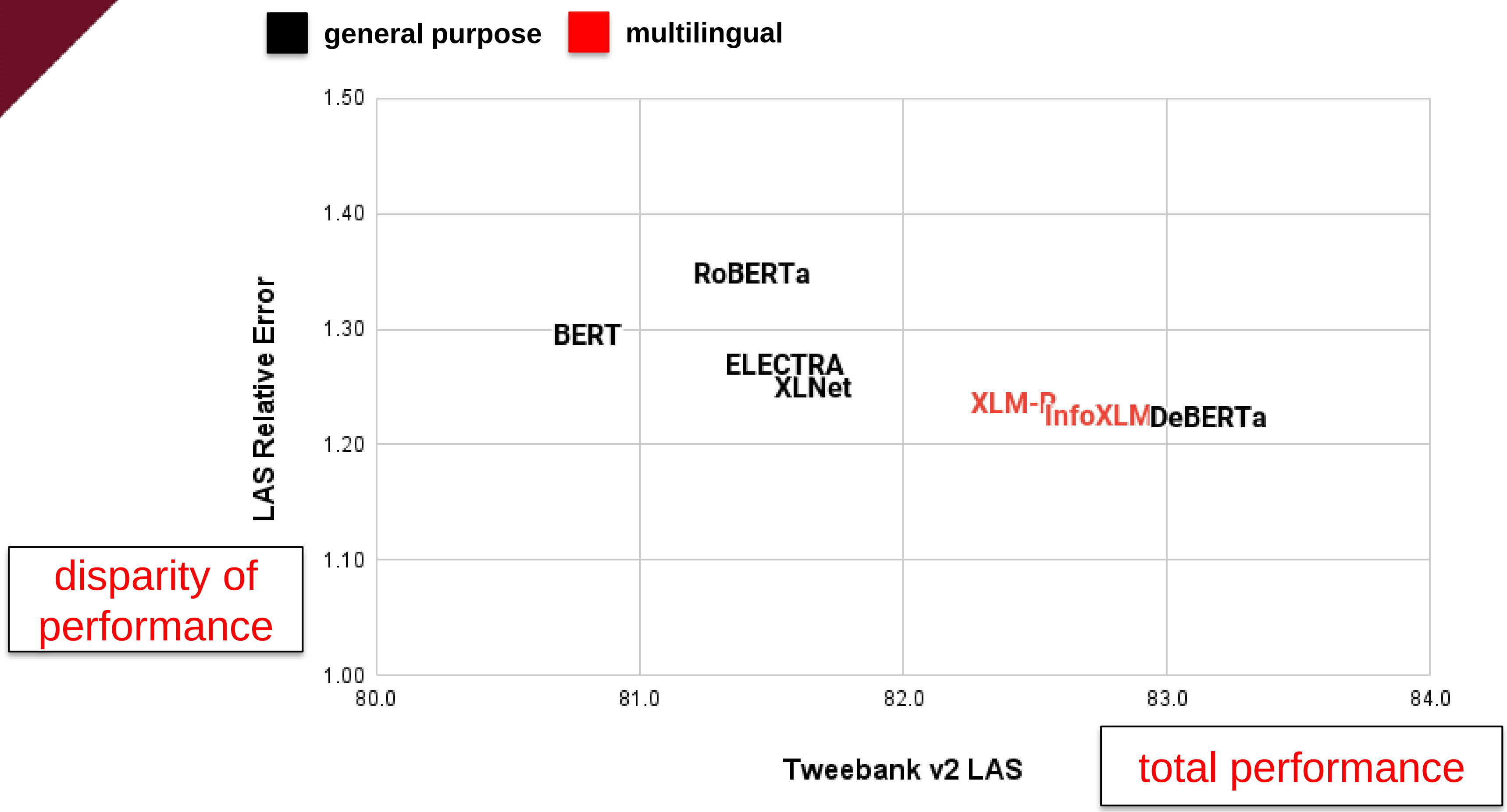

1.50
1.40
1.30
1.20
1.10
1.00
80.0
81.0
82.0
83.0
84.0
LAS Relative Error
Tweebank v2 LAS

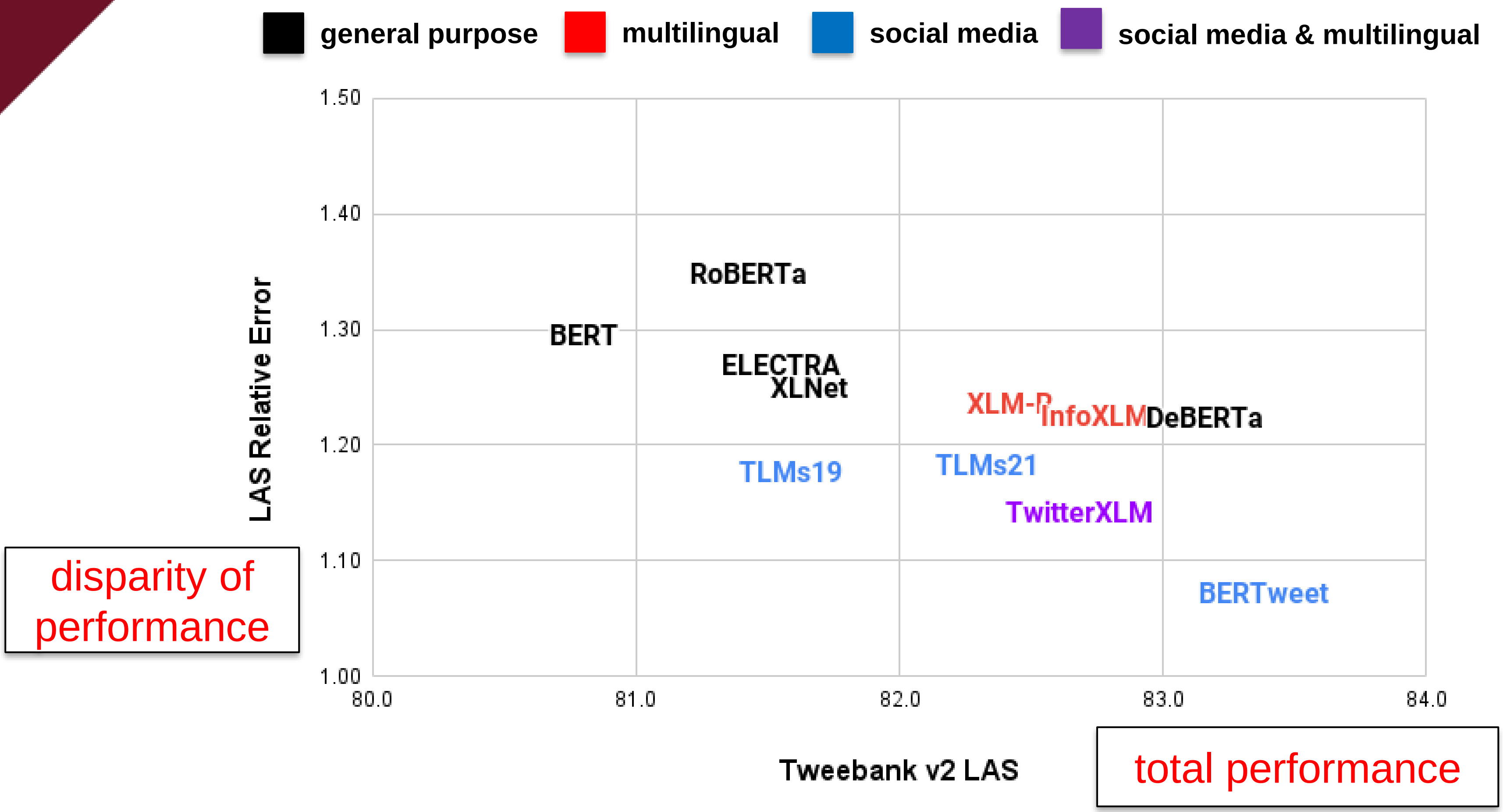

1.50
1.40
1.30
1.20
1.10
1.00
LAS Relative Error
80.0
81.0
82.0
83.0
84.0
Tweebank v2 LAS
disparity of performance
total performance



disparity of performance

total performance


general purpose
multilingual
1.50
1.40
1.30
1.20
1.10
1.00
LAS Relative Error
RoBERTa
BERT
ELECTRA
XLNet
XLM-R
InfoXLM
DeBERTa
80.0
81.0
82.0
83.0
84.0
Tweebank v2 LAS
disparity of performance
total performance



general purpose
multilingual
social media
social media & multilingual
LAS Relative Error
1.50
1.40
1.30
1.20
1.10
1.00
RoBERTa
BERT
ELECTRA
XLNet
XLM-R
InfoXLM
DeBERTa
TLMs19
TLMs21
TwitterXLM
BERTweet
80.0
81.0
82.0
83.0
84.0
Tweebank v2 LAS
disparity of performance
total performance

Belief Extraction using TweetIE

- With a performant and equitable dependency parsing core, we now establish our application of it: a system of belief extraction we call TweetIE. This system extracts relations attributed to specified name entities.

Belief Extraction using TweetIE

- With a performant and equitable dependency parsing core, we now establish our application of it: a system of belief extraction we call TweetIE. This system extracts relations attributed to specified name entities.
- TweetIE infers the following relations using our dependency parses:
  - IS_A: entity is a *what?*
  - HAS_A: entity has a *what*?
  - AS_AGENT: entity does *what?*
  - AS_PATIENT: *what* is done to the entity*?*
  - AS_CONJUNCT: entity and *what* do … / entity and *what* are …

Belief Extraction using TweetIE

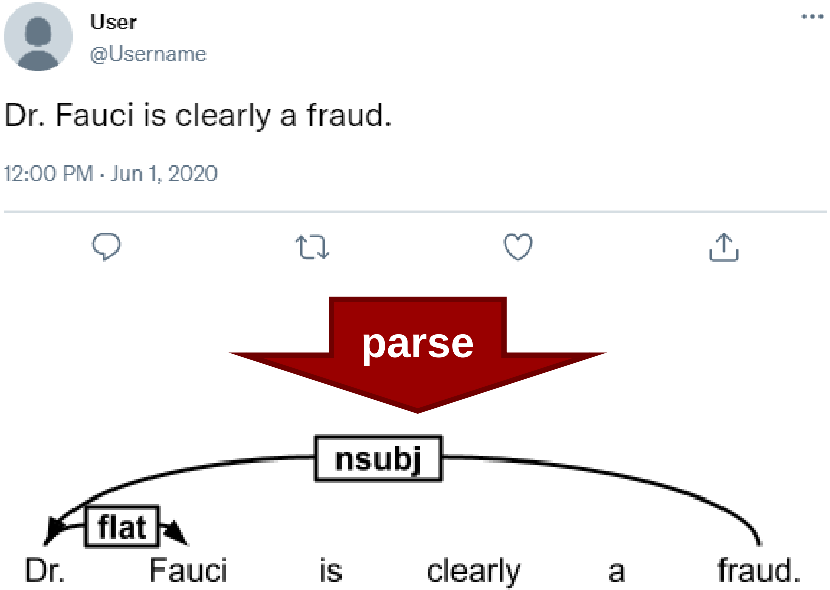
- With a performant and equitable dependency parsing core, we now establish our application of it: a system of belief extraction we call TweetIE. This system extracts relations attributed to specified name entities.
- TweetIE infers the following relations using our dependency parses:
  - IS_A: entity is a *what?*
  - HAS_A: entity has a *what*?
  - AS_AGENT: entity does *what?*
  - AS_PATIENT: *what* is done to the entity?
  - AS_CONJUNCT: entity and *what* do … / entity and *what* are …
- We demonstrate that is has larger yield than alternative systems (OIE frameworks: ReVerb and ClausIE), and perform a precision evaluation, showing that it is more precise than both.

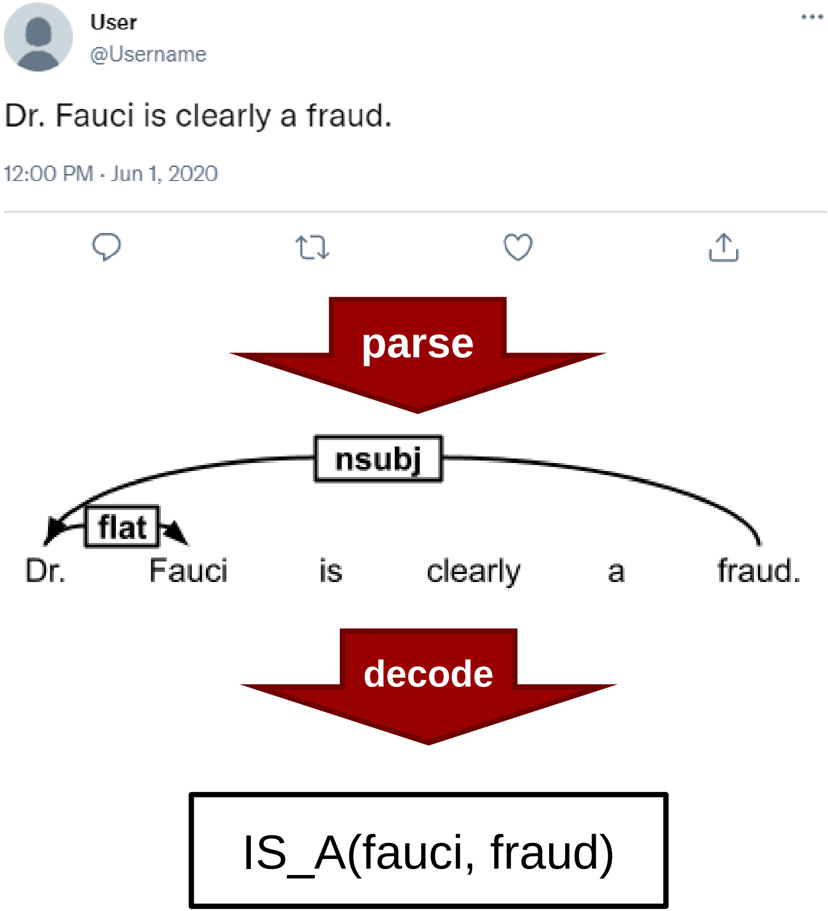
User

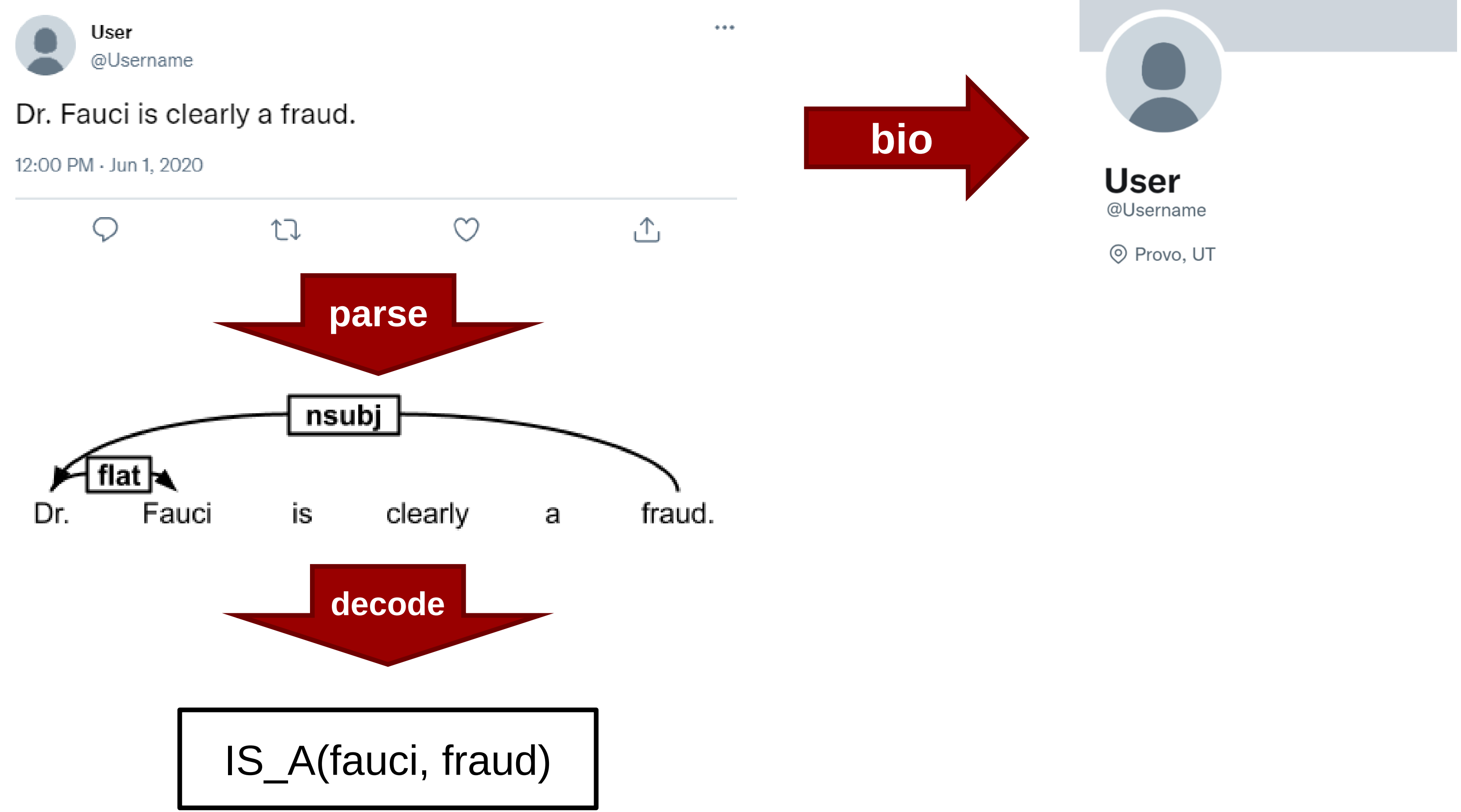
@Username

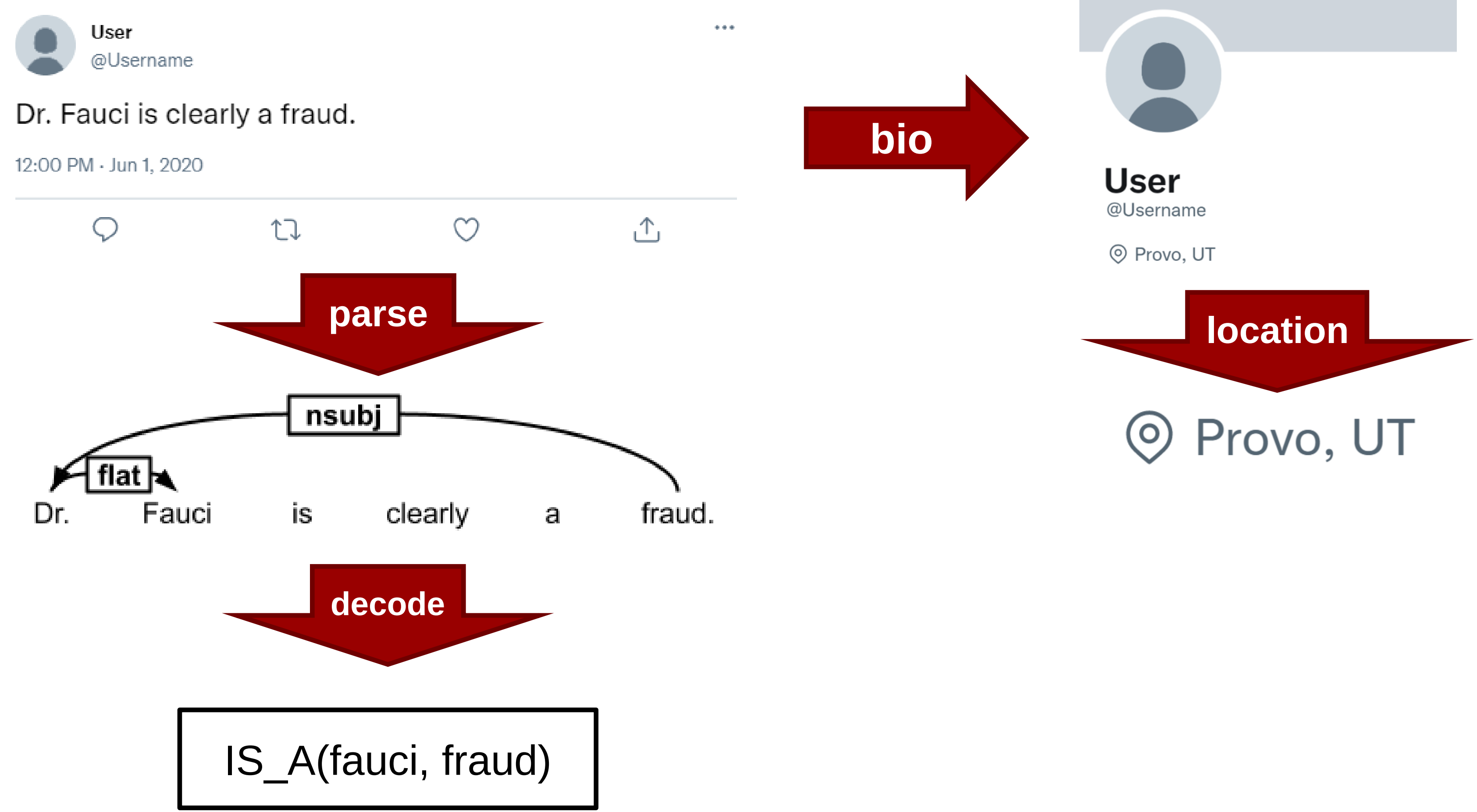
Dr. Fauci is clearly a fraud.

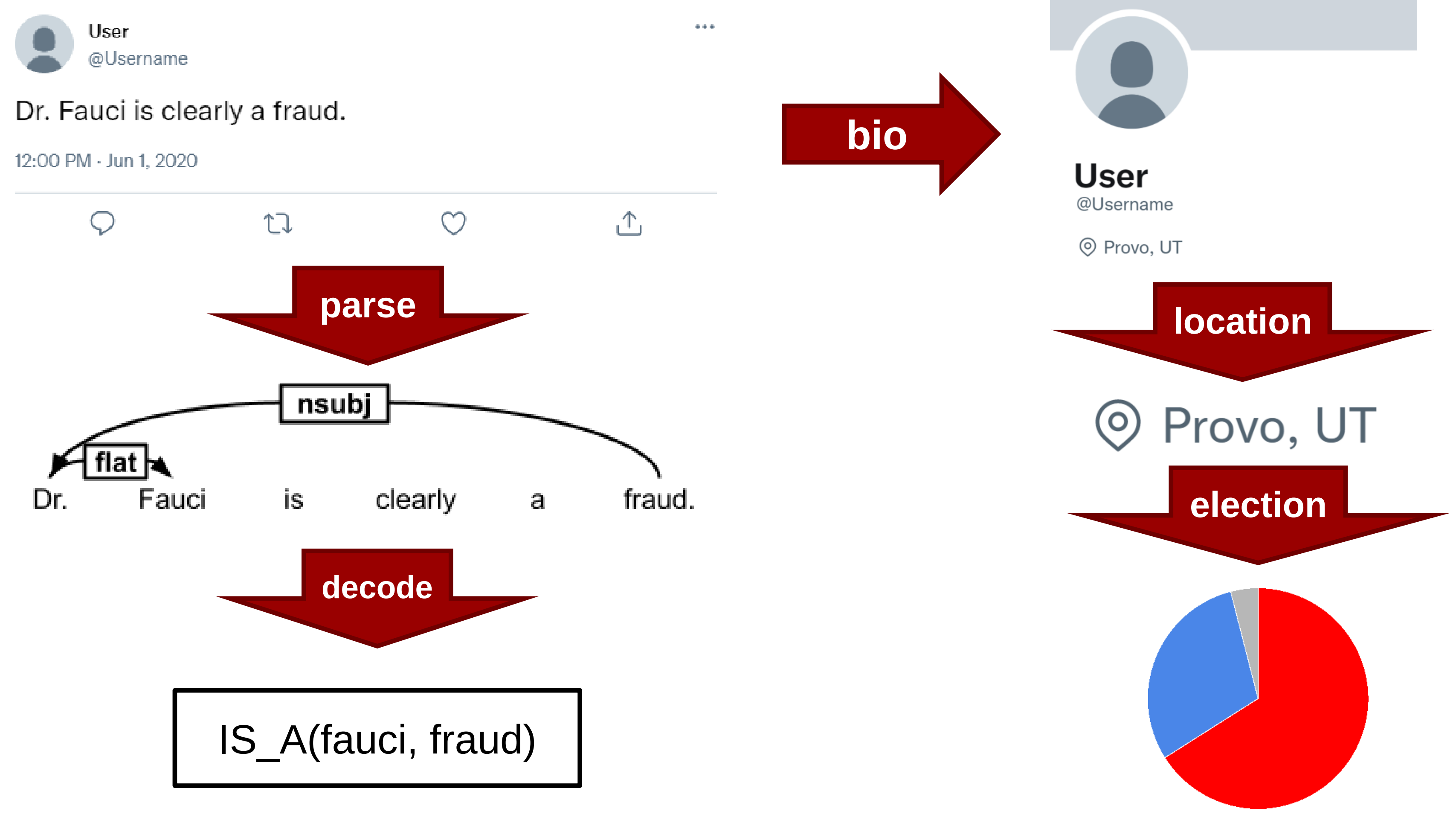
12:00 PM · Jun 1, 2020

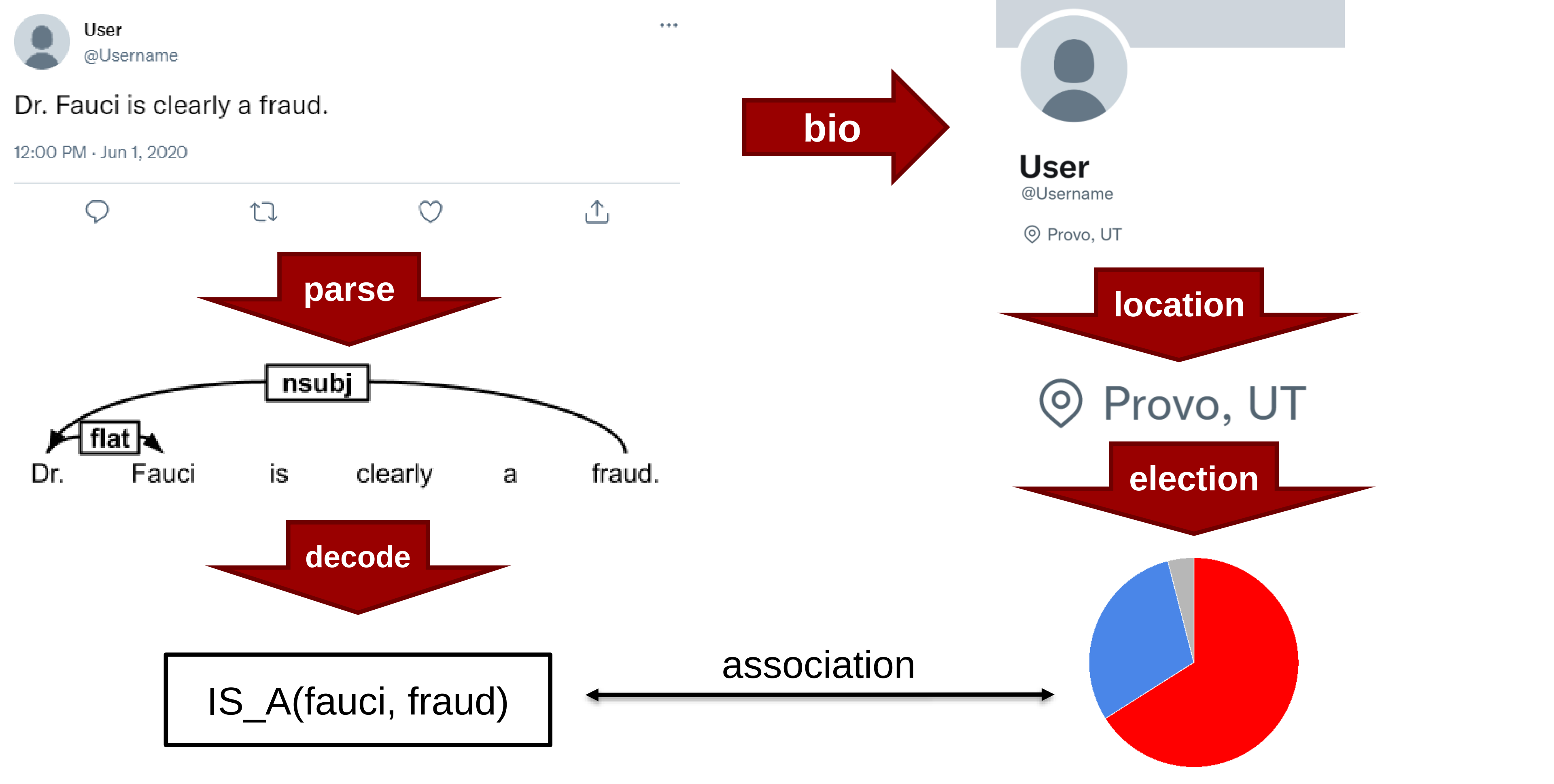

User
@Username
Dr. Fauci is clearly a fraud.
12:00 PM · Jun 1, 2020
parse
nsubj
flat
Dr.
Fauci
is
clearly
a
fraud.


User
@Username
Dr. Fauci is clearly a fraud.
12:00 PM · Jun 1, 2020
parse
nsubj
flat
Dr. Fauci is clearly a fraud.
decode
IS_A(fauci, fraud)


User
@Username
Dr. Fauci is clearly a fraud.
12:00 PM · Jun 1, 2020
bio
User
@Username
Provo, UT
parse
nsubj
flat
Dr. Fauci is clearly a fraud.
decode
IS_A(fauci, fraud)


User
@Username
Dr. Fauci is clearly a fraud.
12:00 PM · Jun 1, 2020
parse
nsubj
flat
Dr. Fauci is clearly a fraud.
decode
IS_A(fauci, fraud)
bio
User
@Username
Provo, UT
location
Provo, UT


User
@Username
Dr. Fauci is clearly a fraud.
12:00 PM · Jun 1, 2020
bio
User
@Username
Provo, UT
parse
nsubj
flat
Dr.
Fauci
is
clearly
a
fraud.
decode
IS_A(fauci, fraud)
location
Provo, UT
election


User
@Username
Dr. Fauci is clearly a fraud.
12:00 PM · Jun 1, 2020
bio
User
@Username
Provo, UT
parse
location
nsubj
flat
Dr. Fauci is clearly a fraud.
Provo, UT
election
decode
IS_A(fauci, fraud)
association

Case Study

- We obtain 75,325 tweets with a location field and the token ‘fauci’ from Twitter Decahose and run TweetIE on them.

Case Study

- We obtain 75,325 tweets with a location field and the token ‘fauci’ from Twitter Decahose and run TweetIE on them.
- We filter the extracted relations based on statistical significance of the mean 2020 US presidential electoral margins inferred from their user bio location fields.

Case Study

- We obtain 75,325 tweets with a location field and the token ‘fauci’ from Twitter Decahose and run TweetIE on them.
- We filter the extracted relations based on statistical significance of the mean 2020 US presidential electoral margins inferred from their user bio location fields.
- This is done using a *t*-statistic.

  In this case, it is the deviation of a belief’s mean political valence from the population’s mean political valence, normalized with respect to the belief’s standard error.

$$t \equiv \frac{\bar{x} - \mu}{s / \sqrt{N}}$$

Case Study

- We obtain 75,325 tweets with a location field and the token ‘fauci’ from Twitter Decahose and run TweetIE on them.
- We filter the extracted relations based on statistical significance of the mean 2020 US presidential electoral margins inferred from their user bio location fields.
- This is done using a *t*-statistic.

  In this case, it is the deviation of a belief’s mean political valence from the population’s mean political valence, normalized with respect to the belief’s standard error.

$$t \equiv \frac{\bar{x} - \mu}{s / \sqrt{N}}$$

- This allows us to gather beliefs that should be disproportionately representative of Biden-leaning or Trump-leaning users.

Relation	**Trump-Leaning** ($t < -2$)	**Biden-Leaning** ($t > 2$)
IS_A(fauci, *property*$_{\text{nom}}$)	murderer**, joke**, hack*, fraud*, rat*, flip*, idiot, flop, state, prison, fake, jail	nih**, hero, md, director, president
IS_A(fauci, *property*$_{\text{adj}}$)	fake*, little*, deep, liberal, wrong, corrupt	beloved, optimistic, best
AS_AGENT(fauci, *verb*)	sweat**, force**, need*, help*, read*, lie*, know*, let*, not_fund*, not_understand*, flip, predict, write, make, stick, hold, prove, want, not_say, admit, not_get, demand, issue, laugh, state, put, spread, pull	speak**, join*, warn*, throw, not_recommend, offer, provide, respond, consider, debunk, fail, reveal
AS_PATIENT(fauci, *verb*)	not_trust***, screw, prosecute, grill, keep to, arrest, expose, lock, do to, remove, accord to, look like, mean, blast, read	know*, feature, discredit, threaten, worship, join, insult
HAS_A(fauci, *object*)	friend*, nih*, family, mind, hand, ex-employee, involvement, fraud, mask	guidance, time
AS_CONJUNCT(fauci, *conj.*)	gates***, obama**, bill gates*, biden*, brix, cdc, rest, covid, nih, company, government	director, experts

Table 5: TweetIE extractions with at least 20 unique users with a county-level political valence t-statistic outside of [-2, 2]. Results are reported in decreasing absolute value t-statistic. * $|t| > 3$, ** $|t| > 4$, *** $|t| > 5$.

Conclusion

We can extract the beliefs of communities through social media, and by doing so pursue a socially-aware form of information extraction.

Thank you for listening.

Paper Available at *github.com/slanglab/TweetIE_WNUT2022*.