Contextual Affective Analysis: A Case Study of People Portrayals in Online #MeToo Stories

Anjalie Field, Gayatri Bhat, Yulia Tsvetkov
Language Technologies Institute
Carnegie Mellon University
{anjalief, ytsvetko}@cs.cmu.edu

Abstract

In October 2017, numerous women accused producer Harvey Weinstein of sexual harassment. Their stories encouraged other women to voice allegations of sexual harassment against many high profile men, including politicians, actors, and producers. These events are broadly referred to as the #MeToo movement, named for the use of the hashtag "#metoo" on social media platforms like Twitter and Facebook. The movement has widely been referred to as "empowering" because it has amplified the voices of previously unheard women over those of traditionally powerful men. In this work, we investigate dynamics of sentiment, power and agency in online media coverage of these events. Using a corpus of online media articles about the #MeToo movement, we present a *contextual affective analysis*—an entity-centric approach that uses contextualized lexicons to examine how people are portrayed in media articles. We show that while these articles are sympathetic towards women who have experienced sexual harassment, they consistently present men as most powerful, even after sexual assault allegations. While we focus on media coverage of the #MeToo movement, our method for contextual affective analysis readily generalizes to other domains.[1]

1 Introduction

In 2006, Tarana Burke founded the #MeToo movement, aiming to promote hope and solidarity among women who have experienced sexual assault (Ohlheiser 2018). In October 2017, following waves of sexual harassment accusations against producer Harvey Weinstein, actress Alyssa Milano posted a tweet with the hashtag #MeToo and encouraged others to do the same. Her message initiated a widespread movement, calling attention to the prevalence of sexual harassment and encouraging women to share their stories.

Tarana Burke has described her primary goal in founding the movement as "empowerment through empathy."[2] However, mainstream media outlets vary in their coverage of these recent events, to the extent that some outlets accuse others of misappropriating the movement. For instance, in January 2018, Babe.net published an article written by Katie Way, describing the interaction between anonymous 'Grace' and famous comedian Aziz Ansari (Way 2018). The article sparked not only instant support for Grace, but also instant backlash criticizing Grace's lack of agency: "The single most distressing thing to me about this story is that the only person with any agency in the story seems to be Aziz Ansari" (Weiss 2018). One widely circulated article, written by Caitlin Flanagan and published in The Atlantic, strongly criticized Way's article and questioned whether modern conventions prepare women to fight back against potential abusers (Flanagan 2018).

The manner in which accounts of sexual harassment portray the people involved affects both the audience's reaction to the story and the way people involved in these incidents interpret or cope with their experiences (Spry 1995). In this work, we use natural language processing (NLP) techniques to analyze online media coverage of the #MeToo movement. In a *people-centric* approach, we analyze narratives that include individuals directly or indirectly involved in the movement: victims, perpetrators, influential commenters, reporters, etc. Unlike prior work focused on social media (Ribeiro et al. 2018; Rho, Mark, and Mazmanian 2018), our work examines the prominent role that more traditional outlets and journalists continue to have in the modern-era online media landscape.

In order to structure our approach, we draw from social psychology research, which has identified 3 primary affect dimensions: *Potency* (strength vs. weakness), *Valence* (goodness vs. badness), and *Activity* (liveliness versus torpidity) (Osgood, Suci, and Tannenbaum 1957; Russell 1980; 2003). Exact terminology for these terms has varied across studies. For consistency with prior work in NLP, we refer to them as **power**, **sentiment**, and **agency**, respectively (Sap et al. 2017; Rashkin, Singh, and Choi 2016). In the context of the #MeToo movement, these dimensions tie closely to the concept of "empowerment through empathy."

The crux of our method is in developing contextualized, entity-centric connotation frames, where polarity scores are generated for words in context. We generate token-level sen-

[1] We provide code and public data at https://github.com/anjalief/metoo_icwsm2019

[2] https://metoomvmt.org/

timent, power, and agency lexicons by combining contextual ELMo embeddings (Peters et al. 2018) with (uncontextualized) connotation frames (Rashkin, Singh, and Choi 2016; Sap et al. 2017) and use supervised learning to propagate annotations to unlabeled data in our #MeToo corpus. Following prior work, we first evaluate these models over held-out subsets of the connotation frame annotations. We then evaluate the specifics of our method, namely contextualization and entity scoring, through manual annotations.

We ultimately use these contextualized connotation frames to generate sentiment, power, and agency scores for entities in news articles related to the #MeToo movement. We find that while the media generally portrays women revealing stories of harassment positively, these women are often not portrayed as having high power or agency, which threatens to undermine the goals of the movement. To the best of our knowledge, this is the first work to introduce *contextual affective analysis*, a method that enables nuanced, fine-grained, and directed analyses of affective social meanings in narratives.

2 Background

We motivate the development of contextual affective analysis as a people-centric approach to analyzing narratives. Entity-centric models, which focus on people or characters rather than plot or events, have become increasingly common in NLP (Bamman 2015). However, most approaches rely on unsupervised models (Iyyer et al. 2016; Chambers and Jurafsky 2009; Bamman, O'Connor, and Smith 2013; Card et al. 2016), which can capture high-level patterns but are difficult to interpret and do not target specific dimensions.

In contrast, we propose an interpretable approach that focuses on power, sentiment, and agency. These dimensions are considered both distinct and exhaustive in capturing affective meaning, in that all 3 dimensions are needed, and no additional dimensions are needed; other affective concepts, such as anger or joy, are thought to decompose into these three dimensions (Russell 1980; 2003). Furthermore, these dimensions form the basis of *affective control theory*, a social psychological model which broadly addresses how people respond emotionally to events and how they attribute qualities to themselves and others (Heise 1979; 2007; Robinson, Smith-Lovin, and Wisecup 2006). Affective control theory has served as a model for stereotype detection in NLP (Joseph, Wei, and Carley 2017).

Furthermore, while automated sentiment analysis has spanned many areas (Pang and Lee 2008; Liu 2012)[3] analysis of power has been almost entirely limited to a dialog setting: how does person A talk to a higher-powered person B? (Gilbert 2012; Prabhakaran and Rambow 2017; Danescu-Niculescu-Mizil et al. 2012). Here, we focus on a *narrative* setting: does the journalist portray person A or person B as more powerful?

In order to develop an interpretable analysis that focuses on sentiment, power, and agency in narrative, we draw from existing literature on *connotation frames*: sets of verbs annotated according to what they imply about semantically dependent entities. Connotation frames, first introduced by Rashkin, Singh, and Choi (2016), provide a framework for analyzing nuanced dimensions in text by combining polarity annotations with frame semantics (Fillmore 1982). We visualize connotation frames in Figure 1 on the left. More specifically, verbs are annotated across various dimensions and perspectives, so that a verb might elicit a positive sentiment for its subject (i.e. sympathy) but imply a negative effect for its object. We target power, agency, and sentiment of entities through pre-collected sets of verbs that have been annotated for these traits:

- Perspective($writer \rightarrow agent$) – Does the writer portray the agent (or subject) of the verb as positive or negative?
- Perspective($writer \rightarrow theme$) – Does the writer portray the theme (or object) of the verb as positive or negative?
- Power – does the verb imply that the theme or the agent has power?
- Agency – does the verb imply that the subject has positive agency or negative agency?

For clarity, we refer to Perspective($writer \rightarrow agent$) as Sentiment($agent$) and Perspective($writer \rightarrow theme$) as Sentiment($theme$) throughout this paper.

These dimensions often differ for the same verb. For example, in the sentence: "She amuses him," the verb "amuses" connotes that *she* has high agency, but *he* has higher power than *she*. Rashkin, Singh, and Choi (2016) present a set of verbs annotated for sentiment, while Sap et al. (2017) present a set of verbs annotated for power and agency. However, these lexicons are not extensive enough to facilitate corpus analysis without further refinements. First, they contain only a limited set of verbs, so a given corpus may contain many verbs that are not annotated. Furthermore, verbs are annotated in synthetic context (e.g., "X amuses Y"), rather than using real world examples. Finally, each verb is annotated with a single score for each dimension, but in practice, verbs can have different connotations in different contexts.

Consider two instances of the verb "deserve":

1. The hero deserves appellation
2. The boy deserves punishment

In the first instance, annotators rate the writer's perspective towards the agent ("hero") as positive, while in the second instance, annotators rate the writer's perspective towards the agent ("boy") as negative (Rashkin, Singh, and Choi 2016). We can find numerous similar examples in articles related to the #MeToo movement, i.e. *She pushed him away* vs *Will one part of the movement's legacy be to push society to find the right words to describe it all?*.

The uncontextualized annotations presented by Rashkin, Singh, and Choi (2016) and Sap et al. (2017) serve as starting points for more in-depth analysis. We build uncontextualized features for verbs to match the uncontextualized annotations, and then use supervised learning to extend the uncontextualized annotations to verbs in context, thus learning contextualized annotations.

[3] http://alt.qcri.org/semeval2016/index.php?id=tasks

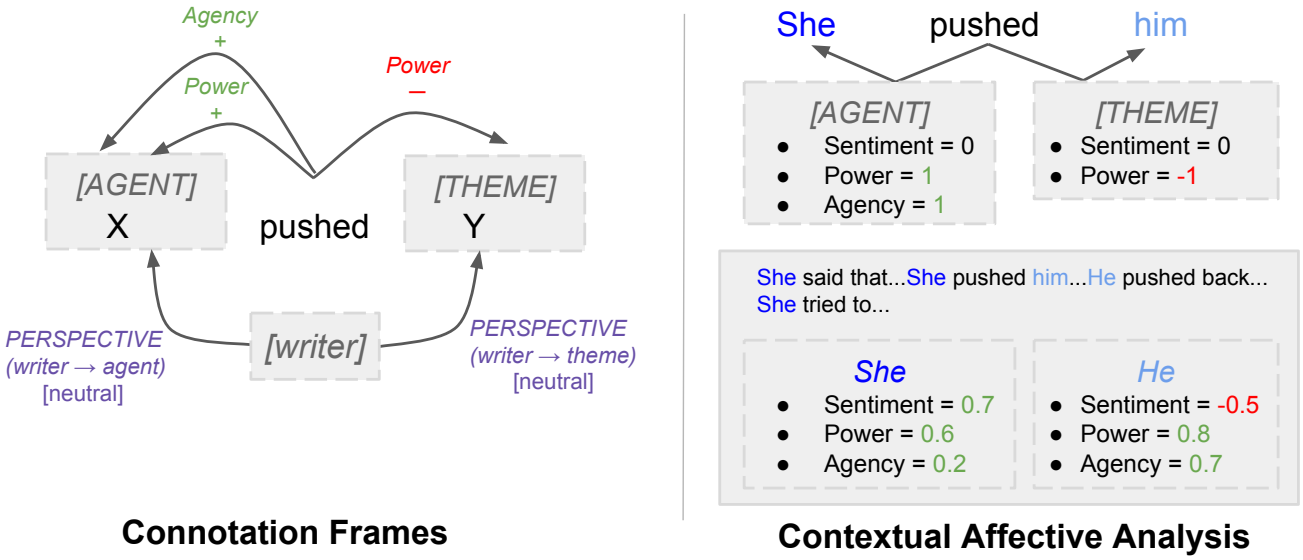

Figure 1: Left, we show off-the-shelf connotation frame annotations (Rashkin, Singh, and Choi 2016; Sap et al. 2017) for the verb "push". Right, we show the proposed adaptation. We adapt connotation frames from a verb-centric formalism to an entity-centric formalism, transferring scores from verbs to entities using a context-aware approach (top right). We then aggregate contextualized scores over all mentions of entities in a corpus (bottom right). This new approach—*contextual affective analysis*—enables us to obtain sentiment, power, and agency scores for entities in unannotated corpora and conduct extensive analyses of people portrayals in narratives, which we exemplify on #MeToo data.

Our work extends the concept of domain-specific lexicons: that words have different connotations in different situations. For instance, over the last century, the word "lean" has lost its negative association with "weakness" and instead become positively associated with concepts like "fitness." These changes in meaning have motivated research on inducing domain-specific lexicons (Hamilton et al. 2016). Using contextual embeddings (Peters et al. 2018), we extend this concept by introducing *context-specific* lexicons: we induce annotations for words in context, rather than just words in domain. To the best of our knowledge, this is the first work to propose contextualized affective lexicons. Our overall methodology uses these contextualized lexicons to obtain power, sentiment, and agency scores for entities in contextual affective analysis.

3 Methodology

In the proposed contextual affective analysis, our primary goal is to analyze how people are portrayed in narratives. To obtain sentiment, power, and agency scores for people in the context of a narrative (a sentence, paragraph, article, or an outlet), we adapt the connotation frames reviewed above as follows. First, since connotation frames are verb-centric—rather than people-centric—we define a mapping from a verb in a sentence to people that are syntactic arguments of the verb. We visualize this mapping in Figure 1. Second, since only a small subset of verbs (<30%) in our corpus is included in the annotations crowdsourced by Rashkin, Singh, and Choi (2016) and Sap et al. (2017), we devise a lexicon induction method to annotate unlabeled verbs using the existing seed of annotations. To contextualize the analysis, the lexicon induction procedure uses contextual ELMo embeddings (Peters et al. 2018). Contextual embeddings are a form of distributed word representations that incorporate surrounding context words. Thus, using the above example, ELMo embeddings provide different representations for "push" in "She pushed him away" and for "push" in "Will one part of the movement's legacy be to push..." In what follows, we detail the components of our methodology. All of our code is publicly available.

We use connotation frames, which are annotations on verbs, to obtain affective scores on people (the agent or theme of these verbs). In our running example "She pushes him away," "She" is the agent and "him" is the theme. Given a verb V, the verb's agent A, the verb's theme T, and a set of connotation frame annotations over V (e.g., $V_{Sentiment(agent)} = +1$), we obtain sentiment, power, and agency scores as follows:

$$
\begin{aligned}
Sentiment(A) &= V_{Sentiment(agent)} \\
Power(A) &= V_{Power} \\
Agency(A) &= V_{Agency} \\
Sentiment(T) &= V_{Sentiment(theme)} \\
Power(T) &= -V_{Power}
\end{aligned}
$$

We obtain a sentiment score within a corpus for an entity E by averaging over all $V_{Sentiment(agent)}$ scores where E is the agent of V and all $V_{Sentiment(theme)}$ scores where E

	Lexicon Size	Training Set Size
Sentiment	948	300
Power	1,714	571
Agency	2,104	701

Table 1: Annotated Lexicon Statistics

is the theme of V. We compute agency and power scores in the same way. Thus, the final entity scores are not merely direct mappings from verb annotations, but an aggregation of all such mappings over the corpus of entity mentions.

We obtain these verb scores by taking a supervised approach to labeling verbs in context with sentiment ($V_{Sentiment(agent/theme)} \in \{-1, 0, +1\}$), power, ($V_{Power} \in \{-1, 0, +1\}$), and agency ($V_{Agency} \in \{-1, 0, +1\}$).

For a given verb in our training set V, we assume that V occurs n times in our corpus, and enumerate these occurrences as $V_1 ... V_n$. For each V_i, we compute the contextualized ELMo embedding $\mathbf{e}_i$. We then "decontextualize" these embeddings by averaging over $\mathbf{e}_1 \dots \mathbf{e}_n$ to obtain a single feature representation $\mathbf{e}$. We consider these decontextualized embeddings to be representative of the off-the-shelf uncontextualized lexicons, and we use them as training features in a supervised classifier.

Then, for a given verb in our corpus T, for each instance where T occurs in our corpus, we use its contextualized ELMo embedding $\mathbf{e}_i$ as a feature to predict an annotation score for T_i. In particular, we use logistic regression with re-weighting of samples to maximize for the best average F1 score over a dev set.[4]

For evaluation, in order to compare our results against existing annotations, we use two methods to obtain uncontextualized annotations from the learned T_i scores for verbs in our test sets. In the first, which we refer to as *type-level*, we average all of the token-level embeddings ($\mathbf{e}_i$) in the test data in the same way as in the training data, and we learn a single annotation for each verb T, rather than learning contextualized annotations. This approach is most similar to prior work. In the second, which we refer to as *token-level*, we predict a separate score for each token-level embedding as described above, and we then take a majority vote over scores to obtain an overall score for each verb.

4 Experimental Setup

Lexicons We provide basic statistics of the annotated lexicons in Table 1. The sentiment frame annotations are reported averages over annotations from 15 crowdworkers. We ternerize these annotations using the same cut-offs as Rashkin, Singh, and Choi (2016): Negative: $[-1, -0.25)$, Neutral: $[-0.25, 0.25]$ and Positive: $(0.25, 1]$. The power frame annotations are formatted as [power_agent, power_equal, power_theme], which we map to $[1, 0, -1]$. Similarly, the agency verbs are formatted as [agency_positive, agency_equal, agency_negative], which we map to $[1, 0, -1]$.

Corpora We gathered a corpus of articles related to the #MeToo movement by first collecting a list of URLs of articles that contain the word "metoo" using an API that searches for articles from over 30,000 news sources.[5] Over two separate queries, we gathered URLs from November 2, 2017 to January 31, 2018 and from February 28, 2018 to May 29, 2018. Next, we used Newspaper3k to obtain the full text of each article.[6]

We then discarded any pages that returned 404 errors, any URLs containing the word "video," and any non-English articles, identifying languages using the Python package langdetect.[7] Finally, we removed duplicate articles by converting each article into a bag-of-words vector and computing the cosine distance between every pair of vectors. If the distance between 2 articles was less than a threshold (0.011, identified by manually examining random samples), we discarded the more recent article. Our final data set consists of 27,602 articles across 1,576 outlets, published between November 2, 2017 and May 29, 2018.

Our data collection method includes any articles that mention the word #MeToo, which includes articles focused on other events that only mention the movement in passing. However, we believe these articles are still relevant for our analysis, as mentioning any entities alongside the movement implicitly associates them with these events, and as discussed in §6, people not directly involved in events can become important entities in the movement.

Preprocessing We tokenize and sentence-split our corpus using the Stanford NLP pipeline. In generating 1024-dimensional ELMo embeddings, we only take embeddings for verbs, performing stemming and POS tagging using spaCy[8], and we keep only the 2nd (middle) ELMo embedding layer. In generating entity scores, we use dependency parsing to identify agents and themes: we consider an entity E to be an agent of V if it is the verb's subject. We consider an entity E to be a theme of V if it is the verb's object or passive subject ("nsubjpass"). We use the Stanford NLP pipeline for dependency parsing, named entity recognition, and co-reference resolution. We find a total of 3,132,389 entity-verb pairs across the corpus, which form the basis of our analysis.

5 Evaluation

We first evaluate our methods on their ability to predict the labels of off-the-shelf connotation frame lexicons. While this task is not our primary objective, it serves as a sanity-check on our feature representations and allows us to compare our method with prior work. Then, we evaluate the

[4] We experimented with other supervised and semi-supervised methods common in lexicon induction including graph-based semi-supervised label propagation and random walk-based propagation (Goldberg and Zhu 2006; Hamilton et al. 2016), but found that logistic regression outperformed these methods on all frames.

[5] https://newsapi.org/

[6] http://newspaper.readthedocs.io/en/latest/

[7] https://pypi.org/project/langdetect/

[8] https://spacy.io/

	Aspect	Frame	Type	Token
Sentiment ($theme$)				
Accuracy	67.56	67.56	63.33	66.67
Macro F1	56.18	56.18	55.63	51.90
Sentiment ($agent$)				
Accuracy	60.54	61.87	60.00	61.33
Macro F1	60.72	63.07	58.26	60.23
		Majority Class	**Type**	**Token**
Power				
Accuracy	-	69.66	69.66	69.49
Macro F1	-	27.37	55.97	54.10
Agency				
Accuracy	-	79.14	70.30	72.74
Macro F1	-	29.45	48.79	50.14

Table 2: Accuracy and F1 Score of lexicon expansion for our methods (Type and Token) compared with prior work (Aspect and Frame). Non-trivial F1 scores demonstrate that our feature representations capture meaningful information about sentiment, power, and agency.

methods in a contextualized setting, assessing their ability to model contextualized verb annotations (§5.1) and contextualized entity annotations (§5.2) by comparing with human annotators.

We divide the annotations into train, dev, and test; for sentiment, we use the same data splits as Rashkin, Singh, and Choi (2016); for power and agency, we randomly divide the lexicons into subsets of equal size. In order to compare the contextualized scores generated by our method with the off-the-shelf annotations, we aggregate the contextualized annotations into uncontextualized annotations for the verbs in the test set, as described in §3.

Table 2 reports results. For comparison, we show the Aspect-Level and Frame-Level models presented by Rashkin, Singh, and Choi (2016) over the sentiment annotations. Our type-level logistic regression is essentially identical to the aspect-level model, the primary difference being our use of ELMo embeddings. Our results are slightly lower, but generally comparable to the results reported by Rashkin, Singh, and Choi (2016); crucially, they are obtained with a model that ultimately allows us to incorporate context. The type-level and token-level aggregation methods perform about the same.

In the absence of prior work on this task for the power and agency lexicons, we show a majority class baseline. Our methods show a strong improvement over F1 scores for the majority class baseline. As for sentiment, the type-level and token-level methods perform similarly. Table 2 generally shows that ELMo embeddings capture meaningful information about power, agency, and sentiment. However, our primary task is not to re-create the word-level annotations in the connotation frame lexicons, but rather to contextualize these lexicons by obtaining instance-level scores over verbs in context. We evaluate these contextualized scores in the following section.

	Verb-level	Sent.-level	Sent.-level training
Sentiment (t)	41.05	44.35	50.16
Sentiment (a)	51.37	52.80	54.11

Table 3: F1 scores for using our method to score contextualized annotations. Predicting sentence-level scores outperforms predicting verb-level scores. Best performance is achieved by also using sentence-level training data, but this is unsustainable in practice.

5.1 Evaluation of Contextualization

In this section, we draw from the original annotations used to create the connotation frame lexicons in order to assess the impact of contextualization. The publicized connotation frames consist of a single score for each verb. However, for the sentiment dimensions, these scores were obtained by collecting annotations over verbs in a variety of simple synthetically-generated contexts and averaging annotations across contexts, i.e. collecting 5 annotations each for "the hero deserves appellation," "the student deserves an opportunity," and "the boy deserves punishment" and averaging across the 15 annotations (Rashkin, Singh, and Choi 2016). Then, for the sentiment lexicons, we can evaluate our method's ability to provide annotations in context by reverting to the original pre-averaged annotations. (We cannot perform the same evaluation for the power and agency lexicons, because they were created by collecting annotations over verbs without any context, i.e. "X deserves Y" (Sap et al. 2017).)

When we ignore context, meaning we treat all 15 annotations over each verb as annotations over the same sample, the inter-annotator agreement (Krippendorff's alpha) for Sentiment($theme$) is 0.20 and for Sentiment($agent$) is 0.28. However, when we treat each sentence as a separate sample (i.e. measuring agreement in annotations over "the hero deserves appellation" separately from annotations over "the boy deserves punishment"), the agreement rises to 0.34 and 0.40 respectively, a $> 40\%$ increase for each trait. The improvement in agreement demonstrates that when annotators disagree about the connotation implied by a verb, it is often because the verb has different connotations in different contexts.

We can then evaluate our method for contextualization by using these sentence-level annotations. More specifically, we use the same train, dev, and test splits as before. However, for verbs in the test set, instead of averaging all 15 annotations for each verb into a single score, we only average over annotations on the same sentence. Thus, our gold test data contains separate scores for "the hero deserves appellation", "the student deserves an opportunity", and "the boy deserves punishment", resulting in approximately 3 times as many test points as the test data in Table 2.

Table 3 shows the results of evaluating our method on

this contextualized test set. In the first column, Verb-level, our model disregards contextualization and predicts a single score for each verb using type-level aggregation, which is equivalent to the method used in Table 2. The primary difference is that in Table 2, we evaluate over uncontextualized annotations (i.e. a single score for "deserve"), while in Table 3, we evaluate over contextualized annotations.

In the second column, Sent.-level, we predict a separate score for each verb in context, rather than using token or type level aggregation over the test data. This column represents our primary method and is the method we use for analysis in §6. For both traits, this method outperforms the aggregation approach shown in the first column.

In the third column, Sent.-level training, we similarly predict a separate score for each context, but we further treat each context as a separate training sample. Thus we both train and evaluate on contextualized annotations. In the Sent.-level and Verb-level columns, we train on uncontextualized annotations, as described in §3.

While training on contextualized annotations achieves the best performance, it is difficult to generalize to other data sets. The sentences used for gathering these annotations were created using Google Syntactic N-grams and designed to be short generic sentences (Rashkin, Singh, and Choi 2016). Thus, they are much simpler than real sentences, and we would not expect them to serve as realistic training data in other domains. In order to use these connotation frame lexicons in a new domain, it would be necessary to annotate a new set of sentences, which defeats the usefulness of off-the-shelf lexicons. Instead, we focus on the second column, Sent.-level, as our primary method, since it is an improvement over existing ways of using off-the-shelf lexicons without requiring new annotations for every task. Overall, Table 3 demonstrates the usefulness of contextualization, as both contextualized approaches outperform the uncontextualized approach.

5.2 Evaluation of Entity Scores

In the previous sections, we evaluated our methods for generating verb annotations. In this section, we evaluate our methods for transferring verb annotations to entity scores, specifically focusing on power. In order to assess entity scoring, we devised an annotation task in which we asked annotators to read articles and rank entities mentioned. We then compare our entity scores against these annotations.

More specifically, we sampled 30 articles from our corpus that all contain mentions of Aziz Ansari. We then provided 2 annotators with a list of 23 entities extracted from these articles and asked the annotators to read each article in order. After every 5 articles, annotators ranked the listed entities by assigning a 1 to the lowest-powered entity, a 7 to the highest-powered entity, and scaling all other entities in between. In this way, since annotators rerank entities every 5 articles, we maximize the number of annotations we obtain, while minimizing the number of articles that annotators need to read. Furthermore, we ensure that we obtain different power scores for different entities by forcing annotators to use the full range of the ranking scale.

Off-the-shelf	**Frequency**	**Ours**
57.1	59.1	71.4

Table 4: Accuracy for scoring how powerful entities are, as compared with manual annotations. We calculate accuracy by assessing if the metric correctly answers "is entity A more powerful than entity B?". Our method outperforms both baseline metrics.

However, we note that this is a very subjective annotation task. Annotators specifically described it as difficult, both in deciding which entities were most powerful and in ranking entities based on the provided articles rather than on outside knowledge. We observed some of this subjectivity in the collected annotations: one annotator consistently ranked abstract entities like "The New York Times" as high-powered, while the other annotator consistently ranked these entities as low-powered. Thus, while we present results as an approximation of how well our methods work, we caution that further evaluation is needed.

From the annotations, we have a ranking for each entity for each 5-article step. Hypothesizing that some entities are more subjective to rank than others, we eliminate all samples where the difference in ranking between the two annotators is greater than 2. We are then left with 81 annotations. The correlation between annotators on this set is statistically significant (Spearman's R=0.55, p-value=1.03e-07). In the following analysis, we average the rank assigned by annotators to obtain a score for each entity.

We ultimately evaluate our methods through pairwise comparisons at each 5-article step. For every pair (A, B) of entities at each time step, we evaluate if entity A is scored as more powerful or less powerful than entity B. We discard samples where the entities were ranked as equal. We compare our method against 2 baseline metrics: (1) the frequency of the entity and (2) power scores assigned by the off-the-shelf connotation frames, rather than our contextualized frames.

Off-the-shelf connotation frames are limited to a subset of verbs in our corpus. Furthermore, our analysis pipeline is dependent on the named entity extraction and co-reference resolution tools used during pre-processing, which we find miss many entity mentions (for instance, when we manually extract entities in the first 10 articles, we identify 28 entities that occur at least 3 times, while the automated pipeline identifies only 4). For fair comparison between our method and the off-the-shelf lexicons, we discard entities that do not occur with at least 3 off-the-shelf power-annotated verbs, as identified by our preprocessing pipeline. After this filtering, we are left with 49 pairwise comparisons.

Table 4 reports results. Our method outperforms both baselines, correctly identifying the higher-ranked entity 71.4% of the time. We note that each annotator individually achieves at most 83.7% accuracy on this task, which suggests an upper limit on the achievable accuracy.

Furthermore, if we perform the same test, eliminating only entities that occur fewer than 3 times in the text, rather than mandating that entities occur with at least 3 off-the-

Most Positive	Most Negative
Kara Swisher	Bill Cosby
Tarana Burke	Harvey Weinstein
Meghan Markle	Eric Schneiderman
Frances McDormand	Kevin Spacey
Oprah Winfrey	Ryan Seacrest

Table 5: The most positively portrayed entities consist primarily of 3rd party commentators on events. The most negatively portrayed entities consist of men accused of sexual harassment.

shelf annotated verbs, our method achieves an accuracy of 63.01% over 73 pairs, while the frequency baseline achieves an accuracy of 53.42%.

In conducting this analysis, we observed that one of the limitations of the power annotations provided by Sap et al. (2017) is that the authors only annotate transitive verbs for power. They hypothesize that a power differential only occurs when an entity (e.g. the agent) has power over another entity (e.g. the theme). However, we do not limit our scoring to transitive verbs, hypothesizing that intransitive verbs can also be indicative of power, even if there is no direct theme. The improved performance of our scoring metric over the off-the-shelf baseline supports this hypothesis.

6 Analysis of Entity Portrayals in #MeToo Movement

In this section, we use the affect scores to analyze how people are portrayed in media coverage of the #MeToo movement. For reference, we provide brief descriptions of the entities mentioned and their connection to the movement in the Appendix. We propose a top-down framework for structuring a people-centric analysis with three primary levels:

1. Corpus-level: we examine broad trends in coverage of all common entities across the entire corpus
2. Role-level: we examine how people in similar roles across separate incidents are portrayed
3. Incident-level: we restrict our analysis to people involved in a specific incident

We present here only a subset of possible analyses. While we focus on the #MeToo movement, our methodology readily generalizes to other research questions and corpora, such as analyzing news coverage of scientific publications (MacLaughlin, Wihbey, and Smith 2018).

6.1 Corpus-Level

By examining portrayals at a corpus level, we can assess the overall media coverage of the #MeToo movement. Whom does the media portray as sympathetic? Did media coverage of events empower individuals?

We examine these questions by computing sentiment, power, and agency scores for the 100 most frequent proper nouns across the corpus, shown in part in Tables 5–7. For brevity, we omit redundant entities (i.e. Donald Trump and President Donald Trump). Table 5 shows the five most positively and negatively portrayed entities. Unsurprisingly, the most negatively portrayed entities all consist of men accused of sexual harassment, lead by Bill Cosby, the first man actually convicted in court following the wave of accusations in the movement. However, the most positively portrayed entities consist not of women voicing accusations or of men facing them, but rather of 3rd party commentators, who were outspoken in their support of the movement. None of these entities were directly involved in cases that arose out of the #MeToo movement, but all of them made widely-circulated comments in support of the accusers.

Highest Power	Lowest Power
The #MeToo movement	Kevin Spacey
Judge Steven O'Neill	Andrea Constand
The New York Times	Uma Thurman
Congress	Dylan Farrow
Facebook	Leeann Tweeden
Twitter	
Eric Schneiderman	
Donald Trump	

Table 6: The entities portrayed as most powerful consist of men and abstract institutions. The lowest powered entities consist of primarily women.

Highest Agency	Lowest Agency
Judge Steven O'Neill	Kara Swisher
Eric Schneiderman	the United States
Russell Simmons	Hollywood
The New York Times	Meryl Streep
Frances McDormand	
CNN	
Donald Trump	
Hillary Clinton	

Table 7: Entities with the most the agency and the least agency consist of a mix of men, women, and abstract institutions.

When we examine power (Table 6), the most powerful entities include abstract concepts, like "the #MeToo movement" and "Twitter". Women are conspicuously absent from the list of high-powered entities. Instead we find men, including ones directly accused of sexual misconduct (Eric Schniederman). In contrast, women dominate the list of lowest powered entities. While Table 6 only shows proper nouns, we also observed that common noun references to women were among the least powerful entities identified (e.g. "a women", "these women").

The agency portrayals are more balanced (Table 7). While Eric Schneiderman and Donald Trump appear among the entities with highest agency, we also find female supporters of the movement: Frances McDormand and Hillary Clinton.

6.2 Role-Level

We next conduct a pairwise analysis, where we directly compare two entities who occupied similar roles in different incidents. Through this analysis, we can identify different ways in which narratives of sexual harassment can be framed. Fur-

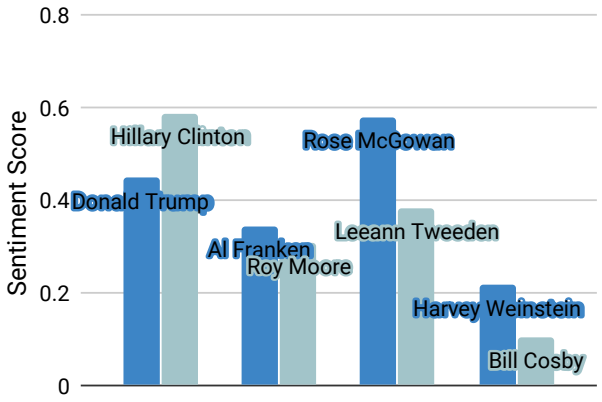

Figure 2: Entities in similar roles are portrayed with different levels of sentiment.

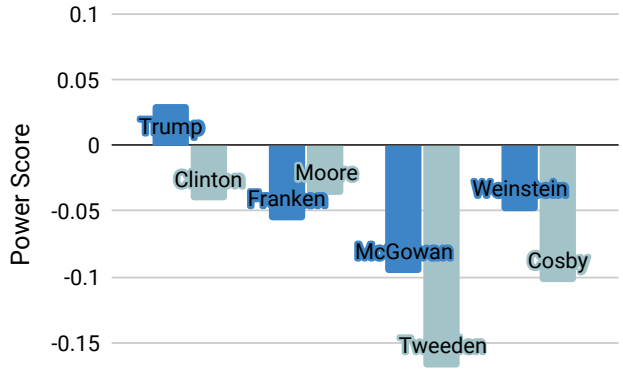

Figure 3: Power scores for comparable entities do not coincide with sentiment scores.

thermore, by decomposing the pair-wise analysis across different outlets, we can identify bias. How do different outlets cover comparable entities differently?

We directly compare sentiment (Figure 2) and power (Figure 3) for several entity pairs. There is a striking difference between the portrayals of Rose McGowan and Leeann Tweeden, two women who both accused high-profile men of sexual assault (Harvey Weinstein and Al Franken respectively). Both women are portrayed with lower power than the men they accused, but Rose McGowan is portrayed with both more positive sentiment and higher power than Leeann Tweeden.

Sample articles about Leeann Tweeden focus on accounts of what happened to her: "The first women to speak out was Leeann Tweeden who said that Franken forcibly kissed her."[9] In contrast, news articles about Rose McGowan focus on statements she made after the fact: "As Rose McGowan, one of the heroes to emerge from the Harvey Weinstein fallout, has mentioned...".[10] We can generalize that in this corpus, Rose McGowan matches more of "survivor" frame, connoting someone who is proactively fighting, while

[9]https://dailym.ai/2WBiA38

[10]https://bit.ly/2TNjtE4

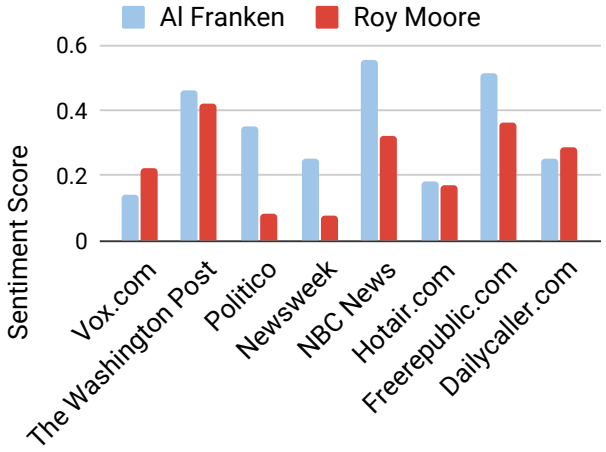

Figure 4: Sentiment scores across left-leaning and right leaning outlets for Democrat Al Franken and Republican Roy Moore do not fall along party lines.
Left-leaning (Democratic) outlets: Vox.com, The Washington Post, Newsweek, NBC News.
Right-leaning (Republican) outlets: Hotair.com, Freerepublic.com, Dailycaller.com.
Centrist: Politico[12]

Leeann Tweeden matches more of a "victim" frame, which connotes helplessness and pity.

Figures 2 and 3 reveal further differences in portrayals. Democrats Hillary Clinton and Al Franken are portrayed more positively than corresponding Republicans Donald Trump and Roy Moore. However, Donald Trump appears as more powerful than every other entity, which coincides with his role as the current U.S. President. In general, politicians accused of sexual harassment (Roy Moore and Al Franken) are portrayed more positively than entertainment industry figures (Harvey Weinsten and Bill Cosby). While few defended entertainment industry figures accused of harassment, politicians received more mixed support and criticism from their own parties. For instance, Donald Trump publicly endorsed candidate Roy Moore, despite the allegations against him.

However, comparing coverage of individuals across the corpus as a whole reveals limited information because it is difficult to separate fact from bias. Does Al Franken receive a more positive portrayal than Roy Moore because his actions throughout the movement were more sympathetic? Or does he receive a more positive portrayal because of a liberal media bias? We can better assess the impact of media bias by comparing how the same entity is portrayed across different outlets. Figure 4 shows sentiment scores for Al Franken and Roy Moore across all outlets that mention both entities at least 10 times.

The coverage of Republican Roy Moore falls broadly along party lines, with the lowest-scoring portrayals occurring in left-leaning outlets Politico and Newsweek. Simi-

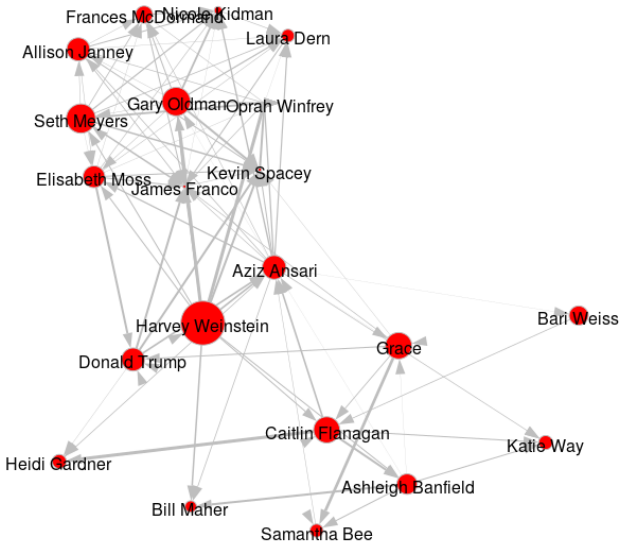

Figure 5: Graphical visualization of power dynamics between entities involved in the accusations against Aziz Ansari

larly, the left-leaning outlets (with the exception of Vox.com) portray Democrat Al Franken more positively than Roy Moore. However, the right-leaning outlets portray both men similarly, notably Freerepublic.com portrays Al Franken more positively than Roy Moore. In reading articles from these outlets, many are sympathetic towards Al Franken, presenting him as a scapegoat, forced out of office by other Democrats without a fair ethics hearing.[13] Our analysis reveals surprising differences in how conservative and liberal outlets react differently to events of the #MeToo movement. Entity portrayals do not necessarily fall along party lines (i.e. liberal outlets portraying liberal politicians positively), but these outlets do focus on different aspects of events.

6.3 Incident-Level

For the incident-level analysis, we return to the example introduced in the beginning of this paper: the accusations against comedian Aziz Ansari. First, we examine the power landscape surrounding Aziz Ansari through a graphical visualization. We develop this graph by drawing from psychology theories of power, which suggest that a person's power often derives from the people around them (French and Raven 1959). We devise a metric to capture the concept of relative power: for any pair of entities, we average the difference between their power scores across all articles in which they are both mentioned. We visualize these differentials in a graph: an edge between two entities denotes that there is at least 1 article that mentions both entities. An edge from entity A to entity B denotes that on average, A is

[13]https://bit.ly/2CCjNAW, https://bit.ly/2COQdao

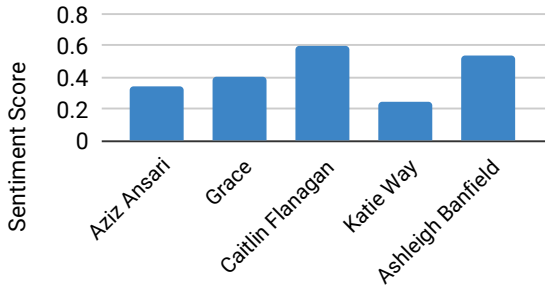

Figure 6: Aziz Ansari and his accuser, Grace, are portrayed with comparable sentiment.

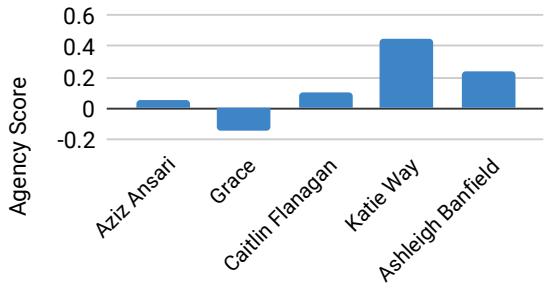

Figure 7: Grace is portrayed with low agency, while journalist Katie Way has very high agency.

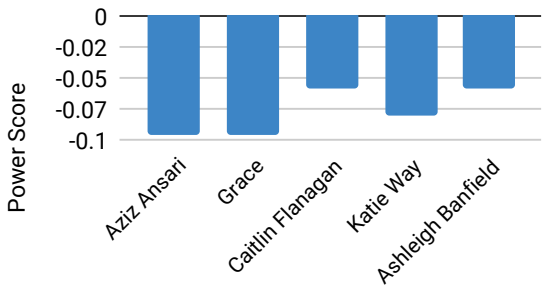

Figure 8: Aziz Ansari and Grace are portrayed with lower power than journalist Ashleigh Banfield.

presented as more powerful than B. The magnitude of that difference is reflected in the edge weight. Finally, we sum all of the edge weights to obtain a score for each node. Thus, a large node for entity A indicates that entity A is often portrayed as more powerful than other entities mentioned in the same articles as A.

We show this graph for articles related to Aziz Ansari by limiting our corpus to articles that mention "Aziz" (555 articles). We take only the 100 most frequently mentioned entities, eliminate all non-people (i.e. Hollywood), and manually group redundant entities (i.e. Donald Trump vs. President Donald Trump). From this graph, we can see two separate clusters, related to two distinct narratives about Aziz Ansari. The cluster of entities in the top left corner, including Seth Meyers, Oprah Winfrey, and Nicole Kidman, is tied to media coverage of the 2018 Golden Globe Awards. All of

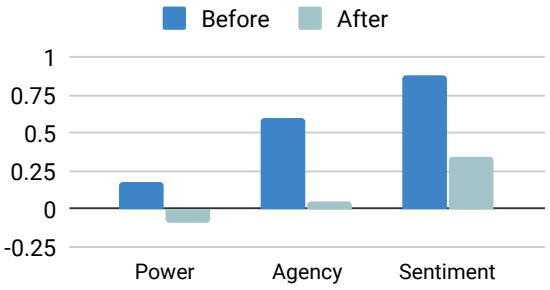

Figure 9: Aziz Ansari is portrayed with lower power, agency, and sentiment in articles published after the allegations against him.

theses entities won awards or gave speeches, including Aziz Ansari, who won Best Actor in a Television Series – Musical or Comedy. Much of the Golden Globes centered on the #MeToo movement, as presenters expressed their support of the movement in speeches or wore pins indicating their support. Aziz Ansari himself wore a "Time's Up" pin, to express his opposition to sexual harassment. Thus, because Aziz Ansari won a prominent award and much of the Golden Globes centered on the #MeToo movement, selecting all articles about Aziz Ansari from our #MeToo corpus results in many articles about the Golden Globes.

In the bottom right corner, we see a second narrative, relating to the accusations against Aziz Ansari. Aziz Ansari and Grace are the primary entities in the narrative. Other frequently mentioned entities include journalists: Katie Way, who authored the original article about the allegations, and Caitlin Flanagan and Ashleigh Banfield, who publicly criticized Katie Way's article. We can see the importance of these journalists in their relative power. For instance, the edge from Caitlin Flanagan to Aziz Ansari indicates that she is often portrayed as more powerful than he is.

We then analyze the sentiment, power, and agency scores for these prominent entities, limiting our data set to articles that mention Aziz Ansari and were published after the Babe.net article that first disclosed the allegations against him (476 articles). Figure 6 portrays sentiment scores. Unlike figures like Harvey Weinstein and Bill Cosby, who have much lower sentiment scores than their accusers, Aziz Ansari has a similar sentiment score to Grace, reflective of the mixed reaction to the article published about them. Katie Way, the journalist who wrote the original article, is portrayed with particularly low sentiment, which coincides with the severe criticism she received for publishing the story. In contrast, Caitlin Flanagan and Ashleigh Banfield, who were both front runners in criticizing Katie Way, were portrayed more positively.

Figure 7 shows the agency scores for the same entities. Grace does have lower agency than Aziz Ansari, which supports the critique that Grace lacks any agency in the narrative. Aziz Ansari and Grace both have lower agency scores in comparison to all 3 journalists. Sentences where Ashleigh Banfield is scored as having high power and agency include: "Banfield slammed the accuser for her 'bad date'."[14] The high agency of these journalists demonstrates the prominent role that journalists have in social movements. In contrast Grace's low agency stems from sentences like "she felt pressured to engage in unwanted sexual acts." [15] The power scores (Figure 8) reflect a similar pattern, Grace and Aziz Ansari are portrayed as less powerful than the 3 journalists.

Finally, we compare sentiment, power, and agency scores for Aziz Ansari in articles published before the Babe.net article (79 articles) and articles published after (476 articles) in Figure 9. As articles about Aziz Ansari before these allegations focus primarily on his Golden Globe victory, these articles portray him with high power, agency, and sentiment. Following the Babe.net article, we see a decrease in Aziz Ansari's power, agency, and sentiment scores.

6.4 Limitations and Ethical Implications

Our analysis is limited by several factors which we identify as areas for future work. First, our metrics for scoring power, agency, and sentiment are based on ternary verb scores. While our work suggests that verbs are informative, we suspect that there are many other relevant features, including other parts of speech like adjectives and apposition nouns as well as high-level features, like whether or not the entity is directly quoted and how early in the article the entity is first mentioned. Additionally, we assign a ternary (positive, negative, neutral) score to each verb, whereas affect scores are likely best measured with continuous values. The incorporation of additional features and continuous-valued verb scores could provide more accurate entity scores. Our work has also highlighted a need for better evaluation metrics for measuring subjective traits like power. As discussed in §5.2, these tasks are difficult for annotators to perform. Best-Worst Scaling, which has been used for creating affective lexicons (Mohammad 2018), may offer a viable framework. Additionally, our data set consists of an imperfect sample of articles covering the #MeToo movement and a different sampling of articles may yield different results.

Furthermore, although our goal in analyzing media coverage is to promote positive journalism and advocate for media that empowers underrepresented people, we acknowledge that our work has the potential to be misused. Tools for analyzing media portrayals can be used to intentionally present biased viewpoints and manipulate public opinion. Furthermore, our analysis of actual people could have unforeseen consequences on them and their public images.

7 Conclusions

We present contextual affective analysis: a framework for analyzing nuanced entity portrayals. While we focus specifically on power, agency, and sentiment in media coverage of the #MeToo movement, our approach is a general framework that can be adapted to other corpora and research questions. Our analysis of media coverage of the #MeToo movement addresses questions like "Whom does the media portray as sympathetic?" and "Whom does the media portray

[14] https://bit.ly/2WFVlVz
[15] https://bit.ly/2HTvNRx

as powerful?" We demonstrate that although this movement has empowered women by encouraging them to share their stories, this empowerment does not necessarily translate into online media coverage of events. While women are among the most sympathetic entities, traditionally powerful men remain among the most powerful in media reports. We further show the prominence of journalists and 3rd party entities commenting on events without being directly involved, not only because their statements can influence perception of the movement, but because by making statements, they become entities in the narrative. Through this analysis, we highlight the importance of media framing: journalists can choose which narratives to highlight in order to promote certain portrayals of people. They can encourage or undermine movements like the #MeToo movement through their choice of entity portrayal.

Acknowledgements

We gratefully acknowledge Alan Black, Hannah Rashkin, Maarten Sap, Emily Ahn, and our anonymous reviewers for their helpful advice. We further thank our annotators. This material is based upon work supported by the NSF Graduate Research Fellowship Program under Grant No. DGE1745016 and by Grant No. IIS1812327 from the NSF. 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 NSF.

Appendix

An alphabetical list of people whom we refer to in connection with the #MeToo movement:

Woody Allen - U.S. movie director; repeatedly accused of sexual assault by daughter Dylan Farrow since 2013

Aziz Ansari - U.S. comedian; accused of sexual misconduct by an anonymous woman in January 2018

Ashleigh Banfield - Canadian-American journalist; criticized the article levelling accusations of sexual assault against Aziz Ansari in an open letter

Tarana Burke - U.S. civil rights activist; started the #MeToo movement to raise awareness about sexual assault

Hillary Clinton - Democratic Party's nominee for President of the United States in the 2016 election, former U.S. Senator and Secretary of State

Andrea Constand - Accused Bill Cosby of sexually assaulting her in 2004; though many women accused Cosby of harassment, only Constand brought her case to court

Bill Cosby - Former comedian; publicly accused of sexual assault or harassment by 60 women; found guilty of assaulting Andrea Constand in April 2018

Caitlin Flanagan - U.S. writer; wrote a widely-circulated article that was critical of the accusations of sexual misconduct against Aziz Ansari

Al Franken - Former U.S. senator; resigned in December 2017 following several allegations of sexual misconduct

Grace - pseudonym employed by the woman whose accusations of sexual misconduct against Aziz Ansari were published in a Babe.net article

Jimmy Kimmel - U.S. television host; delivered a widely reported speech in support of the #MeToo movement as the host of the Oscars in March 2018

Meghan Markle - Former U.S. actress; married Prince Harry in May 2018 (with much media coverage of engagement leading up to the wedding); public supporter of #MeToo movement

Frances McDormand - U.S. actress; prominent because of her support of #MeToo movement during the 2018 Oscars

Rose McGowan - U.S. actress; one of the first women to accuse Harvey Weinstein of sexual misconduct in Oct 2017

Alyssa Milano - U.S. actress; posted a tweet with the hashtag #MeToo and encouraged others who had been sexually harassed to do the same

Roy Moore - U.S. politician; accused by multiple women of sexual misconduct; ran (and lost) for the U.S. Senate in 2017 and was publicly endorsed by Donald Trump

Judge Steven O'Neill - Montgomery Country judge; sentenced Bill Cosby to prison and labeled him a sex offender

Eric Schneiderman - Former New York State Attorney General; resigned in May 2018 following accusations of sexual misconduct by four women

Ryan Seacrest - U.S. television personality and producer; accused of sexual misconduct by his stylist in November 2017

Russel Simmons - U.S. record producer; accused by 18 women of sexual assault

Meryl Streep - U.S. actress who has been accused of long knowing about, but not condemning, Weinstein's alleged sexual misconduct

Kara Swisher - U.S. journalist

Uma Thurman - U.S. actress; detailed charges of sexual assault against Harvey Weinstein in February 2018

Leeann Tweeden - U.S. radio broadcaster; accused Al Franken of sexual misconduct in November 2017

Katie Way - author of an article published at Babe.net which accused Aziz Ansari of sexual misconduct against anonymous 'Grace', and which set off a public dialogue about consent and the definition of sexual misconduct

Harvey Weinstein - Former film producer, accused by over 80 women of sexual misconduct; accusations against him sparked the #MeToo movement

Oprah Winfrey - U.S. media executive and television personality; gave a widely reported speech in support of the #MeToo movement at the Golden Globes in January 2018

References

Bamman, D.; O'Connor, B.; and Smith, N. A. 2013. Learning latent personas of film characters. In *ACL*.

Bamman, D. 2015. *People-Centric Natural Language Processing*. Ph.D. Dissertation, Carnegie Mellon University.

Card, D.; Gross, J.; Boydstun, A.; and Smith, N. A. 2016. Analyzing framing through the casts of characters in the news. In *EMNLP*.

Chambers, N., and Jurafsky, D. 2009. Unsupervised learning of narrative schemas and their participants. In *ACL*.

Danescu-Niculescu-Mizil, C.; Lee, L.; Pang, B.; and Klein-

berg, J. 2012. Echoes of power: Language effects and power differences in social interaction. In *WWW*.

Fillmore, C. J. 1982. Frame semantics. *Cognitive linguistics: Basic readings* 373–400.

Flanagan, C. 2018. The Humiliation of Aziz Ansari. *The Atlantic*.

French, J. R., and Raven, B. 1959. The bases of social power. *Studies in Social Power* 150167.

Gilbert, E. 2012. Phrases that signal workplace hierarchy. In *CSCW*.

Goldberg, A. B., and Zhu, X. 2006. Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In *Workshop on Graph Based Methods for Natural Language Processing*.

Hamilton, W. L.; Clark, K.; Leskovec, J.; and Jurafsky, D. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *EMNLP*.

Heise, D. R. 1979. *Understanding events: Affect and the construction of social action*. CUP Archive.

Heise, D. R. 2007. *Expressive order: Confirming sentiments in social actions*. Springer Science & Business Media.

Iyyer, M.; Guha, A.; Chaturvedi, S.; Boyd-Graber, J.; and Daumé III, H. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *NAACL*.

Joseph, K.; Wei, W.; and Carley, K. M. 2017. Girls rule, boys drool: Extracting semantic and affective stereotypes from twitter. In *CSCW*.

Liu, B. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5(1):1–167.

MacLaughlin, A.; Wihbey, J.; and Smith, D. A. 2018. Predicting news coverage of scientific articles. In *ICWSM*.

Mohammad, S. M. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *ACL*.

Ohlheiser, A. 2018. The woman behind "Me Too" knew the power of the phrase when she created it – 10 years ago. *The Washington Post*.

Osgood, C.; Suci, G.; and Tannenbaum, P. 1957. *The Measurement of Meaning*. Illini Books, IB47. University of Illinois Press.

Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2(1–2):1–135.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *NAACL*.

Prabhakaran, V., and Rambow, O. 2017. Dialog structure through the lens of gender, gender environment, and power. *Dialogue & Discourse* 8(2):21–55.

Rashkin, H.; Singh, S.; and Choi, Y. 2016. Connotation frames: A data-driven investigation. In *ACL*.

Rho, E. H. R.; Mark, G.; and Mazmanian, M. 2018. Fostering civil discourse online: Linguistic behavior in comments of #MeToo articles across political perspectives. In *CSCW*.

Ribeiro, F. N.; Henrique, L.; Benevenuto, F.; Chakraborty, A.; Kulshrestha, J.; Babaei, M.; and Gummadi, K. P. 2018. Media bias monitor: Quantifying biases of social media news outlets at large-scale. In *ICWSM*.

Robinson, D. T.; Smith-Lovin, L.; and Wisecup, A. K. 2006. Affect control theory. In *Handbook of the sociology of emotions*. Springer. 179–202.

Russell, J. A. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39(6):1161.

Russell, J. A. 2003. Core affect and the psychological construction of emotion. *Psychological review* 110(1):145.

Sap, M.; Prasetio, M. C.; Holtzman, A.; Rashkin, H.; Choi, Y.; and Allen, P. G. 2017. Connotation frames of power and agency in modern films. In *EMNLP*.

Spry, T. 1995. In the absence of word and body: Hegemonic implications of "victim" and "survivor" in women's narratives of sexual violence. *Women and Language* 13(2):27.

Way, K. 2018. I went on a date with Aziz Ansari. It turned into the worst night of my life. *Babe.net*.

Weiss, B. 2018. Aziz Ansari is guilty. Of not being a mind reader. *The New York Times*.