OKG: A Knowledge Graph for Fine-grained Understanding of Social Media Discourse on Inequality

Inès Blin
i.blin@vu.nl
Vrije Universiteit Amsterdam, The Netherlands
Amsterdam, The Netherlands
Sony Computer Science Laboratories-Paris, France
Paris, France

Lise Stork
l.stork@vu.nl
Vrije Universiteit Amsterdam, The Netherlands
Amsterdam, The Netherlands

Laura Spillner
laura.spillner@uni-bremen.de
University of Bremen, Germany
Bremen, Germany

Carlo R.M.A. Santagiustina
carlo.santagiustina@univiu.org
Venice International University, Italy
Venice, Italy
Ca'Foscari University of Venice, Italy
Venice, Italy

ABSTRACT

In recent years, social media platforms such as Twitter have allowed people to voice their opinions by engaging in online discussions. The availability of such discussions has garnered interest amongst researchers in analyzing the dynamics on critical topics, such as inequality. Most of the current strategies are, however, limited with respect to conveying the fine-grained opinions of users, focusing on tasks such as sentiment analysis or topic modeling that extract coarse categorizations. In this work, we address this challenge by integrating a Twitter corpus with the output of finer-grained semantic parsing for the analysis of social media discourse. To do so, we first introduce the *OBservatory Integrated Ontology (OBIO)* that integrates social media metadata with various types of linguistic knowledge. We then present the *Observatory Knowledge Graph (OKG)*, a knowledge graph in terms of the ontology, populated with tweets on inequality. We lastly provide use cases showing how the knowledge graph can be used as the backbone of a social media observatory, to facilitate a deeper understanding of social media discourse.

CCS CONCEPTS

• **Computing methodologies → Information extraction**; **Semantic networks**; **Ontology engineering**.

KEYWORDS

Social Media Discourse, Ontology Engineering and Population

ACM Reference Format:
Inès Blin, Lise Stork, Laura Spillner, and Carlo R.M.A. Santagiustina. 2023. OKG: A Knowledge Graph for Fine-grained Understanding of Social Media Discourse on Inequality. In *Knowledge Capture Conference 2023 (K-CAP '23), December 05–07, 2023, Pensacola, FL, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3587259.3627557

K-CAP '23, December 05–07, 2023, Pensacola, FL, USA

ACM ISBN 979-8-4007-0141-2/23/12.
https://doi.org/10.1145/3587259.3627557

1 INTRODUCTION

In recent years, the proliferation of social media platforms such as Twitter has allowed people to voice their opinions by engaging in online discussions. The accessibility of several of these online resources has piqued the interest of researchers and policymakers. They are now eager to capture and discover perspectives and impactful narratives circulating throughout society on critical topics like migration, war, vaccination, inequality, or climate change. Many works have addressed this interest by publishing dashboards [35], social media datasets [8, 13, 14], and by leveraging automated natural language processing (NLP) strategies for the discovery and analysis of online debates [23, 29, 30, 36].

Traditional NLP strategies for the understanding of online debates have focused commonly on sentiment analysis, opinion mining and topic modeling [26, 34]. Such strategies are limited in their capabilities to convey fine-grained analysis of the opinions of individuals or social groups in the form of narratives, arguments or claims, given that statistical strategies are used to label natural language texts that are vague and imprecise in nature. The complexity of Twitter posts, like the usage of slang and acronyms, further complicates precise interpretation.

Fine-grained text analysis techniques, e.g., named entity recognition [12], relation extraction [39], semantic role labeling [21], frame extraction [32], and dependency parsing [19] can provide fine-grained insights into the specific stances communities take in a debate or narrative. Such techniques allow researchers to retrace the provenance of their findings [38], ensuring the validity of results, reproducibility of experiments, and fostering transparency in research. We argue that the integration of tweet metadata with the output of both fine and coarse-grained NLP analyses into a single integrative network, allows researchers to perform increasingly complex analyses to better understand online debates.

With this goal in mind, this work introduces the *Observatory Knowledge Graph (OKG)*, a knowledge graph in terms of the *OBservatory Integrated Ontology (OBIO)*, which integrates tweet metadata with various types of linguistic knowledge and Linked Open Data (LOD), such as named entities, dependencies, and PropBank rolesets. We present use cases that demonstrate how the OKG can aid researchers in understanding online discussions on inequalities [40], e.g., perceived causes, driving factors and effects of inequalities, as well as the relations among mentioned entities (people, places, events, organizations, etc.) Thus, the contributions of this paper are twofold:

(1) the *OBservatory Integrated Ontology (OBIO)*[1], which integrates tweet metadata with linguistic analysis data.
(2) the *Observatory Knowledge Graph (OKG)*, which integrates a Twitter corpus with the output of both fine-grained and coarse-grained NLP analyses, and present relevant use cases. Code and the OKG are available via a github library[2] and Zenodo[3] respectively. Upon request, access to the SPARQL endpoint[4], maintained both by Triply[5] and the IISG[6], can be provided.

Such an holistic picture can offer valuable insights into public perceptions, social trends, and the perceived effectiveness of inequality mitigation efforts. It can inform decision-making processes, guide empirical research efforts in the field, and contribute to a more informed debate about inequality.

2 RELATED WORK

Discourse on social media. We focus on related work that analyzes discourse in social media, and more particularly on Twitter. First, we present work that is the closest to ours, ie., that uses knowledge graphs as intermediate representations. Second, we present work that uses NLP techniques to get insights from social media data, not necessarily in graph format.

The work that is the most similar to ours is TweetsKB [14] and TweetsCOV19 [13], that is a subset of TweetsKB containing Covid-related tweets. They created a RDF(S) model for describing metadata and annotation information for a collection of tweets, and they presented useful use cases such as entity-centric data exploration. The authors of [10] used TweetsKB as a starting point to analyze public attitudes towards controversies, specifically on the topic of migration, and propose the following components in their pipeline: topic modeling, sentiment analysis, hate/speech detection and entity linking to Wikidata. We extend these models with further fine-grained linguistic information retrieved from tweet texts. More specifically, we add information on semantic roles of entities, as well as sentence grammar.

More recently, a lot of attention has been given to analysing social narratives about Covid19 on Twitter. [8] released a multilingual dataset containing tweets about the coronavirus, and [37] monitored the mood of India during the Covid pandemic starting from tweets. Lastly, [2] analyzed trending hashtags on Twitter with a specific use-case on Covid-19, and mapped the output to knowledge bases like Framester [16].

Tweet2story [6] is an NLP pipeline to automatically extract narratives from tweets in the form of simple graphs. Their pipeline includes: actor entity extraction, time entity extraction, event entity extraction, link extraction and semantic role extraction. Our work is complementary, providing a complete ontology for the knowledge graph output. The authors of [22] used the Dutch vaccination debate on Twitter to identify online communities, narratives and interactions. [33] combines an NLP pipeline with network analysis to extract conflicting narrative mechanics from Twitter data.

Knowledge Graph from text. We first present ontologies that were used to model textual data, and more specifically focused on social media data. We then present existing resources.

The NLP Interchange Format (NIF) ontology [18] was designed to integrate text into knowledge graphs, whereas the NLP Annotation Format (NAF) ontology [15] additionally focused on linking linguistic annotations. The OntoLex ontology [24] models lexical data in the semantic web. To represent social media data in the form of a knowledge graph, TweetsKB [14] reused existing ontologies such as the Semantically Interlinked Online Community (SIOC) [5]. The Influence Tracker ontology [27] integrates tweet data with quality metrics about Twitter users. Framester [16] is a frame-based ontological resource that bridges major linguistic resources such as FrameNet and PropBank. In our model and graph, we reuse and extend the existing TweetsKB model, and integrate it with (i) text analysis using NIF, (ii) new metrics that are not in the Influence Tracker ontology, and (iii) Framester PropBank rolesets.

In terms of existing resources, [28] proposes an NLP pipeline to build an event-centric knowledge graph from news data, using frame semantics. Throughout the BioSampo project, researchers extracted knowledge graphs from plenary debates, to analyze parlementary language and culture [31]. TakeFive [1] transforms texts into a frame-oriented knowledge graph, and FRED [17] also parses natural text into linked data.

3 KNOWLEDGE GRAPH CONSTRUCTION

Here, we describe the OBIO ontology (Section 3.1), and the construction and validation of the OKG in terms of the ontology (Section 3.2).

3.1 The Ontology

The goal of the ontology we build is the inclusion of fine-grained semantics, as a semantic layer on top of the tweet texts. We show five example tweets, use these to describe our ontological requirements which then inform the ontology creation process.

3.1.1 Motivating Examples. We present 5 sentences from tweets in Table 1 to show the type of analyses we aim to do with our ontology. Relevant content for the understanding of these sentences can be divided in three categories: **(1) tweet metadata, (2) meaning and (3) grammar**.

The **tweet metadata** category would include information such as the user who posted the content, information on the user, the date of the tweet, etc. It would also include standards metrics such as the number of likes.

[1] https://www.w3id.org/okg/obio-ontology/
[2] https://github.com/muhai-project/okg_media_discourse
[3] https://doi.org/10.5281/zenodo.10034210
[4] https://api.druid.datalegend.net/datasets/lisestork/OKG/services/OKG/sparql
[5] https://triply.cc
[6] https://iisg.amsterdam

Table 1: Sentences from tweets on inequality.

Id	Sentence Content
1	*"We see the end of the **trauma** created by our brutal system of race and **gender inequality** "*
2	*"These **unregulated systems** could cause **discrimination** on a massive scale says **Buolamwini**"*
3	*"How're we expected to behave **rationally** in the face of **brutality, inequality,** and **racism** when they get "scared" of a black person reaching for their wallet at a traffic stop."*
4	*"Nothing wrong with wanting to end **inequality**."*
5	*"I'm glad you think we should judge people on **merit**."*

The **meaning** category would typically include entities from sentences, such as *inequality* or *Buolamwini*. In our approach, we aim to go beyond the mere enumeration of entities in a tweet, and to add another semantic layer on top, specifically extracted rolesets: verbs and their arguments. PropBank [20] details provide such fine-grained analyses. PropBank, short for the Proposition Bank, is a linguistic resource that associates verbs with their arguments, also called semantic roles. An instantiated ensemble with a verb and its filled arguments is called a roleset. For example: sentence 2 from Table 1, the verb *"cause"* triggers a roleset, that has *"unregulated systems"* as subject and *"discrimination"* as object. Likewise in sentence 5, *"think"* triggers a roleset that has *"we should judge people on merit"* as object.

Lastly, the **grammar** category represents the grammatical structure of the sentences, more specifically the dependency relationships. Other than facilitating the entity extraction process, the analysis can provide interesting insights like the nesting of PropBank rolesets, as is the case in sentence 4, which triggered two rolesets: *"wanting"* and *"end"*.

3.1.2 Ontological Requirements (ORs). We derive three high-level ontological requirements from the examples above, directly linked to the three types of analyses that our ontology should enable.

OR1: Model the metadata of the social media ecosystem.
(1) Distinguish between a regular tweet, a repost and a reply.
(2) Each tweet should have a date.
(3) Metrics: sentiment, polarity, subjectivity, number of repost, number of likes.
(4) User attributes: account verified, number of followers, number of accounts followed, location.

OR2: Model the meaning of the tweets' content.
(1) Link tweets to extracted entities, and link them to external ontologies.
(2) Link tweets to PropBank extracted rolesets: triggering verb and its arguments.

OR3: Model the grammar of the tweets' content.
(1) Tweet content should be chunked down by sentences and tokens.
(2) Dependency relations should be added between tokens.
(3) Include: lemma, part-of-speech tag, token index.

3.1.3 Ontology creation. Following best practices in ontology development [11], we aim to re-use existing models and extend them with classes and object properties. Our core starting point is TweetKB [14], that uses the SIOC [5] ontology. We create new classes, data and object properties if they do not already exist. Apart from TweetsKB, we mainly integrate two other ontologies: NIF [18] for integrating text with KGs and Framester [16] for PropBank-related information. The prefix for the OBIO is `obio`[7].

Ontology Presentation. . We show the ontology in Figures 1 and 2. Figure 2 mainly covers OR2-2, while Figure 1 covers the rest. The prefixes are given in the Figures. We describe below how we encode the ontological requirements:

OR1-1: We introduce `obio:RePost` and `obio:Reply` as subclasses of `sioc:Post` for reposts and replies. A repost, or a retweet in the Twitter context, is to share another user's tweets to your followers, while a reply is a direct response to another tweet.

OR1-2: We re-use TweetsKB for this part, with `dc:created`.

OR1-3: We introduce the `obio:post_metrics` data property for metrics related to tweets. We define sub-properties of this property for the differents metrics that are listed: `obio:sentiment_label`, `obio:polarity_score`, `obio:subjectivity_score`, `obio:nb_repost`, and `obio:nb_like`.

OR1-4: We introduce various data properties to describe the following attributes for a user: `obio:is_verified`, `obio:description`, `obio:location`, `obio:follower`, and `obio:following`.

OR2-1: We re-use TweetsKB with `schema:mentions`.

OR2-2: We integrate text data with Propbank annotations and Framester. We attempt to stay as close as possible to the original Propbank annotations in Framester, with minor modifications[8]. For one annotation of a Propbank roleset, we re-use the `wsj:CorpusEntry` class. The main difference is that instead of using blank nodes to describe mapped roles, we use classes from the NIF ontology, such as `nif:Word` and `nif:Phrase`. Lastly, similarly to the existing Datatype Property `wsj:onLemma`[9], we create an Object Property `obio:onToken` that goes from a `wsj:CorpusEntry` to a `nif:String`. To enforce a better integration between OR2-1 and OR2-2, we add additional `nif:superString` links, as is shown in Figure 2.

OR3-1: We re-use the classes `nif:Word` and `nif:Sentence` from the NIF ontology.

OR3-2: We add `obio:dependency_relation` as a sub-property of `nif:inter` to describe the dependency relations between two words in a sentence. Each dependency relation is then added as a sub-property of `obio:dependency_relation`.

OR3-3: We mostly re-use content from the NIF ontology, and add the data property `obio:hasTokenIndex` to link a token to its index in the original tweet.

Together with the code that we submit with this paper, we add a more detailed documentation and visualization of our ontology.

[7] Short for https://www.w3id.org/okg/obio-ontology/.
[8] See http://etna.istc.cnr.it/framesterpage/wsj/wsjpropnetannotations/CE_64700.
[9] See http://etna.istc.cnr.it/framesterpage/wsj/onLemma.

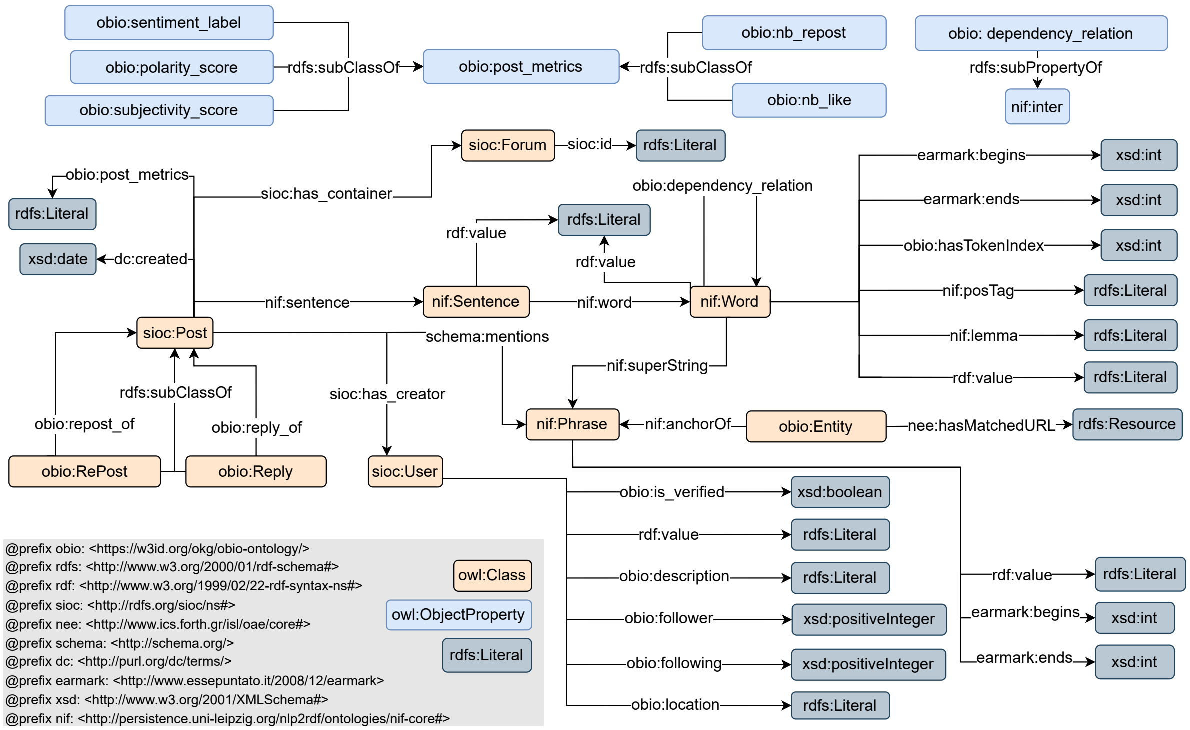

Figure 1: The Observatory Integrated Ontology.

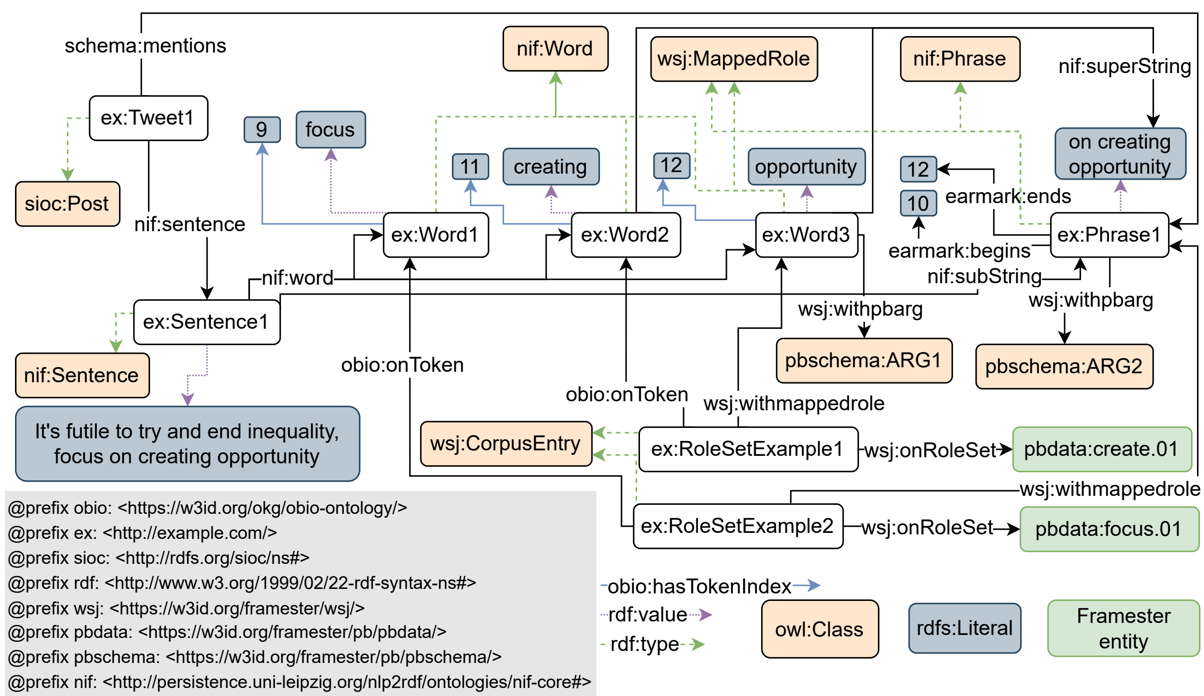

Figure 2: Framester integration.

3.2 Ontology Population and Validation

3.2.1 Tweet, metadata and grammar extraction (OR1 & OR3). The acquisition of social media data about inequality from the Twitter platform was done through the "academictwitteR" R library [3], using the Full-Archive API V2[10] with the following query parameters: "*(inequality OR inequalities) lang:en*". We downloaded the data before the changes of policies[11] in June 2023. We retrieved tweets and retweets published from the end (30th) of May 2020 to the beginning (1st) of May 2023. In this paper, we use a sample published from May 30th to August 27th, 2020. To be compliant with the Twitter policies, we remove user metadata and the texts of tweets and tweet sentences. We also replace user IDs with skolem IRIs through skolemization[12].

For the grammar of the tweet content, we used the output of spaCy[13], that covers all requirements for OR3. We used the `en_core_web_sm` model[14]. For the metadata of the tweets, all requirements of OR1 except OR1-3 were provided by the Twitter API: type of tweet (OR1-1), date (OR1-2) and user attributes (OR1-4). For the sentiment extraction (OR1-3), we use the RoBERTa-base model trained on around 124M tweets from January 2018 to December 2021 directly from Hugging Face[15], which achieved 69.1% accuracy for sentiment analysis. For the polarity and subjectivity metrics (OR1-3), we use TextBlob[16].

3.2.2 Meaning extraction (OR2).

Entities from text (OR2-1). We use the same methodology as [33]. We first extract two types of named entities from the output of spaCy: named entities and noun phrases. A named entity refers to a proper name, whereas a noun phrase is a grammatical construction that includes a noun and its modifiers. These entity mentions are then consolidated into `obio:Entity`, and the mapping to DBpedia is done through DBpedia Spotlight [25].

PropBank Rolesets (OR2-2). To link the tweets to the PropBank rolesets (OR2-2), we use an extended version of the PropBank grammar developed by [4], that uses computational construction grammar to extract semantic frames from text corpora. Such a grammar uses as a basis constructions, that are structured meaning-form pairs. The grammar output the verbs and their semantic roles. We then link the extracted frames to Framester [16]. As an example, the frame `protest.01` from [4] corresponds to the Framester entity `pbdata:protest.01`. For each frame, we create a `wsj:CorpusEntry` as illustrated in Figure 2.

3.2.3 Validation and statistics.

KG statistics. The KG we present in this paper contains 9,243,293 triples and 1,084,882 unique entities. Out of the 10,613 `obio:Entity` entities, 3,592 (33.8%) had a mapping to DBpedia. The average node indegree and outdegree are 8.4 and 8.5 respectively. The minimum, mean and maximum number of entities per tweets are 1, 2.3 and 11 respectively. 2,398 distinct frames were extracted across all tweets. The top 10 extracted frames were `do.02` (8,029), `pandemic.01` (2,067), `need.01` (1,675), `work.01` (1,674), `see.01` (1,583), `do.01` (1,268), `solve.01` (1,242), `address.02` (1,212), `say.01` (1,179), and `fight.01` (1174).

We describe the distribution of class types in Table 2. The most prevalent classes in OKG come from the NIF ontology, which is expected since each tweet was chunked down into sentences and tokens. OKG introduces 136,391 new corpus entries for PropBank rolesets. There are 62,015 original posts, 34,551 reposts and 4,837 replies for a total number of 42,108 different users.

The KG contains 22,852 tweets with a negative label, 19,070 with a neutral one and 3,927 with a positive one. This represents a ratio of around 6 between the negatively labeled tweets and the positively labeled ones. Furthermore, the average numbers of reposts are 7,015, 827 and 2,769 for the negatively, neutral and positively labeled tweets respectively, which represent a ratio of around 3 between the positive and negative ones. Likewise, the average numbers of like are 2.8, 1.8 and 2.1 respectively, with a ratio of 1.3. We observe the negatively labeled tweets tend to be more numerous, to get more repost and more likes than the other tweets.

Table 2: KG Class Distributions.

Class	Number
`nif:String`	787,187
`nif:Word`	573,377
`wsj:MappedRole`	277,502
`wsj:CorpusEntry`	136,391
`nif:Phrase`	107,756
`nif:Sentence`	106,054
`sioc:Concept`	104,123
`sioc:Post`	62,015
`sioc:Forum`	45,472
`sioc:Container`	45,472
`sioc:User`	42,108
`foaf:OnlineAccount`	42,108
`obio:RePost`	34,551
`obio:Entity`	10,613
`obio:Reply`	4,837

Data validation. The resulting graph was validated against a set of data quality criteria [7], specifically *accuracy* and *consistency* given that data was integrated via automated scripts. To check the graph we used SPARQL queries and SHACL shapes. For *accuracy*, we checked the syntactic validity of all literals using a set of SHACL shapes, and a SPARQL ASK query was created to check whether matches to DBpedia were valid URIs, whether all words were indeed included in their superstrings (`nif:superString`), and whether instances of `pbschema:ARG1` and `pbschema:ARG2` were valid Framester IRIs. For *consistency*, we checked for schema correctness using a set of SHACL shapes, e.g., every tweet has exactly one creator. Errors found through SPARQL and SHACL validation were corrected to improve the quality of the graph. The SPARQL queries and SHACL shapes can be found in the Github repository[17].

[10]https://developer.twitter.com/en/docs/twitter-api/data-dictionary/object-model/tweet
[11]https://developer.twitter.com/en/developer-terms/policy#4-d
[12]https://www.w3.org/TR/rdf11-concepts/#section-skolemization
[13]https://spacy.io
[14]Details on accuracy performance can be found at https://spacy.io/models/en.
[15]https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest
[16]https://textblob.readthedocs.io/en/dev/
[17]https://github.com/muhai-project/okg_media_discourse

4 USE CASES

In this section, we present five use cases that reflect examples of questions researchers and policymakers can answer with the OKG. We first present relevant questions that can be answered using sentiment analysis and named entity recognition, similarly to other Twitter resources, e.g., TweetsKB [14]. Second, we present questions solely facilitated by the OKG, through the integration of tweet metadata with finer-grained semantic layers such as PropBank rolesets and dependency relationships. For each use case, we describe their relevance and show the SPARQL queries used to to each use case. For readability we omit prefixes, which can be found in Figures 1 and 2.

4.1 Top Entities Grouped per Tweet Sentiment

The following SPARQL query lists the frequency of sentiment labels per entity mention:

```
SELECT ?entity (COUNT(?t) as ?nb_t) WHERE {
    ?t schema:mentions ?entityMention ;
       obio:sentiment_label ?label .
    ?entity nif:anchorOf ?entityMention .
}
GROUP BY ?entity ?label
ORDER BY DESC(?nb_t)
```

Listing 1: Top entities per tweet label.

Table 3 shows the top 10 entities mentioned in tweets per label. Some entities are mentioned across all sentiment labels, such as "Economic Inequality", "Racism", or "Poverty". Other entities from the top 10 only appear in one type of tweets, like "Donald Trump" or "Capitalism" for the negatively labeled tweets, or "Coronavirus Disease" for the positive ones.

4.2 Entities co-occurrence grouped per tweet sentiment

The following SPARQL query lists co-occurrence of entity mentions, grouped per sentiment label:

```
SELECT ?sent_label ?ent_1 ?url_1 ?ent_2 ?url_2
       (COUNT(?t) as ?nb_t ) WHERE {
?t schema:mentions ?ent_m_1, ?ent_m_2 ;
   obio:sentiment_label ?sent_label .
?ent_1 nif:anchorOf ?ent_m_1 .
?ent_2 nif:anchorOf ?ent_m_2 .
OPTIONAL {?ent_1 nee:hasMatchedURL ?url_1 .}
OPTIONAL {?ent_2 nee:hasMatchedURL ?url_2 .}
FILTER (STR(?ent_1) < STR(?ent_2))
}
GROUP BY ?sent_label ?ent_1 ?url_1 ?ent_2 ?url_2
ORDER BY DESC(?nb_t)
```

Listing 2: Pairs of entities co-occurring.

Table 4 shows the top 5 pairs of entities that co-occur in tweets, grouped per sentiment label. Among the negatively labeled tweets, "Economic Inequality" appears in nearly all the pairs, related to other abstract concepts such as "Racism" and "Poverty". In neutral tweets, entities that appear the most are "Scientific American Mind", "Happiness" and "Health". In positive tweets, the set of entities is more diverse, with no clear repetitions.

As presented in Section 3, our ontology was extended from TweetsKB [14]. OKG can consequently be used for similar use cases than the ones that make use of both metadata information and entities extracted from tweets. For the next uses, we rather focus on the analysis that are enabled by the integration of the PropBank rolesets and the dependency relationships.

4.3 PropBank Rolesets per tweet sentiment

The following SPARQL query lists rolesets found in tweets, grouped per sentiment label:

```
SELECT ?sent_label ?rs_pb
      (COUNT(?rs_inst) as ?nb_rs) WHERE {
?rs_inst wsj:onRoleSet ?rs_pb ;
         obio:onToken ?lu_token .
?sent nif:word ?lu_token .
?t nif:sentence ?sent ;
   obio:sentiment_label ?sent_label .
}
GROUP BY ?sent_label ?rs_pb
ORDER BY DESC(?nb_rs)
```

Listing 3: Number of PropBank rolesets per sentiment label.

Table 5 shows the top 10 frames that appear in tweets, grouped per sentiment label. There are differences across the sentiment labels. Whereas the frames for the negative labels relate more to (needed) actions or observations, such as `solve.01` or `need.01`, the frames for both the neutral and positive labels seem more motivational, such as `support.01`, `fight.01` or `thank.01`. Some frames most frequently appear regardless of the sentiment label, such as `do.02` and `work.01`.

4.4 Semantic Roles linked to Entities

The following SPARQL query lists rolesets linked to entity mentions:

```
SELECT ?rs_pb ?ent ?pbarg
       (COUNT(?rs_inst) as ?nb_rs) WHERE {
VALUES ?link_ss {nif:word nif:subString}
?rs_inst wsj:onRoleSet ?rs_pb ;
         obio:onToken ?lu_token ;
         wsj:withmappedrole ?role_string .
?sent ?link_ss ?role_string .
?t nif:sentence ?sent .
?role_string wsj:withpbarg ?pbarg .

?t schema:mentions ?ent_m .
?role_string nif:superString ?ent_m .
?ent nif:anchorOf ?ent_m .
}
GROUP BY ?rs_pb ?ent ?pbarg
ORDER BY DESC(?nb_rs)
```

Listing 4: Linking semantic roles to entities.

The output of the query is shown in Table 6, which shows the semantic roles that contain the most entities. For instance, "Global Warming" is associated to the roleset `change.01` 112 times in the dataset, with argument `ARG1`. In PropBank, the argument role `ARG1` often represents the "Theme" or "Patient" of a predicate, whereas `ARG0` typically represents the "Agent" or "Experiencer". Most of the entities are strongly related to issues having to do with inequality: various domains of inequality, such as housing or economic inequality, global warming and racial segregation. Lastly, the number of occurrences is not that high compared to the size of OKG, and in particular the size of the corpus entries and the entities. One explanation is that the integration of the entities and the semantic

Table 3: Top 10 entities in tweets, grouped per label. Freq. corresponds to the frequency of occurrence, and Perc. refers to the percentage of occurrence compared to the total number of tweets.

Negative			Neutral			Positive		
Entity	Freq.	Perc.	Entity	Freq.	Perc.	Entity	Freq.	Perc.
Economic Inequality	609	2.7	Economic Inequality	231	1.2	Racism	75	1.9
Racism	458	2.0	Racism	193	1.0	Economic Inequality	51	1.3
Poverty	234	1.0	Poverty	76	0.4	Poverty	25	0.6
America	117	0.5	Pandemic	63	0.3	Gender Inequality	21	0.5
Institutional Racism	113	0.5	Scientific American Mind	52	0.3	Black Lives Matter	20	0.5
Pandemic	111	0.5	Gender Inequality	51	0.3	Institutional Racism	19	0.5
Donald Trump	99	0.4	Severe Acute Respiratory Syndrome Coronavirus 2	49	0.3	Coronavirus Disease	19	0.5
Gender Inequality	96	0.4	Black Lives Matter	49	0.3	Podcast	14	0.4
Capitalism	94	0.4	Institutional Racism	48	0.3	United Kingdom	13	0.3
Social Inequality	88	0.4	United Kingdom	48	0.3	Web Conferencing	13	0.3

Table 4: Top 5 pair of entities co-occurring in tweets, grouped per sentiment label. Freq. corresponds to the frequency of occurence.

Negative			Neutral			Positive		
Entity 1	Entity 2	Freq.	Entity 1	Entity 2	Freq.	Entity 1	Entity 2	Freq.
Economic Inequality	Racism	42	Scientific American Mind	Terms	45	Coronavirus Disease	Economist	11
Economic Inequality	Poverty	27	Scientific American Mind	Happiness	45	Institutional Racism	Sociology	5
Capitalism	Economic Inequality	23	Scientific American Mind	Health	45	Hurricane Floyd	Minnesota	4
Poverty	Racism	21	Health	Terms	45	130 Crore Indians	British Association For Immediate Care	4
Economic Inequality	Institutional Racism	20	Happiness	Terms	45	British Association For Immediate Care	Modi Govt	4

Table 5: Top 10 PropBank rolesets from tweets, grouped per sentiment label. Freq.:frequency.

Negative		Neutral		Positive	
Roleset	Freq.	Roleset	Freq.	Roleset	Freq.
do.02	7369	do.02	1290	do.02	582
pandemic.01	1690	need.01	1002	work.01	361
see.01	1256	address.02	809	thank.01	280
do.01	1230	work.01	777	fight.01	278
solve.01	1120	fight.01	746	attack.01	277
need.01	1052	support.01	699	admire.01	276
work.01	961	use.01	605	see.01	257
expect.01	933	change.01	573	support.01	222
cut.02	932	pandemic.01	570	love.01	200
say.01	913	do.01	569	help.01	189

Table 6: Top 10 entities included in PropBank rolesets. Freq.: frequency, Neg./Neut./Pos.: negative/neutral/positive.

Roleset	Entity	Pbarg	Freq.	Neg.	Neut.	Pos.
change.01	Global Warming	ARG1	112	64	38	10
act.02	United States Congress	ARG0	38	38	0	0
right.05	Human Rights	ARG1	32	20	8	4
understand.01	I Understand 1941 Song	ARG0	24	16	6	2
end.01	Ibm	ARG0	22	8	14	0
grow.01	Economic Inequality	ARG1	20	18	2	0
segregate.01	Racial Segregation	ARG3	18	10	8	0
house.01	Housing Inequality	ARG1	18	8	6	4
warm.01	Global Warming	ARG1	16	10	6	0
work.01	All Facial Recognition Work	ARG1	16	6	10	0

roles can be further refined, as there are cases where entities are substrings of arguments that are not yet linked.[18]

4.5 Relationships between rolesets

The following SPARQL query lists pairs of rolesets and their dependency relations:

```
SELECT ?rs_pb_1 ?rs_pb_2 ?dep_prop
       (COUNT(?sent) as ?nb_s) WHERE {
?rs_inst_1 wsj:onRoleSet ?rs_pb_1 ;
           obio:onToken ?lu_token_1 .
?rs_inst_2 wsj:onRoleSet ?rs_pb_2 ;
           obio:onToken ?lu_token_2 .
?sent nif:word ?lu_token_1 , ?lu_token_2 .
?dep_prop rdfs:subPropertyOf obio:dependency_relation .
?lu_token_1 ?dep_prop ?lu_token_2 .
FILTER(!CONTAINS(str(?dep_prop), "dependency_relation"))
FILTER(!CONTAINS(str(?dep_prop), "aux"))
}
GROUP BY ?rs_pb_1 ?rs_pb_2 ?dep_prop
ORDER BY DESC(?nb_s)
```

Listing 5: Dependency relationships between rolesets.

[18]We plan to release a larger scale dataset later, which will include more rolesets.

Table 7 shows the frames that were appearing the most in tweets with a direct dependency relationship, with the number of occurrences and their distribution across the sentiment labels. Unlike Table 6, where most of the content came from the negative tweets,

Table 7 shows that some rolesets are specifically associated with positive or neutral tweets, despite their lower number. This is the case for `study offers`, `get` and `provide`, and `offering suggestions`.

In this section, we provided examples of use cases enabled by OKG. Use case 4.1 and 4.2 first provided examples of use cases analyzing entities, co-occurrences of entities and sentiment analysis, outlining which actors/objects appear the most in tweets, and are more likely to play an important roles in the tweets' narratives.

Table 7: Most frequent relationships between rolesets.

Roleset 1	Roleset 2	Dependency	Freq.	Neg.	Neut.	Pos.
right.05	human.02	amod	62	40	16	6
see.01	build.01	ccomp	44	32	12	0
take.01	act.02	dobj	44	22	12	10
cut.02	educate.01	compound	40	40	0	0
offer.01	study.01	nsubj	26	0	4	22
get.01	provide.01	conj	26	0	4	22
rise.01	call.02	compound	26	0	24	2
help.01	move.01	ccomp	26	0	4	22
offer.01	suggest.01	dobj	26	0	4	22
address.02	issue.02	dobj	26	12	12	2

We then provided use cases to gain a deeper understanding of opinions about entities using finer-grained semantic layers like PropBank rolesets and dependency relationships. In use case 4.3, we analyzed commonly used rolesets in tweets, revealing that neutral and positive tweets often contain motivational verbs. In use case 4.4, we explored the entities frequently appearing in PropBank rolesets, uncovering connections to various forms of inequality, such as housing inequality and global warming. In use case 4.5, we examined the most frequent dependency relationships between PropBank rolesets, discovering links between specific rolesets in positive tweets, despite their low proportion. These use cases demonstrate how integrating tweets with finer-grained parsing allows us to analyze the roles of specific entities in events. Similar insights can be obtained for real-world use cases such as the understanding of online debates on COVID-19 vaccination [9].

5 CONCLUSION

We first present the *OBservatory Integrated Ontology (OBIO)* that integrates social media metadata with various types of linguistic knowledge such as entities and PropBank rolesets. We then populate this ontology with the *Observatory Knowledge Graph (OKG)*, with tweets extracted on the topic of inequality. We lastly present several use cases that show how adding finer-grained semantic layer can help improve the overall understanding on social media discourse. The paper focuses on the topic of inequality but the method is generic and can be applied to other topics.

The work we present is based on a small sample of tweets to emphasise the usefulness of semantic layers. We plan to release a larger-scale KG in the future. Moreover, we aim at further curating the output of the natural language processing, as well as to evaluate the utility of the OKG in a real-world use case with expert users, such as social scientists.

Lastly, we plan to integrate better the entities and the PropBank rolesets to extract further information from the latter. Since we use the NIF ontology, some parts of the OKG remain in text format, hence we aim to improve the current representations within the KG.

ACKNOWLEDGMENTS

This work was funded by the European MUHAI project (Horizon 2020 research and innovation program) under grant agreement number 951846, the Sony Computer Science Laboratories-Paris, the Vrije Universiteit Amsterdam, the University of Bremen, and the Venice International University. C.S. acknowledges financial support from PON R&I 2014–2020 (FSE REACT-EU). We thank Frank van Harmelen, Annette ten Teije and Ilaria Tiddi for fruitful discussions.

REFERENCES

[1] Mehwish Alam, Aldo Gangemi, Valentina Presutti, and Diego Reforgiato Recupero. 2021. Semantic role labeling for knowledge graph extraction from text. *Progress in Artificial Intelligence* 10 (2021), 309–320.
[2] Mehwish Alam, Manuel Kaschura, and Harald Sack. 2020. Apollo: Twitter Stream Analyzer of Trending Hashtags: A case-study of# COVID-19.. In *ISWC (Demos/Industry)*. 64–69.
[3] Christopher Barrie and Justin Chun-ting Ho. 2021. academictwitteR: an R package to access the Twitter Academic Research Product Track v2 API endpoint. *Journal of Open Source Software* 6, 62 (2021), 3272.
[4] Katrien Beuls, Paul Van Eecke, and Vanja Sophie Cangalovic. 2021. A computational construction grammar approach to semantic frame extraction. *Linguistics Vanguard* 7, 1 (2021), 20180015.
[5] John G Breslin, Stefan Decker, Andreas Harth, and Uldis Bojars. 2006. SIOC: an approach to connect web-based communities. *International Journal of Web Based Communities* 2, 2 (2006), 133–142.
[6] Vasco Campos, Ricardo Campos, Pedro Mota, and Alípio Jorge. 2022. Tweet2Story: a web app to extract narratives from twitter. In *European Conference on Information Retrieval*. Springer, 270–275.
[7] Gustavo Candela, Pilar Escobar, Rafael C Carrasco, and Manuel Marco-Such. 2022. Evaluating the quality of linked open data in digital libraries. *Journal of Information Science* 48, 1 (2022), 21–43.
[8] Emily Chen, Kristina Lerman, Emilio Ferrara, et al. 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR public health and surveillance* 6, 2 (2020), e19273.
[9] Qingqing Chen and Andrew Crooks. 2022. Analyzing the vaccination debate in social media data pre-and post-COVID-19 pandemic. *International Journal of Applied Earth Observation and Geoinformation* 110 (2022), 102783.
[10] Yiyi Chen, Harald Sack, and Mehwish Alam. 2022. Analyzing social media for measuring public attitudes toward controversies and their driving factors: a case study of migration. *Social network analysis and mining* 12, 1 (2022), 135.
[11] G Cota et al. 2020. The landscape of ontology reuse approaches. *Appl. Practices Ontol. Des., Extraction, Reason* 49 (2020), 21.
[12] Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke Van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. 2015. Analysis of named entity recognition and linking for tweets. *Information Processing & Management* 51, 2 (2015), 32–49.
[13] Dimitar Dimitrov, Erdal Baran, Pavlos Fafalios, Ran Yu, Xiaofei Zhu, Matthäus Zloch, and Stefan Dietze. 2020. Tweetscov19-a knowledge base of semantically annotated tweets about the covid-19 pandemic. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2991–2998.
[14] Pavlos Fafalios, Vasileios Iosifidis, Eirini Ntoutsi, and Stefan Dietze. 2018. Tweetskb: A public and large-scale rdf corpus of annotated tweets. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*. Springer, 177–190.
[15] Antske Fokkens, Aitor Soroa, Zuhaitz Beloki, Niels Ockeloen, German Rigau, Willem Robert Van Hage, and Piek Vossen. 2014. NAF and GAF: Linking linguistic annotations. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*. 9–16.
[16] Aldo Gangemi, Mehwish Alam, Luigi Asprino, Valentina Presutti, and Diego Reforgiato Recupero. 2016. Framester: A wide coverage linguistic linked data hub. In *Knowledge Engineering and Knowledge Management: 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings 20*. Springer, 239–254.
[17] Aldo Gangemi, Valentina Presutti, Diego Reforgiato Recupero, Andrea Giovanni Nuzzolese, Francesco Draicchio, and Misael Mongiovì. 2017. Semantic web machine reading with FRED. *Semantic Web* 8, 6 (2017), 873–893.

[18] Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating NLP using linked data. In *The Semantic Web–ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II 12*. Springer, 98–113.

[19] Rupinder Paul Khandpur, Taoran Ji, Steve Jan, Gang Wang, Chang-Tien Lu, and Naren Ramakrishnan. 2017. Crowdsourcing cybersecurity: Cyber attack detection using social media. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1049–1057.

[20] Paul Kingsbury and Martha Palmer. 2003. Propbank: the next level of treebank. In *Proceedings of Treebanks and lexical Theories*, Vol. 3. Citeseer.

[21] Xiaohua Liu, Kuan Li, Ming Zhou, and Zhongyang Xiong. 2011. Collective semantic role labeling for tweets with clustering. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

[22] Roel O Lutkenhaus, Jeroen Jansz, and Martine PA Bouman. 2019. Mapping the Dutch vaccination debate on Twitter: Identifying communities, narratives, and interactions. *Vaccine: X* 1 (2019), 100019.

[23] Thomas Marcoux and Nitin Agarwal. 2021. Narrative Trends of COVID-19 Misinformation.. In *Text2Story@ ECIR*. 77–80.

[24] John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The Ontolex-Lemon model: development and applications. In *Proceedings of eLex 2017 conference*. 19–21.

[25] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*. 1–8.

[26] Alexander Pak, Patrick Paroubek, et al. 2010. Twitter as a corpus for sentiment analysis and opinion mining.. In *LREc*, Vol. 10. 1320–1326.

[27] Gerasimos Razis and Ioannis Anagnostopoulos. 2014. Semantifying twitter: The influence tracker ontology. In *2014 9th International Workshop on Semantic and Social Media Adaptation and Personalization*. IEEE, 98–103.

[28] Marco Rospocher, Marieke van Erp, Piek Vossen, Antske Fokkens, Itziar Aldabe, German Rigau, Aitor Soroa, Thomas Ploeger, and Tessel Bogaard. 2016. Building event-centric knowledge graphs from news. *Journal of Web Semantics* 37 (2016), 132–151.

[29] Hassan Saif, Yulan He, and Harith Alani. 2012. Semantic sentiment analysis of twitter. In *The Semantic Web–ISWC 2012: 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I 11*. Springer, 508–524.

[30] Carlo Romano Marcello Alessandro Santagiustina and Massimo Warglien. 2022. The architecture of partisan debates: The online controversy on the no-deal Brexit. *PLoS one* 17, 6 (2022), e0270236.

[31] Laura Sinikallio, Senka Drobac, Minna Tamper, Rafael Leal, Mikko Koho, Jouni Tuominen, Matti La Mela, Eero Hyvönen, et al. 2021. Plenary debates of the Parliament of Finland as Linked Open Data and in Parla-CLARIN markup. In *3rd Conference on Language, Data and Knowledge, LDK 2021*. Schloss Dagstuhl-Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing.

[32] Anders Søgaard, Barbara Plank, and Hector Alonso. 2015. Using frame semantics for knowledge extraction from twitter. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.

[33] Laura Spillner, Carlo R. M. A. Santagiustina, Thomas Mildner, and Robert Porzel. 2022. Towards conflictual narrative mechanics. In *Proceedings of the IJCAI/ECAI Workshop on Semantic Techniques for Narrative-based Understanding*.

[34] Lara Tavoschi, Filippo Quattrone, Eleonora D'Andrea, Pietro Ducange, Marco Vabanesi, Francesco Marcelloni, and Pier Luigi Lopalco. 2020. Twitter as a sentinel tool to monitor public opinion on vaccination: an opinion mining analysis from September 2016 to August 2017 in Italy. *Human vaccines & immunotherapeutics* 16, 5 (2020), 1062–1069.

[35] Ming-Hsiang Tsou, Chin-Te Jung, Chris Allen, Jiue-An Yang, Jean-Mark Gawron, Brian H Spitzberg, and Su Han. 2015. Social media analytics and research test-bed (SMART dashboard). In *Proceedings of the 2015 international conference on social media & society*. 1–7.

[36] Andranik Tumasjan, Timm Sprenger, Philipp Sandner, and Isabell Welpe. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the international AAAI conference on web and social media*, Vol. 4. 178–185.

[37] Akhila Sri Manasa Venigalla, Sridhar Chimalakonda, and Dheeraj Vagavolu. 2020. Mood of India during Covid-19-An interactive web portal based on emotion analysis of Twitter data. In *Conference companion publication of the 2020 on computer supported cooperative work and social computing*. 65–68.

[38] Tom Willaert, Paul Van Eecke, Katrien Beuls, and Luc Steels. 2020. Building social media observatories for monitoring online opinion dynamics. *Social Media+ Society* 6, 2 (2020), 2056305119898778.

[39] Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li, and Tao Wang. 2021. Multimodal relation extraction with efficient graph alignment. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5298–5306.

[40] Florian Zollmann, Jeffery Klaehn, Tina Sikka, Kristin Comeforo, Daniel Broudy, Mandy Tröger, Elizabeth Poole, Alison Edgley, and Andrew Mullen. 2018. The propaganda model and intersectionality: integrating separate paradigms. *Media Theory* 2, 2 (2018), 213–239.