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

Chloe Eggleston
University of Massachusetts Amherst
cegglestonumass.edu

Brendan O'Connor
University of Massachusetts Amherst
brenocon@cs.umass.edu

Abstract

In this paper, we utilize recent advancements in social media natural language processing to obtain state-of-the-art syntactic dependency parsing results for social media English. We observe performance gains of 3.4 UAS and 4.0 LAS against the previous state-of-the-art as well as less disparity between African-American and Mainstream American English dialects. We demonstrate the computational social scientific utility of this parser for the task of socially embedded entity attribute analysis: for a specified entity, derive its semantic relationships from parses' rich syntax, and accumulate and compare them across social variables. We conduct a case study on politicized views of U.S. official Anthony Fauci during the COVID-19 pandemic.[1]

1 Introduction

Corpora of social media text contain wide ranges of beliefs that researchers may seek to analyze. But numerous studies have found significant challenges in applying natural language processing (NLP) techniques to social media, ranging from inconsistent spelling practices to continuously evolving terminology (Baldwin, 2012; Eisenstein, 2013).

Under the now-ubiquitous modeling paradigm of pretrained transformers (Peters et al., 2018; Devlin et al., 2019; Bender et al., 2021; Bommasani et al., 2021), it is crucial to include social media content in a language model pretraining corpus. BERTweet (Nguyen et al., 2020), a language model trained entirely on English Twitter, has shown state-of-the-art results in classification (Barbieri et al., 2020), part-of-speech (POS) tagging (Nguyen et al., 2020), and named entity recognition (NER) (Jiang et al., 2022) on social media English.

In addition, treebanks have been annotated to cover this specific variety of English. Tweebank v2

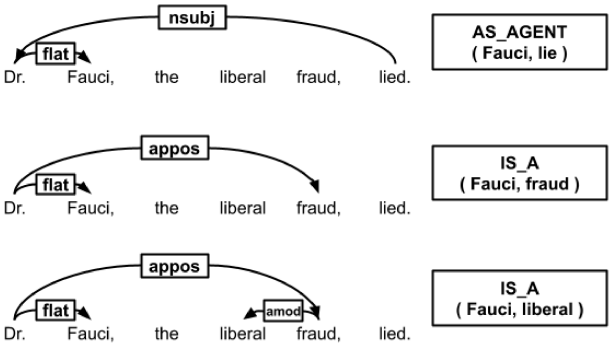

Figure 1: Examples of dependencies and TweetIE's entity attribute extraction system (§4).

(Liu et al., 2018) consists of 3,550 English tweets annotated according to Universal Dependencies (Nivre et al., 2020), and Jiang et al. (2022) add NER tags following the four-class CoNLL 2003 guidelines (Tjong Kim Sang and De Meulder, 2003).

Other work has considered the impact of demographic and dialectical factors on social media NLP. Blodgett et al. (2016, 2018) investigate linguistic variation of African-American English (AAE) on Twitter from aggregate user demographics, developing a small 500 tweet Universal Dependencies corpus half of which consists of tweets heavily using AAE. On this AAE subset, dependency parsers encounter worse performance than on Mainstream American English (MAE), and a similar AAE-MAE dialect disparity is widespread in other areas of NLP (e.g. Koenecke et al., 2020; Ziems et al., 2022).

Social media NLP advances could enable novel techniques in computational social science. Retrieval and representation of the beliefs and opinions of various groups and ideologies is of clear importance to many social sciences, with applications ranging from misinformation studies (Ayoub et al., 2021) to political science and economics (Ash et al., 2021).

With these goals in mind, we train a state-of-

[1]Code for this paper is available at: https://github.com/slanglab/TweetIE_WNUT2022

Proceedings of the 2022 COLING Workshop: The 8th Workshop on Noisy User-generated Text (W-NUT 2022), pages 38–50

the-art social media dependency parser, evaluating social media English performance, as well as AAE dialect disparity, among eleven alternative pretrained models (§3). To illustrate dependency parsing's utility for social media analysis, we implement a rule-based semantic attribute extractor to analyze authors' views toward an entity (Figure 1; §4), and evaluate it in a case study of political narratives surrounding the U.S. official Dr. Anthony Fauci during the COVID-19 pandemic—we compare extractions against the authors' social variable of geolocated election results (§5). We find our TweetIE system has better yield and higher precision for this task, compared to using previous open information extraction systems.

2 Related Work: Social Semantic Extraction

Natural language processing has been used to extract social insight from corpora in humanistic and social scientific study. Archak et al. (2007); Ghose et al. (2007) analyze the economic impact of dependency parse-extracted adjective modification from product reviews and seller feedback, associating perceived attributes with monetary prices. Narrative analysis of fictional characters has used dependency parses to extract attributes associated with character archetypes (Bamman et al., 2013); our semantic relation extractor follows and extends their approach. These dependency-based systems can be viewed as expanding on widely used collocation methods that tabulate words appearing near an entity (Baker, 2006); for example, Blinder and Allen (2016) use words directly before an entity (a rough adjective modifier extractor) to analyze attributes ascribed to immigrants in political discourse.

In the NLP context, outside of computational social science, open information extraction (OIE) is a related semantic approach that extracts relational tuples without a predefined schema, often applied to large heterogenous corpora, such as web data (Banko et al., 2007), typically using off-the-shelf NLP technologies such as part-of-speech (POS) tagging, named entity recognition (NER), semantic role labelling, and dependency parsing (Mausam, 2016). Our TweetIE information extractor uses a rule system working directly from dependency parses, following the approach of argument extraction and normalization systems PropS (Stanovsky et al., 2016) and PredPatt (White et al., 2016); the latter performs well on OIE benchmarks (Zhang et al., 2017). We share PredPatt's motivation to rely on Univerisal Dependencies parses, which have coverage and availability across many language varieties, including social media English here. This contrasts favorably to the domain-dependent limitations of machine-learned semantic role labeling (Carreras and Màrquez, 2005) and semantic dependency parsing (Oepen et al., 2014).

3 Dependency Parsing

3.1 Approach

Dependency parsing is typically performed by either transition-based (Covington, 2001; Nivre, 2003) or graph-based (Eisner, 1996) models, and can utilize representations including word embeddings, recurrent neural networks (Kiperwasser and Goldberg, 2016), and/or transformers (Grünewald et al., 2021). For experiments we use SuPar,[2] a Python library for syntactic and semantic parsing, to implement a graph-based transformer dependency parser using a deep biaffine attention (Dozat and Manning, 2017) layer, fine tuned from a HuggingFace-compatible pretrained transformer language model (Wolf et al., 2020). Due to its comparative performance (§3.3), we select BERTweet-base for the pretrained model for our final parser, fine-tuned[3] on Tweebank v2. Our experiments use the Tweebank v2 splits from its supplied "converted" CoNLL-compatible variant. We use "Twitter-Stanza (TB2)" for tokenization, since it achieves state-of-the-art results on Tweebank v2 tokenization (98.64 F1) (Jiang et al., 2022).[4]

Overall performance results are averaged over three seeds, shown in the last row of Table 1. Our results outperform the BiLSTM baselines featured in (Liu et al., 2018) by 3.4 unlabelled attachment score (UAS) and 4.0 labelled attachment score (LAS), as well as the previous state of the art, spaCy-XLM-RoBERTa, a transition-based parser using the multilingual transformer XLM-R (Conneau et al., 2020).

[2] https://github.com/yzhangcs/parser

[3] Hyperparameters tested (selections underlined): epochs=(50, 75, 100), warmup rate=(0.1, 0.15, 0.2), lr = (1e-5, 5e-6, 1e-4), projective=(false, true)

[4] SuPar provides an option to use either projective (Eisner, 2000; Zhang et al., 2020), or non-projective (matrix tree: Koo et al., 2007; Ma and Hovy, 2017) parsing; we use projective parsing, finding it attains slightly better performance (+0.3 UAS, +0.2 LAS from preliminary experiments), presumably since non-projectivity is rare in English (Peng and Zeldes, 2018).

This software platform easily allows us to compare training treebanks and pretrained language models, which we next explore for their impact on overall social media performance as well as dialect disparity.

System	UAS	LAS
TweeboParser (Kong et al., 2014)	81.4	76.9
Deep Biaffine (Dozat and Manning, 2017)	81.8	77.7
Ensemble Model (Liu et al., 2018)	83.4	79.4
spaCy-XLM-RoBERTa (Jiang et al., 2022)	83.8	79.4
SuPar-BERTweet (this work)	**87.2**	**83.4**

Table 1: Performance (in F1) of systems on Tweebank v2 test set. First four rows are from Liu et al. (2018) and Jiang et al. (2022).

3.2 Impact of Training Treebank

In order to measure the impact of treebanks on performance in this domain, we fine-tune RoBERTa-base (Liu et al., 2020) on three different treebanks, and measure its respective performance on Tweebank v2's test set using the CoNLL evaluation script. In order to ensure compatibility with this script and the ability to evaluate cross-treebanks, we drop the corpora-specific dependency subtypes.

We select the Georgetown University Multilayer Corpus (GUM) (Zeldes, 2017) and English Web Treebank (EWT) (Silveira et al., 2014). These include user-generated content and are 2.5 and 4.5 times larger than Tweebank v2 respectively. Despite their increased size, both see significant performance drops when evaluated on Tweebank v2 (Table 2).

	In-Domain		*Tweebank v2*	
Fine-tuning Corpus	**UAS**	**LAS**	**UAS**	**LAS**
GUM	92.9	90.9	66.6	57.1
EWT	90.7	89.6	70.2	61.5
Tweebank v2	85.7	81.4	**85.7**	**81.4**

Table 2: Performance (in F1) of SuPar-RoBERTa when trained on a given corpus, and its checkpoint with best dev split performance evaluated against the associated (in-domain) test split, as well as Tweebank v2.

3.3 Impact of Pretrained Model Selection

In addition to fine-tuning corpora, we observe a noticeable performance impact with respect to the models used, suggesting that pretraining has a role as well.

We evaluate the performance of eleven transformer models on Tweebank v2. BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2020), ELECTRA (Clark et al., 2020), XLNet (Yang et al., 2019), and DeBERTa v3 (He et al., 2021) are general purpose English transformers. XLM-R (Conneau et al., 2020) adapts RoBERTa to multilingual corpora, and InfoXLM (Chi et al., 2021) improves upon XLM-R with mutual information-improved loss function for cross-lingual context. TimeLMs (Loureiro et al., 2022) fine-tunes RoBERTa, training continually with larger temporal range, yield checkpoints for 2019 and 2019-2021 respectively. BERTweet is a RoBERTa model trained from scratch on Twitter. XLM-T (Barbieri et al., 2022) fine-tunes XLM-R on multilingual Twitter.

Model	UAS	LAS
General Purpose Models		
BERT-base-uncased	85.0	80.8
RoBERTa-base	85.7	81.4
ELECTRA-base	85.6	81.6
XLNet-base-cased	85.8	81.7
DeBERTa-v3-base	87.1	83.2
Multilingual Models		
XLM-R-base	86.2	82.4
InfoXLM-base	86.5	82.7
Social Media Models		
TimeLMs-2019	85.7	81.6
TimeLMs-2021	86.3	82.3
BERTweet-base	**87.2**	**83.4**
Multilingual Social Media Models		
XLM-T-base	86.5	82.0

Table 3: Performance (in F1) of SuPar dependency parsers using various pretrained transformers, fine-tuned and evaluated on the Tweebank v2 train and test splits, with the epoch of the best dev split performance being selected.

Table 3 indicates that stronger performance can be achieved through either better representations in modeling or through more social media pretraining, as seen respectively with DeBERTa v3 and BERTweet, one having the highest GLUE score (Wang et al., 2018; He et al., 2021), and the other trained entirely on Twitter.

3.4 Performance on Non-Majority English

One key challenge of working with social media text is the lack of adherence to any standardized dialect of a language, and the inclusion of significant minority dialects, such as high prevalence of African American English (AAE) (Jones, 2015; Blodgett et al., 2016). AAE dependency parsing includes significant challenges from recognizing null copulas to correctly understanding phonologically

Model	*Tweebank v2* MAE	AAE	R.E.	*TwitterAAE Deps* MAE	AAE	R.E.
General Purpose Models						
BERT	84.03	78.93	1.32	74.24	67.31	1.27
RoBERTa	84.40	78.61	1.37	75.46	67.50	1.32
ELECTRA	84.35	80.73	1.23	74.18	67.31	1.27
XLNet	84.41	79.85	1.29	75.72	69.75	1.25
DeBERTa-v3	**85.63**	82.44	1.22	77.08	71.90	1.23
Multilingual Models						
XLM-R	85.14	81.56	1.24	74.07	68.06	1.23
InfoXLM	85.17	82.11	1.21	74.44	68.19	1.24
Social Media Models						
TLMs19	84.22	81.33	1.18	76.23	72.22	1.17
TLMs21	84.87	82.30	1.17	76.91	72.38	1.20
BERTweet	85.42	**84.38**	**1.07**	**78.10**	**76.55**	**1.07**
Multilingual Social Media Models						
XLM-T	84.86	82.62	1.15	76.14	72.94	1.13

Table 4: MAE/AAE Performance (in LAS F1) and Relative Error of the models from Table 3, trained on Tweebank v2, and evaluated on Tweebank v2 test split and TwitterAAE deps.

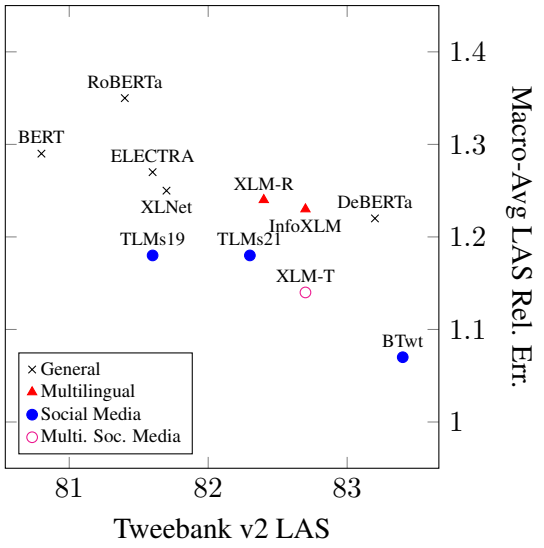

Figure 2: Graph of the performance of the models presented in Table 3 in LAS and macro-average of the relative error on the MAE/AAE split Tweebank v2 test set and TwitterAAE dependencies.

driven alternative spellings (Blodgett et al., 2018).

We evaluate the ability of the previously listed dependency parsing models by using the relative error of their performance on Mainstream American English (MAE) and AAE test sets,

$$\text{LASRelErr} = \frac{1 - \text{LAS}_{\text{AAE}}}{1 - \text{LAS}_{\text{MAE}}} \quad (1)$$

which attains 1 if accuracy is equal across dialects. We have found this to be always greater than 1.0 in our experiments, indicating performance is worse for the minority dialect, AAE.

In order to measure disparity on the fine-tuning source, we measure the relative error of both the TwitterAAE dependencies and use the TwitterAAE demographic dialect inference model to partition the Tweebank v2 test set into splits based on whether there was higher proportion MAE or AAE, yielding 951 and 249 tweets respectively. We also measure this on the TwitterAAE dependencies, which provides 250 tweets of both MAE and AAE respectively.

Table 4 and Figure 2 display the disparities between MAE and AAE performance on Tweebank v2 and TwitterAAE dependencies. This form of demographic evaluation offers insight on a key question that is not visible in the UAS / LAS scores alone: whether the performance gains come from overfitting on the majority dialect or increased performance across dialects.

We observe the social media models to have less LAS relative error than the general purpose models, with BERTweet, the model exposed to the most social media content, having less relative error than any model. As seen in Table 4, its state-of-the-art performance in Tweebank v2 does not suggest that it has the best performance with the syntax of standard English; it actually underperforms DeBERTa-v3, and only outperforms in total due to the 2 LAS difference on AAE. The relative error suggests that BERTweet's performance only adds on average 7% more error to a AAE sample compared to standard English, while general purpose models like DeBERTa v3 and RoBERTa add around 22.5% and 34.5% more, despite being fine-tuned on the same corpora.

The implications suggest that social media transformers capture the syntax not only better than their general purpose counterparts, regardless of architecture improvements, but also do it in a more equitable manner. This is important for applications sensitive to demographic effects.

4 TweetIE: Belief Extraction from Dependencies

A well-performing social media dependency parser, along with pre-existing POS and NER taggers, enable novel applications for computational social science. We apply these technologies for a belief extraction system, which decodes these syntactic structures into simple semantic representations and presents information applicable for computational social scientific purposes, specifically the delin-

eation of beliefs to communities represented by social variables. We call this system **TweetIE**.

4.1 Design Principles

In order to preserve the benefits of the domain-specific dependency parsing system while maintaining a simple overall system, we seek to:

- Infer relations using dependency parses, NER tags, and POS tags, not through lexicons that might only cover standard English.
- Focus on relations regarding a named entity and its attributes.
- Minimize the number of arguments for relations to allow for accumulation and comparison across social variables.

4.2 Target Entities and Pronoun Coreference

We focus our extraction based on the attributes of a single named-entity in a given tweet, through either specifying a name, or using an @ mention of that user's account. In the case of names of persons or organizations, we take into account the specified token, and expand it using the *flat* relation and the span of any BIO NER tags. If the root of this span is a *conj* dependency or if any relevant predicates have *conj* dependencies, we distribute dependency relations over them, as done in the CCprocessed/Enhanced++ variants of Stanford (De Marneffe and Manning, 2008) and Universal (Schuster and Manning, 2016) Dependencies.

In order to capture common forms of anaphora such as possessive pronoun usage, we implement a simple precision-oriented coreference system for binary gendered target entities. The user specifies the target's gender, and the system seeks any personal pronouns with the target as the antecedent. It first determines whether the target's mention(s) are in second person (denoted by the *vocative* relation) or third person (otherwise). It attributes pronouns of the determined person and specified gender to the target if there are no other entities (denoted by "PER" NER tags) mentioned in the text before it that are potentially applicable (as in they agree with regards to grammatical person).

To evaluate this system, we annotated a random sample of 100 tweets for whether their POS-tagged pronouns refer to the target entity of our later case study, Dr. Anthony Fauci (see Section 5). Our system achieved 33/39 (84.6%) precision and 33/52 (63.5%) recall.

4.3 Relations

We limit our focus to the following semantic relations:

4.3.1 IS_A

The IS_A relation covers any nominal or adjectival properties stated to directly pertain to the target entity, represented using the following patterns:[5]

1. target $\overset{\text{nsubj}}{\longleftrightarrow}$ property$_{nom}$
2. property$_{adj}$ $\xrightarrow{\text{nsubj}}$ target
3. target $\overset{\text{appos}}{\longleftrightarrow}$ property$_{nom}$
4. target $\xrightarrow{\text{compound}}$ property$_{nom}$
5. target $\xrightarrow{\text{amod}}$ property$_{adj}$
6. target $\overset{\text{nsubj}}{\longleftrightarrow}$ property$_{nom}$ $\xrightarrow{\text{amod}}$ property$_{adj}$
7. target $\overset{\text{appos}}{\longleftrightarrow}$ property$_{nom}$ $\xrightarrow{\text{amod}}$ property$_{adj}$

Patterns 1 and 2 detect subject-complement linking through copular clauses, even when explicit copulas are omitted. Pattern 3 detects appositions, and Pattern 4 detects titles that do not make up fully formed appositions (ex: "*President* Obama").

Pattern 5 detects adjective modifiers. Patterns 6 and 7 detect adjective modifiers of previously captured nominal properties, hoping to capture intersective adjectives (ex: "Trump is a *famous* person").

4.3.2 HAS_A

The HAS_A relation pertains to any object possessed the target entity, implemented through possessive modification.

1. object$_{nom}$ $\xrightarrow{\text{nmod:poss}}$ target

4.3.3 AS_AGENT, AS_PATIENT

The AS_AGENT and AS_PATIENT relations pertain to actions performed by the target entity and performed upon the target entity respectively.

1. active verb $\xrightarrow{\text{nsubj}}$ target$_{agent}$
2. active verb $\xrightarrow{\text{obj}}$ target$_{patient}$
3. passive verb $\xrightarrow{\text{nsubj:pass}}$ target$_{patient}$
4. passive verb $\xrightarrow{\text{obl}}$ target$_{agent}$
5. active verb $\xrightarrow{\text{obl}}$ target$_{patient}$ $\xrightarrow{\text{case}}$ prep.

[5] H$\rightarrow$D represents a relation from a head H to its dependency D, while X$\longleftrightarrow$Y indicates a relation in either direction.

Patterns 1 and 2 account for active tense verbs, while 3 and 4 account for passive tense verbs, which are distinguished from active tense by the presence of a *nsubj:pass* dependency.

Pattern 5 consists of when the target acts as an adjunct of the verb using a preposition, and is lexicalized through appending the preposition to the verb (ex: "I *stand with* Obama", "He *listens to* Bill Gates").

4.3.4 AS_CONJUNCT

The AS_CONJUNCT relations pertains to any nominal conjoined with the target entity. If this nominal consists of a named-entity, it is expanded in the same manner as the target entity (through *flat* dependencies and BIO NER spans).

1. target $\overset{\text{conj}}{\longleftrightarrow}$ conjunct

Although this has no explicit semantic meaning, it suggests that the two hold a latent semantic relationship, such as co-hypernymy (Snow et al., 2004).

4.4 Negation

A theoretical concern for this mode of semantic extraction deals with the presence of negative polarity adverbs. Intuitively when comparing these extractions across social variables, this form of negation should not be accumulated in the same case as the original clause.

However, dependency relations describing negative polarity do not exist in the current version of Universal Dependencies, with the *neg* relation being removed in Universal Dependencies v2 (Nivre et al., 2020). In order to account for this, we check previous version of treebanks for user-generated content with this relation: specifically EWT v1.4. In this treebank, the *neg* relation only covers the following tokens: ['no', 'not', 'never', 'nt', 'n't'].

We utilize this list by adding a negative polarity to any relation extracted that is modified by any of those tokens. This is implemented by prepending the extraction's argument with 'not_', an approach used in sentiment analysis (Das and Chen, 2007). A word list in this vein has clear limitations - it does not cover social media variations in spelling, yet it allows us to capture this quality on its most common variants.

4.5 Evaluation

TweetIE can either be evaluated through the accuracy of each component, or qualitatively through how well its outputs model the social variables. On a component level, its accuracy depends foremost upon the performance of its dependency parsing, NER, and POS models.

The performance of the dependency parsing has been described in Section 3. For POS and NER tagging use Jiang et al. (2022)'s state-of-the-art-models: "HuggingFace-BERTweet (TB2+EWT)" for POS (which achieved 95.38 UPOS accuracy on Tweebank v2) and "HuggingFace-BERTweet (TB2+W17)" for NER (which achieved 74.35 F1 on Tweebank-NER).

Finally, we examine externally validity by investigating the model's ability to capture social context in the following case study.

5 Case Study: COVID-19 Polarization

A key source of variation in opinion is with respect to political ideology, and social media is rife with arguments about political figures specifically. In this section, we show TweetIE's ability to capture the ideological attributes of said figures, specifically the attributes social media users ascribe to Dr. Anthony Fauci, director of the National Institute of Allergy and Infectious Diseases, who is a key figure in United States COVID-19 discourse. While TweetIE could be used to study a network of entities and their relations, we find focusing on a single entity is a useful and insightful first step.

5.1 Corpora Design and Configuration

We collect a corpus of tweets from Twitter Decahose with the token 'fauci' spanning from March 1, 2020 to December 31, 2021. We filter to messages with geographic location information: either from a tweet's official API geotag, or from its author having a self-described *user.location* text field consisting of a city and state in postal code notation (e.g. "Minneapolis, MN"). We look up these fields using the US Census Bureau's Place boundary shapefiles,[6] and as a proxy for political valence, each valid place is paired with its county's Biden-Trump margin, the difference of Joe Biden's versus Donald Trump's percentage votes won in the 2020 U.S. presidential election (MIT Election Data & Science Lab, 2018).[7] Additionally, we discard any tweets from verified users or users with over 10,000 followers in order to capture conversational

[6] https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.2020.html

[7] For Alaska we use the state-level result, since it does not provide county-level results.

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. * $|t| > 3$, ** $|t| > 4$, *** $|t| > 5$.

dialogue rather than statements by reporters and officials.

5.2 Results and Qualitative Evaluation

We obtain 75,325 tweets, which have an electoral margin average of 22.8 and standard deviation of 33.9. TweetIE yields 13,532 unique triples of *relation*(Fauci, *token*), which we call unique extractions. The counts of these sum to 99,633 total extractions overall. In order to improve aggregation, we lowercase and normalize the *token* terms with NLTK's WordNetLemmatizer (Loper and Bird, 2002), and remove stopwords from NLTK's English stopword list.

For each tuple that is expressed by at least 20 unique users, we use a one-sample student's *t* statistic to determine if the mean author-geography political sentiment of the tuple is significantly different than the corpus population's. We require $|t| > 2$ as a rough filter for traditional statistical significance.[8] This method for term ranking is appropriate for the continuous variable of political sentiment. Since words' frequencies greatly vary, rare terms tend to be sentiment average outliers; the t statistic's normalization by standard error helps control for an expression's sample size.[9]

This results in 110 expressions have test statistics greater than 2 or less than -2, shown in Table 5. These reflect common political narratives concerning Fauci and his COVID-19 response. Political scientific work has found liberal respondents to be more trusting in COVID-19 experts such as Fauci than conservatives (Kerr et al., 2021), as well as more hesitant towards COVID-19 vaccination (Khubchandani et al., 2021), whose development and production Fauci was involved with.

The notable considerations of Fauci as a joke or a fraud, or that he lies or is not trusted, reflect lack of trust in Fauci by the Trump-leaning. Likewise, suggesting that Fauci is a hero or beloved, as well as emphasizing what he says or his warnings show trust in Fauci from the Biden-leaning.

There are elements of COVID-19 related right-wing conspiracism in the Trump-leaning extractions as well. Common antecedents of COVID-19 conspiracism include the notions of a fraudlent pandemic, vaccination as a weapon, suspicions of the government, pharmaceutical industry, Democrats, and Bill Gates (van Mulukom et al., 2022). In our analysis this theme surfaces in Gates' appearance as a frequent conjunct; furthermore, many Trump-leaning extractions indicate Fauci as a murderer for his involvement in vaccination, or as someone who should be prosecuted, arrested, or put in prison. A shortcoming of our token-based approach can be seen with the bigram "deep state", a key narrative element, being split into two separate IS_A statements, which would be better viewed together.

[8] Under the central limit theorem, $|t| > 1.96$ corresponds to p-value < 0.05. Given multiple hypothesis testing issues we do not propose a formal significance test interpretation, though false discovery rate or other methods could be applied (Bamman et al., 2012).

[9] Social science NLP has often ranked terms by analogous confidence measures of term frequency versus a discrete social variable, such as χ^2 (Gentzkow and Shapiro, 2010) or log-odds posterior confidence (Monroe et al., 2008).

5.3 Alternative Systems

To demonstrate TweetIE's value over open information extraction (OIE) systems for this task, we evaluate two other systems against the Fauci corpus. These are ReVerb, a lexical pattern and POS-based system (Fader et al., 2011), and ClausIE, a Stanford Dependencies based system (Del Corro and Gemulla, 2013). ReVerb was selected to represent systems that do not require a parser, while ClausIE is the state-of-the-art system on the BenchIE OIE benchmark (Gashteovski et al., 2022). Like other OIE systems, these extract <Arg1, Relation, Arg2> tuples where relations and arguments are (normalized) strings from the sentence. While some work has sought to use OIE triples for social insight (Ash et al., 2021), we map them to IS_A, AS_AGENT, and AS_PATIENT for comparability.[10]

ReVerb is an OIE system that extracts relations using POS tags, noun phrase chunks, and lexical constraints; its output OIE triples have normalized values. If the relation is normalized to "be", and the target entity is in one of the arguments, we extract the other argument as IS_A. Otherwise if the target entity is in Argument 1, the relation is extracted as AS_AGENT, and if in Argument 2, AS_PATIENT.

ClausIE parses a sentence using Stanford Dependencies, using pattern detectors to eventually arrive at final OIE triples ("propositions"). While the relations are short, unfortunately the arguments can be very long phrases, and cannot be accumulated for counts or social variable aggregates. For a fair and generous comparison, we utilize ClausIE's intermediate representation of "clause" tuples, which are based on one of seven syntactic patterns such as copular clauses (SVC) or monotransitives (SVO); these are tuples of syntactic head words.[11] For IS_A, we take all detected copular clauses with the target entity in the subject or complement role, recording the remaining of the two as an IS_A extraction. For AS_AGENT, we extract the verb argument of any non-copular clause with the target entity in the subject role. We do the same for AS_PATIENT if the target entity is in the complement or object roles. We normalize these outputs in the same way as TweetIE.

As neither ReVerb nor ClausIE use coreference resolution, we present TweetIE with and without coreference enabled for comparison.

The systems share common extractions; the top ten IS_A share *fraud, one, liar, expert, doctor, man*, the top five AS_AGENT share *say* and *tell*, and the top five AS_PATIENT share *fire* and *trust*.

This suggests that they all can capture similar phenomena in the dataset, yet the amount of information they actually extract (total yield) varies significantly. Over these three patterns, ReVerb yields 16,980 total extractions, ClausIE yields 43,097, $\text{TweetIE}_{\text{no-coref}}$ yields 61,484, and TweetIE yields 74,572. TweetIE's superior yield is important, as the statistical inference over social variables is reliant on the ability to extract on a scale large enough to be representative; the smaller yield from ReVerb is likely to be inadequate. This occurs in our social analysis criteria of requiring terms to have at least 20 unique users and a t-statistic outside of [-2,2]. For IS_A, AS_AGENT and AS_PATIENT respectively, ReVerb yields 1/1/2, ClausIE yields 12/22/6, $\text{TweetIE}_{\text{no-coref}}$ yields 23/28/22, and TweetIE yields 26/39/22.

In addition, ClausIE struggled to understand @ mentions, and they appeared as extractions of every variety instead of extraneous vocative mentions (second most common IS_A and AS_AGENT, most common AS_PATIENT). We attribute this to ClausIE's reliance on a parser not trained on a social media domain without the benefit of transformer modeling.

Finally, we perform a precision evaluation to judge which systems' extractions more accurately reflect semantic implications of the text. We randomly sample 250 tweets and annotate whether each semantic tuple from ReVerb, ClausIE, and TweetIE is present in or directly implied by the text. The annotator (first author) was presented with the text of the tweet, along with the outputs of all systems in a random order (with system names hidden). Each output was labelled as implied or not implied; for each system we report the precision and its 95% confidence interval from bootstrapped standard errors, from 100,000 simulations of resampling at the tweet level. This results in ReVerb having a precision of $73.8 \pm 12.5\%$ (31/42), ClausIE having a precision of $66.1 \pm 8.5\%$ (84/129), and TweetIE having the highest precision at $83.5 \pm 4.7\%$

[10] While IS_A requires adaptation from the OIE framework, AS_AGENT and AS_PATIENT relations can be viewed as a Davidsonian-style binarization of an OIE triple: e.g. <*Fauci, hate, us*> is equivalent to *AGENT(hate, Fauci)* $\wedge$ *PATIENT(hate, us)*, at least assuming a Dowty (1991)-style proto-role theory of what OIE Arg1 and Arg2 mean.

[11] A shortcoming of this approach is that ClausIE only applies coordination handling to the final OIE triples; it was not clear to us if it was possible to backport this feature to the clause tuples.

(187/222).

The difference between TweetIE and ClausIE is statistically significant ($p < 0.001$). Thus TweetIE is able to achieve its higher yield but without any cost to precision, presumably due to its modeling and rule improvements.

6 Conclusion and Future Work

The annotations from Tweebank v2 and the performance improvements from BERTweet have lead to significant advancements in social media dependency parsing, with performance gains of 3.4 UAS and 4.0 LAS, as well as significantly lessening how much performance lags for the non-standard language variety of African-American English.

These achievements enable downstream applications of syntactic parsing on social media data, of which we note information extraction as being especially utilizable for computational social scientific means. We outline a process to decode these dependency parses into aggregatable semantic structures, for comparisons with social variables that one may seek to study.

We show how one can model political narratives with respect to named entities with a case study on elements and actions attributed to Dr. Anthony Fauci on social media during the COVID-19 pandemic. Through this, we replicate findings in social scientific literature on the topic, and we have similar extractions to pre-existing open information extraction yet with increased yield, enabling more substantial computational social scientific analyses.

Future work can build upon these foundations by extending these techniques to beliefs spanning multiple entities, by considering additional social variables, or by taking into account temporal effects through timestamps. This could allow for the observation of more complex phenomena, such as actions from an entity towards another entity or the adoption and decline of beliefs over time.

Acknowledgements

We are thankful for feedback from the UMass NLP and AAE groups, as well as comments from the reviewers. This work was supported by National Science Foundation grants 1845576 and 2042939. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Nikolay Archak, Anindya Ghose, and Panagiotis G. Ipeirotis. 2007. Show me the money!: deriving the pricing power of product features by mining consumer reviews. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 56–65.

Elliott Ash, Germain Gauthier, and Philine Widmer. 2021. Relatio: Text semantics capture political and economic narratives.

Jackie Ayoub, X. Jessie Yang, and Feng Zhou. 2021. Combat covid-19 infodemic using explainable natural language processing models. *Information Processing & Management*, 58(4):102569.

Paul Baker. 2006. *Using Corpora in Discourse Analysis*. Bloomsbury Discourse. Bloomsbury Academic.

Timothy Baldwin. 2012. Social media: Friend or foe of natural language processing? In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 58–59, Bali, Indonesia. Faculty of Computer Science, Universitas Indonesia.

David Bamman, Brendan O'Connor, and Noah A. Smith. 2012. Censorship and deletion practices in Chinese social media. *First Monday*, 17(3).

David Bamman, Brendan O'Connor, and Noah A. Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361, Sofia, Bulgaria. Association for Computational Linguistics.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, page 2670–2676, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Francesco Barbieri, Luis Espinosa-Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. In *Proceedings of LREC*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Scott Blinder and William L. Allen. 2016. Constructing immigrants: Portrayals of migrant groups in british national newspapers, 2010–2012. *International Migration Review*, 50(1):3–40.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2018. Twitter Universal Dependency parsing for African-American and mainstream American English. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Melbourne, Australia. Association for Computational Linguistics.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 152–164, Ann Arbor, Michigan. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Michael A. Covington. 2001. A fundamental algorithm for dependency parsing. In *In Proceedings of the 39th Annual ACM Southeast Conference*, pages 95–102.

Sanjiv R. Das and Mike Y. Chen. 2007. Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388.

Marie-Catherine De Marneffe and Christopher D Manning. 2008. Stanford typed dependencies manual. Technical report, Stanford University.

Luciano Del Corro and Rainer Gemulla. 2013. Clausie: Clause-based open information extraction. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, page 355–366, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia. Association for Computational Linguistics.

Jason Eisner. 2000. *Bilexical Grammars and their Cubic-Time Parsing Algorithms*, pages 29–61. Springer Netherlands, Dordrecht.

Jason M. Eisner. 1996. Three new probabilistic models for dependency parsing: An exploration. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Kiril Gashteovski, Mingying Yu, Bhushan Kotnis, Carolin Lawrence, Mathias Niepert, and Goran Glavaš. 2022. BenchIE: A framework for multi-faceted fact-based open information extraction evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4472–4490, Dublin, Ireland. Association for Computational Linguistics.

Matthew Gentzkow and Jesse M. Shapiro. 2010. What drives media slant? Evidence from U.S. daily newspapers. *Econometrica*, 78(1):35–71.

Anindya Ghose, Panagiotis Ipeirotis, and Arun Sundararajan. 2007. Opinion mining using econometrics: A case study on reputation systems. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 416–423, Prague, Czech Republic. Association for Computational Linguistics.

Stefan Grünewald, Annemarie Friedrich, and Jonas Kuhn. 2021. Applying occam's razor to transformer-based dependency parsing: What works, what doesn't, and what is really necessary. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 131–144, Online. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Hang Jiang, Yining Hua, Doug Beeferman, and Deb Roy. 2022. Annotating the tweebank corpus on named entity recognition and building nlp models for social media analysis. In *Proceedings of the Language Resources and Evaluation Conference*, pages 7199–7208, Marseille, France. European Language Resources Association.

Taylor Jones. 2015. Toward a Description of African American Vernacular English Dialect Regions Using "Black Twitter". *American Speech*, 90(4):403–440.

John Kerr, Costas Panagopoulos, and Sander van der Linden. 2021. Political polarization on covid-19 pandemic response in the united states. *Personality and Individual Differences*, 179:110892.

Jagdish Khubchandani, Sushil Sharma, James H. Price, Michael J. Wiblishauser, Manoj Sharma, and Fern J. Webb. 2021. Covid-19 vaccination hesitancy in the united states: A rapid national assessment. *Journal of Community Health*, 46(2):270–277.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional LSTM feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. 2014. A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1001–1012, Doha, Qatar. Association for Computational Linguistics.

Terry Koo, Amir Globerson, Xavier Carreras, and Michael Collins. 2007. Structured prediction models via the matrix-tree theorem. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 141–150, Prague, Czech Republic. Association for Computational Linguistics.

Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A. Smith. 2018. Parsing tweets into Universal Dependencies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 965–975, New Orleans, Louisiana. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. RoBERTa: A robustly optimized BERT pretraining approach.

Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2017. Neural probabilistic model for non-projective MST parsing. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 59–69, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Mausam Mausam. 2016. Open information extraction systems and downstream applications. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 4074–4077. AAAI Press.

MIT Election Data & Science Lab. 2018. County Presidential Election Returns 2000-2020.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin' Words: Lexical feature selection and evaluation for identifying the con tent of political conflict. *Political Analysis*, 16(4):372.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Joakim Nivre. 2003. An efficient algorithm for projective dependency parsing. In *Proceedings of the Eighth International Conference on Parsing Technologies*, pages 149–160, Nancy, France.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Dan Flickinger, Jan Hajič, Angelina Ivanova, and Yi Zhang. 2014. SemEval 2014 task 8: Broad-coverage semantic dependency parsing. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 63–72, Dublin, Ireland. Association for Computational Linguistics.

Siyao Peng and Amir Zeldes. 2018. All roads lead to UD: Converting Stanford and Penn parses to English Universal Dependencies with multilayer annotations. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 167–177, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Sebastian Schuster and Christopher D. Manning. 2016. Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2371–2378, Portorož, Slovenia. European Language Resources Association (ELRA).

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Rion Snow, Daniel Jurafsky, and Andrew Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press.

Gabriel Stanovsky, Jessica Ficler, Ido Dagan, and Yoav Goldberg. 2016. Getting more out of syntax with props.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Valerie van Mulukom, Lotte J. Pummerer, Sinan Alper, Hui Bai, Vladimíra Čavojová, Jessica Farias, Cameron S. Kay, Ljiljana B. Lazarevic, Emilio J.C. Lobato, Gaëlle Marinthe, Irena Pavela Banai, Jakub Šrol, and Iris Žeželj. 2022. Antecedents and consequences of covid-19 conspiracy beliefs: A systematic review. *Social Science & Medicine*, 301:114912.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal decompositional semantics on Universal Dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1713–1723, Austin, Texas. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Sheng Zhang, Rachel Rudinger, and Benjamin Van Durme. 2017. An evaluation of PredPatt and open IE via stage 1 semantic role labeling. In *IWCS 2017 — 12th International Conference on Computational Semantics — Short papers*.

Yu Zhang, Zhenghua Li, and Min Zhang. 2020. Efficient second-order TreeCRF for neural dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3295–3305, Online. Association for Computational Linguistics.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. VALUE: Understanding dialect disparity in NLU. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3701–3720, Dublin, Ireland. Association for Computational Linguistics.