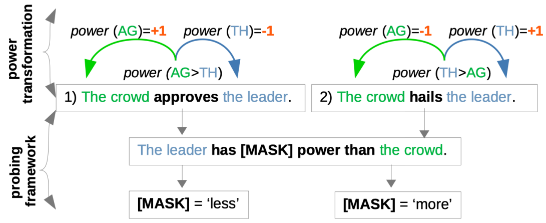
Probing Power by Prompting: Harnessing Pre-trained Language Models for Power Connotation Framing

Shima Khanehzar[†] **Trevor Cohn**[†] **Gosia Mikolajczak**[⋆] **Lea Frermann**[†]
[†] School of Computing and Information Systems, The University of Melbourne
[⋆] The Global Institute for Women's Leadership, The Australian National University
skhanehzar@student.unimelb.edu.au
{tcohn,lfrermann}@unimelb.edu.au
Gosia.Mikolajczak@anu.edu.au

Abstract

Subtle changes in word choice in communication can evoke very different associations with the involved actors. For instance, a company 'employing workers' evokes a more positive connotation than the one 'exploiting' them. This concept is called *connotation*. This paper investigates whether pre-trained language models (PLMs) encode such subtle connotative information about *power differentials* between involved entities. We design a probing framework for power connotation, building on Sap et al. (2017)'s operationalization of connotation frames. We show that zero-shot prompting of PLMs leads to above chance prediction of power connotation, however fine-tuning PLMs using our framework drastically improves their accuracy. Using our fine-tuned models, we present a case study of power dynamics in US news reporting on immigration, showing the potential of our framework as a tool for understanding subtle bias in the media.[1]



Prelude	The crowd approves the leader.
Hypothesis	The leader has MASK power than the crowd.
Connotation	AG$\succ_{power}$TH
Prelude	The crowd hails the leader.
Hypothesis	The leader has MASK power than the crowd.
Connotation	AG$\prec_{power}$TH

Figure 1: Top: Connotation frames from Sap et al. (2017)'s dataset for the predicates 'approve' and 'hail', which evoke different power differentials between arguments 'crowd' and 'leader'. Predicate 'approve' implies that the leader needs approval of the *powerful crowd*; while 'hail' suggests that the crowd praises the *powerful leader*. Colored arrows show how we map from predicate- to entity-centric connotation. Bottom: Our PLM probing framework comprising a *prelude* (introducing predicate and arguments) and a *hypothesis* (expressing the implied power differential).

1 Introduction

An author's choice of words evokes associations and reactions from a reader that go beyond the literal meaning they express. This underlying level of meaning is called connotation and often carries social, cultural, or emotional implications for listeners or readers (Sonesson, 1998). In high-stakes settings such as opinion polls or news reporting, it has been shown that subtle changes in word choices can influence responses or opinions (Kahneman and Tversky, 1984; Rashkin et al., 2016). For example, the choice of the term 'undocumented workers' vs. 'illegal aliens' to describe immigrants can elicit different levels of prejudice toward that group (M.S., 2010).

Different connotations can inject multiple layers of meaning into a word, phrase, or passage. This paper focuses on *connotations of power* arising from descriptions of entities. Power dynamics are omnipresent at all levels of society, between individuals, groups, political actors, or institutions, and the subjective power of an entity can be expressed through the choice of words used to describe their actions (Sap et al., 2017). For example, in Figure 1, sentence (1) implies the leader requires approval of the *powerful crowd*; while sentence (2) implies the crowd praises the *powerful leader*.

Following the successful application of *probing* pre-trained language models (PLMs) to a variety of tasks ranging from grammatical structure (Koto et al., 2021; Kulmizev et al., 2020; Kassner and Schütze, 2020; Sinclair et al., 2022) to commonsense reasoning (Lin et al., 2020), we extend the probing paradigm to connotative power and study

[1] Source code is available at https://github.com/shinyemimalef/Probing-Power-by-Prompting

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 873–885
May 2-6, 2023 ©2023 Association for Computational Linguistics

the extent to which PLMs can be used to reliably predict power dynamics between two entities.

We do so by drawing on the recently proposed notion of *connotation frames* (Rashkin et al., 2016; Sap et al., 2017; Ma et al., 2020), a structured framework that captures connotative associations evoked by a predicate (verb) about its agent (subject) and theme (object). Figure 1 illustrates the concept of the *power connotation frame*, which for a given predicate predicts whether the agent (subject) has more, less, or equal power compared to the theme (object). Specifically, we devise a probing framework to test PLMs for this formalization of connotative power which carefully controls for confounds relating to the specific agents and themes and probe structure. We show that, while PLMs cannot reliably predict power dynamics in a zero-shot setting, it is possible to fine-tune PLMs using our probes to predict power connotation with close to 80% macro F1.

Using our fine-tuned PLMs, we present a case study of power dynamics in US news reporting on the topic of immigration. News reporting is a particularly prominent example of documents where power dynamics are often at the core of the issue under discussion, and there is often an intrinsic motivation, especially in partisan news outlets, to present information in a biased way. We draw connections between power dynamics and emphasis framing and surface subtle bias in the context of immigration. In sum, our main contributions are:

- We propose a method to (i) disentangle connotation frames implied by the predicate from its arguments and the sentence structure; and (ii) quantify predicate connotation frames in PLMs.

- We probe the zero-shot ability of common PLMs to capture connotation frames and find poor performance, suggesting that this subtle signal is not captured in the representations.

- We show that the power prediction performance of PLMs can be drastically improved by fine-tuning on a small set of labeled instances, achieving F1 scores close to 80%.

- Using the best model, we present a case study on the news reporting on immigration and analyze how the power of immigrants and immigration services are portrayed in US news outlets with different political leaning.

2 Background

In this section, we provide an overview of the relevant literature on connotative framing and its formalization, probing PLMs, and how connotative framing and PLMs combined can be harnessed to study media bias.

2.1 Connotation Frames

Connotation frames were introduced by Rashkin et al. (2016) as a formalism for examining the sentiment and presupposed facts about actors and themes, as implied by the actions and events that they engage in (i.e., their predicates). The original framework cover several connotations, incld-ing the writer's perspective (e.g., being sympathetic/antagonistic towards the agent), effect (e.g., the theme has been hurt), and the mental states projected onto agents and themes (e.g., being unhappy). Later work by Sap et al. (2017) extended the set of connotation frames to include 'power', which denotes the relative authority levels of the agent and theme implied by the predicate, and 'agency' defined as the capability of the agent to progress their own narrative. While our paper focuses on *power*, we note that our methodology can be extended to other connotation dimensions.

Sap et al. (2017) published a data set of English verbs manually annotated with power levels: The agent has either more power (AG$\succ_{power}$TH; the writer implies the agent has a level of control over the theme), or less power (AG$\prec_{power}$TH; the theme is implied to be more authoritative). We use a subset of their data in this work.[2] Unlike previous approaches to connotation frame prediction, which used annotated data to train supervised classifiers (Field et al., 2019; Rashkin et al., 2016), we use the method of probing to study and extend PLMs' ability to predict connotative power.

In earlier application of connotation frames to study various biases, researchers studied entities' depicted sentiment polarities in news articles (Rashkin et al., 2016), gendered bias in movies Sap et al. (2017), or people portrayals in #metoo stories Field et al. (2019). In our case study, we connect power connotation with emphasis frames (Card et al., 2015) for a more nuanced analysis of media bias. We show that our method is applicable to complex input sentences, diverging from settings in prior work, which largely considered isolated,

[2] The original data also includes an 'equal power' option, which we do not use here.

short text snippets (such as (agent,predicate,theme) tuples), and abstracted away from specific entities by using placeholders.

2.2 Probing PLMs

Probing language models for different types of knowledge has become a widely-used approach to understanding the knowledge encoded in PLMs (Liu et al., 2023). Typically, a PLM is presented with a prompting input and either completes the input or predicts the most likely token to fill in a masked position. Auto-prompting (Shin et al. (2020); PLMs generating prompts) and continuous prompting (Qin and Eisner (2021); prompting with an embedding) have been proposed recently, reducing the requirement of natural prompt engineering, but also transparency. While probing is a flexible framework, previous work showed that results can be sensitive to prompt wording and structure (Elazar et al., 2021a), and recommended considering such confounds by careful construction of controlled probe sets. Probing can be used as a zero-shot knowledge querying strategy or a way of fine-tuning to train PLMs to more accurately complete probes (Liu et al., 2023). Here, we extend probing to subtle connotative power associations arising from transitive verbs. Following previous work (Trichelair et al., 2019; Elazar et al., 2021b), we introduce several layers of controls, including paraphrased prompts and controlled sets of entities (i.e., predicate arguments) to ensure our findings are robust.

2.3 Media Bias and Framing

Automatically predicting framing bias has attracted recent attention in the NLP community, ranging from article-level frame prediction (Card et al., 2016; Field et al., 2018; Khanehzar et al., 2019, 2021) to tweet analysis (Mendelsohn et al., 2021). Previous works (Card et al., 2016; Khanehzar et al., 2021) showed characteristics of entities could help predict the article-level emphasis frames. In a case study (§ 7) on the topic of immigration in major US news outlets, we analyze the interplay of power connotation and article-level emphasis frames. Immigration has been a contested issue in the US, with the proponents and opponents actively trying to steer public opinion towards their stance by emphasizing selective and often simplified aspects of the topic (Farris and Silber Mohamed, 2018; Lawlor and Tolley, 2017; Ommundsen et al., 2014).

3 Probing PLMs for Connotative Power

We propose a probing framework to assess how much connotative information PLMs acquired during pretraining, including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and ALBERT (Lan et al., 2020). Similar to the masked token prediction objective during PLM pre-training, we have PLMs predict a masked token which in context reveals the power dynamics between an agent and theme in a given sentence. Our probe formulation is closely tied to Sap et al. (2017)'s annotation instructions to align the ground-truth data and information elicited from the PLMs.

We directly query the underlying masked language model in PLMs to compute power connotations associated with predicates. In particular, we formulate our probing template reflecting the definition of the power connotation:

Probe 1 *AG P TH. TH has MASK power than AG.*

Here, P is a placeholder for a predicate (details in §5.2). AG and TH are placeholders for concrete entities in the agent and theme position, which during probing and fine-tuning, are instantiated with common English names (details in §3.1). In the remainder of this paper, we will refer to the first sentence of our template as *prelude* which introduces the verb and associated entities; and to the second sentence as *hypothesis*. Figure 1 (bottom) shows two instatinated probes.

The PLMs are probed with the context of an instantiated probe with $\{P{=}p, AG{=}ag, TH{=}th\}$:

$$P(\text{MASK}{=}m|p, ag, th),$$

for the predicted probability of two possible values in the MASK position, namely the masked target word taking value $m{\in}\{\text{'more'}, \text{'less'}\}$. For example, if $P(m{=}\text{'more'}) > P(m{=}\text{'less'})$, the PLM predicts higher power of theme (th) than agent (ag) of the verb p, (AG$\prec_{\text{power}}$TH). To reduce the impact of specific instantiations of entities, we compute the final power connotation score for p as the average score over a large numger of entity combinations in the AG and TH positions:

$$S^{p}_{\text{probe}}(m) = \frac{1}{K} \sum_{\substack{ag \in A \\ th \in T}} P(\text{MASK}{=}m|p, ag, th) \tag{1}$$

where A and T are the set of all candidate entities for the AG and TH positions, respectively, and $K{=}|A||T|$.

3.1 Removing confounds

As discussed in §2.2, PLM probing is susceptible to confounds, including probe structure and choice of arguments, both of which can effect connotation frame prediction. We introduce measures to control three confounds (C1–C3).

C1: Probe structure In addition to Probe 1 we propose a second, semantically equivalent, probe with different structure.

Probe 2 *AG P TH. TH is MASK powerful than AG.*

We use different probe formulations to ensure that the model predictions are not incidental to a particular probe wording; and that the model is *consistently correct*. To this end, we adapt Elazar et al. (2021b)'s group evaluation (§5.3) score which only accepts a model prediction as correct if the correct class is predicted for multiple variants of the probe.

C2: Entity semantics We aim to capture power differentials as implied by the predicates that relate to two entities. As such, we do not want the predicted scores to be impacted by intrinsic power associated with the entities (e.g., high power 'presidents' vs. low power 'immigrants'). To this end, we constrain our entities to common English names. We use 16 female and male names from Nosek et al. (2002),[3] because (1) unlike general groups or entities (immigrants, agencies, politicians, etc.) names are largely free of a priori connotation which could impact our connotation scores; (2) all names are in the PLM vocabulary, removing discrepancies in subword tokenization; and (3) the set of names has been tested in implicit association tests (IAT and WEAT) and calibrated to only differ in gender, removing confounding effects of class or frequency (Jentzsch et al., 2019). The same set of names has been previously used in the context of gender bias mitigation (Gupta et al., 2022). To remove the gender confound we instantiate all possible $\{ag, th\}$ tuples from our 16 names, and reporting per-verb power connotations, obtaining $16 \times 15 = 240$ name combinations for each of our two probe templates, resulting in 480 instantiated probes per verb. We finally average probe predictions as in equation 1.

C3: PLM representation bias PLMs encode demographic biases, including stereotypical gender associations (Sun et al., 2019). Thus, in addition to controlling the entity set, we normalize power prediction scores by the PLM's a priori power assessment of the agent and theme names in the absence of a predicate. This normalization is also important for our case study in §7, where we use our probing framework with a fine-tuned PLM on less constrained contexts involving real-world entities in news articles. Inspired by previous work (Trichelair et al., 2019; Elazar et al., 2021b), we subtract from our connotation score the a priori relative power between the agent and theme in the absence of a specific predicate connecting them. To do so, we predict the MASK token given only the *hypothesis*:[4]

Prior 1 *TH has MASK power than AG.*

Like for the full probes, we obtain the probability of m with $m{\in}\{\text{`more'}, \text{`less'}\}$ from our PLMs and average over all entity combination, obtaining S_{prior}:

$$S^{p}_{\text{prior}}(m) = \frac{1}{K}\sum_{\substack{ag\in A\\ th\in T}} P(\text{MASK}{=}m|p, ag, th)$$

where P is computed using the Prior 1 template.

3.2 The Power Probing Score

The final power probing score (PPS) is the difference of the log predicted probability of the candidate MASK terms by the prior and the full probe:

$$\text{PPS}^{p}(m) = \log S^{p}_{\text{probe}}(m) - \log S^{p}_{\text{prior}}(m),$$

where $m \in \{\text{`more'}, \text{`less'}\}$. We finally derive a connotation score (CS) as a binary indicator variable:

$$CS^{p} = \begin{cases} +1 & \text{if } PPS^{p}(\text{`less'}){>}PPS^{p}(\text{`more'}) \\ -1 & \text{otherwise,} \end{cases} \tag{2}$$

noting that CS=+1 directly corresponds to Sap et al. (2017)'s (AG$\succ_{power}$TH) and CS=-1 to (AG$\prec_{power}$TH).

4 Fine-tuning PLMs for connotative power prediction

As we will show in §6.1, PLMs do not encode power connotation naturally (in the zero-shot setting). We therefore aim to instill connotative knowledge into PLMs through task-specific fine-tuning.

[3] Male: {John, Paul, Mike, Kevin, Steve, Greg, Jeff, Bill}; Female: {Amy, Joan, Lisa, Sarah, Diana, Kate, Ann, Donna}

[4] See Table 3 in the Appendix for a complete list of priors.

Probe 1	AG P TH. TH has MASK power than AG.
No-Ent 1	P. has MASK power than.
Part-Sent 1	TH has MASK power than AG.

Table 1: Probe 1 (top) and two derived baseline templates, where agent (AG), theme (TH), and predicate (P) are instantiated as explained in §3.1. Probe 2 baselines are constructed analogously (cf., Appendix Table 3).

We use the probes introduced in §3 to construct a connotation prediction task as masked token prediction-based fine-tuning. The input to the model is the instantiated probe with instantiated ag, th, and p and with the target masked word m. The model is trained to predict 'more' or 'less', in correspondence with the gold standard power connotation value for the predicate p, CS^p. Training then proceeds as normal, backpropagating gradients through the full model, allowing the model to focus better on aspects of the encoding related to power connotation.

5 Experimental Setup

5.1 Baselines

We include a random baseline and a majority-class baseline to compare with PLMs. In addition, we provide two baselines to investigate the underlying language models' biases towards context and arguments of template sentences, by constructing inputs that are structurally similar to our probes, but do not carry any signal towards power connotations. This allows us to test (1) the extent of bias in the PLMs measured by deviation of the baselines from a random model; and (2) how much actual power connotative knowledge (over and above biases) the models contain in a zero-shot setting, or after fine-tuning, by comparing the respective models tested with the probes against these baselines. Both baselines are adapted from prior work on controlled debiasing (Elazar et al., 2021b).

No-Entity Baseline This baseline quantifies a priori PLM bias towards P in the absence of any entity. Probe 1 is reduced to No-Ent 1 in Table 1 (middle). An unbiased model should predict 'more' and 'less' with equal probability as the prompts contain no signal of the power differential.

Partial-Sentence Baseline This baseline quantifies prior biases related to the arguments by removing the predicate. We reduce the Probe 1 to only the *prelude* (similar to our prior in §3), see Part-Sent 1 in Table 1. Again, there is no information about power, and we expect an unbiased model to show random performance. Both baselines predict CS^p analogous to equations 1 and 2, but conditioning on their respective contexts.[5]

5.2 Predicate Selection

We use a subset of Sap et al. (2017)'s dataset of power-annotated verbs, keeping only transitive verbs applicable to human actors and themes based on manual filtering. We retain 300 verbs, with 67% of AG$\succ_{power}$TH and 33% verbs with AG$\prec_{power}$TH.

5.3 Evaluation metrics

We report class-wise and macro-averaged F1 scores to provide a detailed view of model performance and acknowledge the label skew in the ground truth data. We report standard (single-instance) evaluation for each test probe separately, resulting in 200 verbs × 2 probes = 400 instances; as well as the stricter grouped evaluation discussed below.

Group evaluation The more stringent *group evaluation* ensures that predictions are consistent across our semantically equivalent probes. A prediction CS^p for a given predicate p is accepted as correct only if the predicted connotation is correct for both probe 1 and probe 2. We compute the connotation score for each (ag, p, th) with probe 1 and probe 2 templates, and assign the worst-performing score. This group evaluation lowers the chance of random or coincidental correct predictions.

6 Results

We first test PLMs for power connotation knowledge in a zero-shot setting (§ 6.1), before we turn to task-specific fine-tuning (§ 6.2). See Appendix A for parameter settings.

6.1 Zero-shot Setting

We perform five separate runs. Each time, we obtain a stratified sample of 200 verbs selected from the full dataset of 300 to test the model. We report mean and variance across the five runs.[6]

Table 2 (top left) shows all three PLMs in the *single evaluation* metric. The 'No-Entity' and 'Partial-Sentence' baselines degrade to the majority classifier, which demonstrates that the PLMs contain

[5] See Table 3 in the Appendix for Probe 2 baselines.

[6] This evaluation protocol mirrors the fine-tuning setup in §6.2, allowing the comparison of variance between settings.

		Single Evaluation			Group Evaluation
	Setup	F1+	F1-	macro F1	macro F1
	Random	57.20 ± 2.60	39.74 ± 2.90	48.49 ± 2.40	21.56 ± 2.46
	Majority	80.36	0	40.18	40.18
Zero-shot	No-Entity*	80.36	0	40.18	40.18
	Partial-Sentence*	80.36	0	40.18	40.18
	BERT-Prob	75.41 ± 1.00	49.51 ± 3.00	41.64 ± 1.30	32.67 ± 1.80
	ALBERT-Prob	75.22 ± 0.40	44.54 ± 1.40	59.88 ± 0.80	40.84 ± 1.71
	RoBERTa-Prob	49.72 ± 0.16	2.19 ± 1.18	25.95 ± 0.60	25.00 ± 0.39
Fine-tuned	BERT-No-Ent	46.01 ± 14.61	12.81 ± 6.14	29.41 ± 4.38	23.35 ± 7.49
	ALBERT-No-Ent	20.58 ± 2.36	34.17 ± 2.94	27.38 ± 0.47	18.91 ± 0.43
	RoBERTa-No-Ent	58.42 ± 4.69	22.34 ± 5.76	40.38 ± 0.70	29.74 ± 7.69
	BERT-Part-Sent	42.88 ± 12.44	10.62 ± 4.52	26.74 ± 8.84	20.02 ± 14.80
	ALBERT-Part-Sent	19.11 ± 27.02	46.17 ± 4.62	32.64 ± 11.20	16.48 ± 11.65
	RoBERTa-Part-Sent	72.68 ± 10.86	13.21 ± 18.69	42.95 ± 3.91	26.79 ± 18.94
	BERT-Prob	85.36 ± 0.96	51.59 ± 4.37	68.48 ± 2.62	66.30 ± 2.57
	ALBERT-Prob	84.45 ± 0.29	66.49 ± 0.12	75.43 ± 0.29	74.02 ± 0.28
	RoBERTa-Prob	86.56 ± 0.98	70.73 ± 0.86	78.65 ± 0.88	77.18 ± 1.51

Table 2: Class-wise and macro F1 score of *power* connotation frame predictions with both single- (left) and group evaluation (right) for random and majority baseline, zero-shot setting and after task specific fine-tuning. *Results for these baselines are identical across all PLMs (and to the majority baseline) as they consistently predicted 'more' in this setting. Results are the mean and standard deviation over 5× repeated random subsampling of 200 verbs as a test set.

significant biases. An unbiased model would show a random performance.

Overall, in the zero-shot setup using the probe templates, the ALBERT model outperforms the BERT and RoBERTa models and random baseline in macro F1. Surprisingly, RoBERTa performs very poorly compared to the other two PLMs. We speculate that this might be due to the next sentence prediction loss, which is not part of RoBERTa training but is included BERT and ALBERT, thus explaining their ability to learn short-range dependencies between adjacent sentences in the probe templates. The main finding in this experiment is that none of the models contain power connotation information, as their performances are all close to random. Next, we show that PLMs can be effectively fine-tuned on this task.

6.2 Fine-tuned Setting

We evaluate the performance of fine-tuned PLMs using our proposed masked predication based approach (§4). As in the zero-shot experiments, we perfome 5 separate runs, each time using a stratified sample of 100 verbs for model fine-tuning, and the remaining 200 verbs for testing. We report mean and variance across these five runs.

Table 2 (bottom) shows that the fine-tuned models perform significantly better than the zero-shot versions across architectures. This holds in both the *single* (left) and the stricter *group evaluation* (evaluation). In the group evaluation, we observe the performance drops slightly compared to single evaluation after fine-tuning, and to a lesser extent than in the zero-shot setting. This suggests consistency of model predictions, increasing our confidence that the PLMs indeed capture connotative power associations with the predicates. Although RoBERTa performed poorly in the zero-shot setting, it is the strongest model after fine-tuning.

All fine-tuned models outperform all baselines by a large margin. Furthermore, we compared the performance of our fine-tuned models with the best previously reported approach for power connotation prediction — the logistic regression model of Field et al. (2019). The logistic regression model obtains macro F1 of 60%, substantially lower than our fine-tuned results.

Next, we analyze the impact of the training set size, and the robustness and consistency of our probing score.

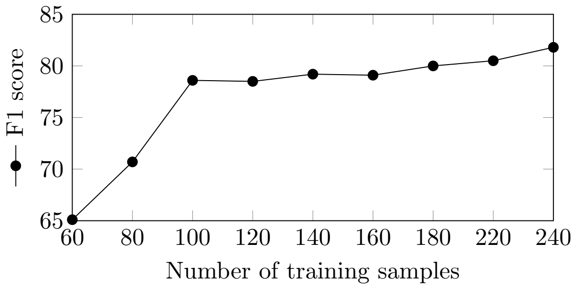


Figure 2: Performance (macro F1) of the fine-tuned RoBERTa model in single evalution with training sets of varying size.

Impact of the training set size Our main results are based on fine-tuning on only 100 labeled instances. We investigated the effect of varying the number of training samples from $\{60, 80, \ldots, 220, 240\}$, and a fixed test set of the remaining 60 instances for RoBERTa. The benefits of using more training samples plateaued after 100 samples as shown in Figure 2. Pre-trained models can efficiently learn to capture power connotation information from few training samples.

Robustness For each predicate, we calculated the variance of the probability of the masked token taking values 'more' or 'less', normalizing their scores to sum to 1 and averaging across all possible combinations of names in the agent and theme position (n=240). Overall, variance is low (mean=0.003). The predicates with the highest variance (0.005) include 'question' and 'reach', while those with lowest variance (0.001) include 'shut' and 'ruin'. Intuitively, the former (high-variance) verbs are more ambiguous than the latter.

Consistency We checked whether, given a predicate, the predicted filler for the masked token is the same across possible combinations of names in the agent and theme position. Across all test predicates, we find that 88% of the time the prediction is consistent among the 240 argument combinations. These final two analyses suggest that our probing score is not sensitive to the confounding variables of the agent or theme position or gender.

7 Case study

We employ our best-performing fine-tuned model RoBERTa to explore subtle power dynamics in US news reports on immigration. In particular, we study the power connotations implied in the descriptions (i.e., actions and events) involving prominent *entities* in the articles. In doing so, we draw connections between power connotation and emphasis frames (§7.1) as well as power connotation and issue stance (§7.2). We first describe our dataset and the entity extraction process, before explaining how we map the predicate-centric probing framework from §3 to an entity-centric measure.

Dataset We use the Media Frames Corpus (MFC, Card et al. (2015)), which contains 6.7k news articles about immigration,[7] manually labeled with one of 15 emphasis frames (including `Legality`, `Political`, `Economic`, ...), as well as an additional 42k unlabeled news articles. The articles were sourced from 13 U.S. news outlets published between 1969 and 2017. A subset of the immigration articles includes manual labels of 'stance' which indicates whether the article author is supporting, neutral about, or opposing immigration.[8]

Entity extraction We identify the most common entities within each article using an off-the-shelf transformer-based semantic role labeling model[9] (Shi and Lin, 2019) which has been previously used to identify key entities in MFC articles (Khanehzar et al., 2021). For each sentence in a document, we collect spans corresponding to the main verbs (`predicate`), and their first (ARG0) and second (ARG1) arguments, and then apply a coreference resolution model (Lee et al., 2018) to group the arguments into entities. We also use NER and string matching to find all mentions of each argument and consider only entities mentioned >3 times in an article. Figure 5 in the Appendix B shows the most common entities in the dataset.

Entity-level power connotation We apply our power probing framework as introduced in §3, and add a final step to map predicate-level CS^p to an entity-level CS. Similar to Field et al. (2019), and as illustrated in Figure 1, if $CS^p = +1$ (AG$\succ_{power}$TH), we consider connotation score of the agent as positive CS(AG)=+1 and the theme as negative CS(AG)=-1, and vice versa for predicates with $CS^p = -1$. This approach enables us to obtain power scores for entities in unannotated documents. To obtain the power score for each entity in each article, we average the power score

[7] The MFC covers four other issues, but immigration is the most prominent issue spanning the longest time period.

[8] Pro: 2740, Anti: 1685, Neutral: 982, None: 1350

[9] From AllenNLP, trained on OntoNotes5.0

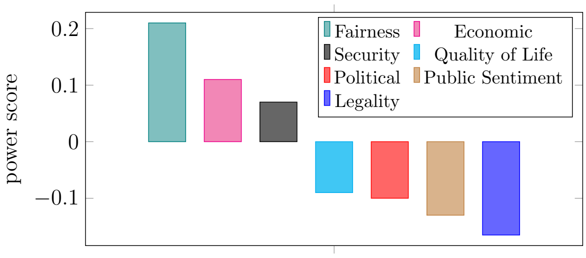


Figure 3: The power of 'immigrants' in articles with different emphasis frames (see legend). The majority of differences are significant, see Appendix C for details.

predicted for each mention of the entity (in the context of a transitive predicate) in that article. The entity's overall power score across the corpus is the average of its power scores across all articles.

7.1 The Power of Immigrants across Different Emphasis Frames

Immigration is a contested issue in the US with politicians, lobby groups, and news agencies selectively framing the topic in a way that supports their stance (Farris and Silber Mohamed, 2018; Lawlor and Tolley, 2017; Ommundsen et al., 2014). We examine how the news outlets portray the key entity involved in this discourse, namely 'the immigrants'. Specifically, we relate power connotations to emphasis frames in the MFC. We applied the method described above to all 6.7k frame-labelled immigration articles, and considered all mentions of the term 'immigrant(s)' and their co-referents.

Figure 3 shows that immigrants are depicted as relatively powerful when the articles are using the Economic, Fairness and Equality, Security and Defense frames, and fairly powerless when the articles are using the Legality, Constitutionality, Jurisdiction, Public Sentiment, Quality of Life, and Political frames. This suggests that news outlets generally approve of immigrants' role in contributing to the economy as confirmed by example mentions like 'Low-skilled immigrants *provide* cheap child care' and 'More immigrants than ever before *start* their own companies'. The gold standard power connotation label of *provide* and *start* is AG$\succ_{power}$TH, and our fine-tuned model predicted the same label with confidence for both cases.

Another example of depicting immigrants as relatively high in power are articles adopting the Security and Defense frame. These articles portray immigrants' power to disrupt the societal order. Examples include 'Immigrants *break* a security cordon', 'immigrants *overwhelm* the guards' and 'immigrants *overstay* their visas'. Conversely, articles using the Fairness and Equality frame tend to portray immigrants in an unfavorable light, albeit with high power associated with their actions, implying malicious intentions. Examples include 'immigrants *bring* crime and drugs', and 'detained immigrants *fabricate* accounts'.

These examples show that our operationalization of power complements the positive vs negative sentiment dimension, and in general, is not a stable trait of an entity, but rather depends on the more general frame adopted by an article. Additionally, in emphasis frames depicting immigrants as low in power, for example, in the Legality frame, immigrants are often associated with either positive or neutral actions such as 'immigrants *comply with* Federal laws', or 'immigrants *renounce* the citizenship of their native countries'. Among the frames depicting the immigrants as powerless, we noted that articles adopting the Quality of Life frame generally imply that immigrants, especially undocumented immigrants, face difficulties in life. Examples include 'immigrants *suffer* the alleged brutality', 'immigrants *fear* they will be hounded and deported'. Under the political frame, articles generally portray politicians' actions or relations towards immigrants, often attributing less power to immigrants. Examples include 'illegal immigrants *toil* for governor', 'Rudy accused Mitt *employing* illegal immigrants'.

7.2 Power of Immigrants vs. Immigration Services and Issue Stance

Immigration services such as ICE and INS[10] play a significant role in executing and designing immigration policies in the US and are prominent in the public discourse. Notorious for hardline approaches, these agencies are often criticized by the more liberal news outlets for abusing their legal power (Omokha, 2022). We compare the connotative power associated with *immigrants* and *immigration services* in news articles with a supportive, opposing or neutral stance on immigration, as manually labeled in the MFC.

Figure 4 shows, unsurprisingly, that immigrants are generally portrayed as less powerful than immigration services. At the same time, we observe

[10] Immigration and Customs Enforcement (ICE) and Immigration and Naturalization Service (INS)

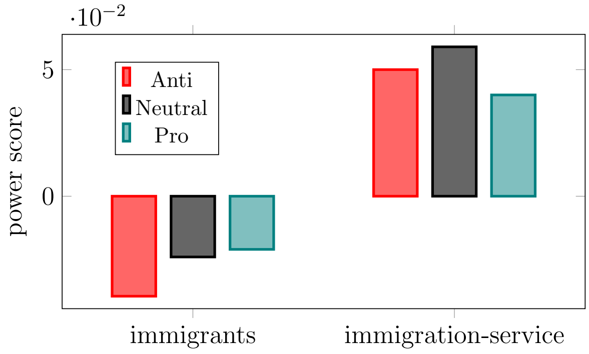


Figure 4: Power of ‘immigrants’ and ‘immigration services’ in articles with an anti-, pro-, and neutral- immigration stance. See Appendix C for significance results.

different trends for the two entities depending on the author’s stance on immigration: immigrants are portrayed as ‘most powerless’ in anti-immigration articles, but slightly less so in neutral and pro-immigration articles. In pro-immigration articles, immigrants more often take action (instead of being acted upon). Illustrative examples from pro-immigration articles include ‘immigrants *navigate* the byzantine rules of permanent residency’, and ‘immigrants *discover* the power of citizenship’, and ‘people should *join* immigrants in the continuing fight for civil rights and human dignity’. Examples of depictions of immigrants as powerless in anti-immigration articles include ‘officers *detain* a total of thirty Haitian immigrants’ and ‘immigration officials *arrest* illegal immigrants’. Immigration services, on the other hand, are portrayed as powerful in both pro- and anti-immigration articles, but less so in neutral ones. Immigration services in anti-immigration articles are often involved in strong actions toward immigrants: for instance, the predicates ‘arrest’ and ‘deport’ appear 1.8 times more often in anti-immigration articles than in pro-immigration articles.

8 Conclusion

We introduced a framework to (i) disentangle connotation frames implied by the predicate from its arguments and the sentence structure; and (ii) quantify predicate connotation frames in PLMs. Using the proposed framework, we investigated the capability of pre-trained language models to understand connotative information focusing on power dynamics between involved entities, both in a zero-shot setting, where performance was overall poor, and fine-tuning, which lead to drastic improvements. Our framework can be applied to investigate other connotation frames such as agency and sentiment, and their relationships to media bias, although defining the probing template may be challenging due to the subtle nature of connotation frames.

Finally, in a case study, we showed how our model can be used to detect subtle differences in the implied power dynamics between entities. Our findings highlight the potential of our framework as a tool for understanding subtle bias in the media. Future work could use our framework to analyze language in various forms of media, e.g. social media posts and TV programming, to identify patterns and trends in language usage for conveying power and other connotations.

It is worth investigating whether the improved performance in connotation predictions would enhance the model’s reading comprehension (Rajpurkar et al., 2016) capability at a higher level. Naturally, we would like to see that the model can infer which entity is more powerful or wields more authority without the text explicitly stating that fact. Previous work has shown that different layers in PLMs specialize in various meanings (de Vries et al., 2020). Future work should investigate the extent each layer in a PLM contributes to encoding connotation frames.

Our framework can also be used to study character roles by analyzing the verbs or verb phrases used to describe the characters and the connotations that they carry. For example, a news article about a political leader who is advocating for tougher immigration policies might mention the event of ‘cracking down on illegal immigration’, evoking a negative connotation and portraying the leader as a ‘villain’. A different article on the same entity might include an event of ‘helping refugees find safety’ assigning the leader the role of ‘hero’ with a more positive connotative association. Both articles frame the immigrants as ‘victims’. By analyzing the language used to describe characters and the connotations it carries, it is possible to gain insight into their framing and roles in the broader narratives around an issue.

9 Limitations

We identify several limitations and shortcomings in our work as potential areas for future work. While the proposed probing framework could be used to investigate other connotation frames, we focused only on the power dimension. We plan to extend our work to other connotation dimensions and explore their relationships with media bias.

While our case study demonstrates that predicate connotative information can be used to depict entity-level power connotation, we recognize that many other relevant features might do so. These include low-level features like adjectives and noun phrases, as well as high-level ones like the position of the entity in the article, and whether or not the entity is directly quoted. Future studies could compare how these different features contribute to predictions of connotation frames.

We acknowledge that our model, which predicts power dynamics on predicate signals alone, simplifies the construct of social power. This might lead to inaccurate predictions. Furthermore, the automatic analysis of portrayed power of real world entities or groups bears a risk of misuse to justify discriminatory practices or policies on the basis of the power portrayal in the media.

To mitigate these risks, it is essential to consider the potential consequences of the model's predictions and ensure they are used ethically and responsibly, including considering the potential impact on marginalized groups and taking steps to minimize any potential biases in the model's predictions. Additionally, human oversight and interpretation of the results is important, especially when they are used in decision-making processes that have a social impact.

Ethics Statement

The original news articles from MFC (Card et al., 2015) used in this work were obtained from Lexis Nexis under the institutional licence held by the University of Melbourne.

Acknowledgements

We thank the anonymous reviewers for their helpful feedback and suggestions. This article was written with the support from the graduate research scholarship from the Melbourne School of Engineering, University of Melbourne provided to the first author. This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne, established with the assistance of LIEF Grant LE170100200.

References

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Dallas Card, Justin Gross, Amber Boydstun, and Noah A. Smith. 2016. Analyzing framing through the casts of characters in the news. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1410–1420, Austin, Texas. Association for Computational Linguistics.

Wietse de Vries, Andreas van Cranenburgh, and Malvina Nissim. 2020. What's so special about BERT's layers? a closer look at the NLP pipeline in monolingual and multilingual models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4339–4350, Online. Association for Computational Linguistics.

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

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021a. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Yanai Elazar, Hongming Zhang, Yoav Goldberg, and Dan Roth. 2021b. Back to square one: Artifact detection, training and commonsense disentanglement in the Winograd schema. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10486–10500, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Emily M Farris and Heather Silber Mohamed. 2018. Picturing immigration: How the media criminalizes immigrants. *Politics, Groups, and Identities*, 6(4):814–824.

Anjalie Field, Gayatri Bhat, and Yulia Tsvetkov. 2019. Contextual affective analysis: A case study of people portrayals in online #metoo stories. In *Proceedings of the Thirteenth International Conference on Web and Social Media, ICWSM 2019, Munich, Germany, June 11-14, 2019*, pages 158–169. AAAI Press.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies. In *Pro-*

ceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3570–3580, Brussels, Belgium. Association for Computational Linguistics.

Umang Gupta, Jwala Dhamala, Varun Kumar, Apurv Verma, Yada Pruksachatkun, Satyapriya Krishna, Rahul Gupta, Kai-Wei Chang, Greg Ver Steeg, and Aram Galstyan. 2022. Mitigating gender bias in distilled language models via counterfactual role reversal. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 658–678, Dublin, Ireland. Association for Computational Linguistics.

Sophie Jentzsch, Patrick Schramowski, Constantin Rothkopf, and Kristian Kersting. 2019. Semantics derived automatically from language corpora contain human-like moral choices. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 37–44, New York, NY, USA. Association for Computing Machinery.

Daniel Kahneman and Amos Tversky. 1984. Choices, values, and frames. *American psychologist*, 39(4):341.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Shima Khanehzar, Trevor Cohn, Gosia Mikolajczak, Andrew Turpin, and Lea Frermann. 2021. Framing unpacked: A semi-supervised interpretable multi-view model of media frames. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2154–2166, Online. Association for Computational Linguistics.

Shima Khanehzar, Andrew Turpin, and Gosia Mikolajczak. 2019. Modeling political framing across policy issues and contexts. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 61–66, Sydney, Australia. Australasian Language Technology Association.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Discourse probing of pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3849–3864, Online. Association for Computational Linguistics.

Artur Kulmizev, Vinit Ravishankar, Mostafa Abdou, and Joakim Nivre. 2020. Do neural language models show preferences for syntactic formalisms? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4077–4091, Online. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Andrea Lawlor and Erin Tolley. 2017. Deciding who's legitimate: News media framing of immigrants and refugees. *International Journal of Communication*, 11(0).

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868, Online. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. PowerTransformer: Unsupervised controllable revision for biased language correction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7426–7441, Online. Association for Computational Linguistics.

Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. Modeling framing in immigration discourse on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263, Online. Association for Computational Linguistics.

Matthew R. Pearson M.S. 2010. How "undocumented workers" and "illegal aliens" affect prejudice toward mexican immigrants. *Social Influence*, 5(2):118–132.

Brian A Nosek, Mahzarin R Banaji, and Anthony G Greenwald. 2002. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, 6(1):101.

Reidar Ommundsen, Knud S. Larsen, Kees van der Veer, and Dag-Erik Eilertsen. 2014. Framing unauthorized immigrants: The effects of labels on evaluations. *Psychological Reports*, 114(2):461–478. PMID: 24897900.

Rita Omokha. 2022. 'They don't have any humanity': Black immigrants in Ice custody report abuse and neglect.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation frames: A data-driven investigation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 311–321, Berlin, Germany. Association for Computational Linguistics.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334, Copenhagen, Denmark. Association for Computational Linguistics.

Peng Shi and Jimmy Lin. 2019. Simple BERT models for relation extraction and semantic role labeling. *CoRR*, abs/1904.05255.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Arabella Sinclair, Jaap Jumelet, Willem Zuidema, and Raquel Fernández. 2022. Structural Persistence in Language Models: Priming as a Window into Abstract Language Representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050.

Probe 2	AG P TH. TH is MASK powerful than AG.
Prior 2	TH is MASK powerful than AG.
No-Ent 2	P. is MASK powerful than.
Part-Sent 2	TH is MASK powerful than AG.

Table 3: Prior and baselines for Probe 2. The agent (AG), theme (TH), and predicate (P) are instantiated as explained in §3.1.

Göran Sonesson. 1998. Denotation and connotation. *Encyclopedia of semiotics*, pages 187–89.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Paul Trichelair, Ali Emami, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 2019. How reasonable are common-sense reasoning tasks: A case-study on the Winograd schema challenge and SWAG. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3382–3387, Hong Kong, China. Association for Computational Linguistics.

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

A Parameter Settings

Zero-shot Setting We use the pre-trained model implemented in HuggingFace's Transfomers (Wolf et al., 2020) with base version[11] and default model parameters.

Fine-tuning Settings We perform stochastic gradient descent with mini-batches of 64 sentences. We use the Adam optimizer (Kingma and Ba, 2015) with the default parameters, except for the learning rate, which we set to 10^{-5}. We train the models for five epochs or until we reach convergence.

[11] 'bert-base-uncased', 'roberta-base', 'albert-base-v2'

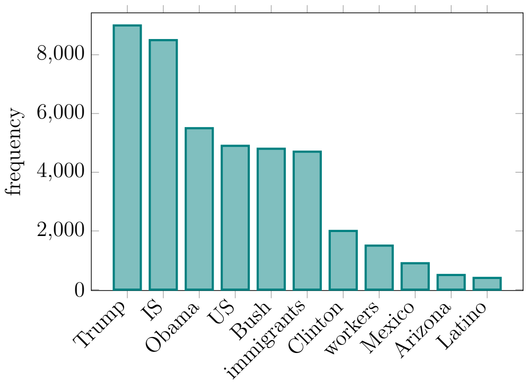


Figure 5: Most common entities such as Immigration Services (IS) and United States (US) in the immigration articles from the MFC (Card et al., 2015).

B Most Common Entities in MFC Immigration

To complement the list of entities extracted by SRL and coreference resolution, we also used NER to extract entities tagged as PERSON, and *string matching* to find all mentions of each entity. We consider only entities mentioned more than three times in each article as the most common entities within each article. Figure 5 shows the most common entities, with the number of their appearance in the immigration articles from the MFC dataset.

C Statistical Significance of the Power Scores in Case Studies

In the case studies in §7, the overall power scores of entities are reported by averaging the power score across the whole corpus. We test the significance of pairwise differences in power score using a z-test assuming a binomial distribution of labels (CS = +1 or -1). Below we discuss the result for each case study separately.

In case study in §7.1 we compare the power score of ‘immigrants’ in articles with different emphasis frames (as shown in Fig 7.1). The majority of comparisons of the power score are significant (as shown in Fig 4); for example, the p-value between frames `Fairness` vs. `legality` is 0.0004, and `Fairness` vs. `Public Sentiment` is 0.001. Overall, differences of power scores between frames of > 0.1 are significant at $p \leq 0.05$. The difference of power scores between frames `Economic` vs. `Security` are not significant ($p = 0.1$).

In the case study in §7.2 the comparisons between power score of ‘immigrants’ vs. ‘immigration services’ in articles with different issue stances (as shown in Fig 4 as same color across blocks) are significant with a p-value of ≤ 0.005. Additionally, the comparison of the power score for each entity in different issue stances (pairwise comparison of bars within blocks) is almost significant with a p-value of $p \leq 0.05$.

	F	E	S	Q	Po	Pu	L
F		*	*	*	*	*	*
E	*		-	*	*	*	*
S	*	-		*	*	*	*
Q	*	*	*		-	-	*
Po	*	*	*	-		-	*
Pu	*	*	*	-	-		-
L	*	*	*	*	*	-	

Table 4: Shows the comparisons of the power score of ‘immigrants’ in different emphasis frames including **F**airness, **E**conomic, **S**ecurity, **Q**uality of Life, **Po**litical, **Pu**blic Sentiment, and **L**egality. The * indicates the difference is significant at $p \leq 0.05$ (z-test) while ‘-’ means insignificant.