Funding Sources and Values of NLP Research

Stanford CS224N Custom Project

Vyoma Raman
vyoma@stanford.edu

Parth Sarin
psarin@stanford.edu

Patricia Wei
patwei@stanford.edu

Abstract

We apply a combination of modeling techniques to investigate how funding sources of natural language processing (NLP) research have influenced what the research values, finding trends over time and level of influence. To identify funding sources, we implemented seven techniques: Regex parsing, fine-tuned RoBERTa classification, and large language model (LLM) prompting to predict the funding sources of each paper, finding that optimized prompting on Mistral does has the best performance, an accuracy of 95% and loss of 0.11. We again applied LLMs to identify values in papers with classification and generation prompts. Then, we implemented Latent Dirichlet Analysis (LDA) and BERTopic to analyze the differences between language used to convey these values.

Mentor: Caleb Ziems.
Sharing project: Vyoma built out a theory of power in NLP for CS 324H.
Team Contributions: All of us helped with data annotation. Vyoma selected and pre-processed our dataset, used Gemini to identify values in the papers, and ran LDA on the results. Patricia wrote the Regex parser (with Parth's help!). Parth and Patricia built the RoBERTa model and augmented the data, and Parth also built the language model classifier and ran BERTopic.
Code: `https://github.com/parthsarin/cs224n-final`

1 Introduction

In the 1950s, following the deliberate efforts of Vannevar Bush of MIT and Fred Terman of Stanford, the field of computer science received substantial funding from the military and corporations. This significantly impacted the direction of purportedly independent research. The U.S. government has gone so far as to intimidate outspoken scientists disagreeing with their priorities, and corporations have done similar: Meta revoked NYU researchers' access to studying the company's role in the January 6th insurrection and Amazon suppressed internal research about "racist logics" in the company's algorithms (Lardner, 1992; Whittaker, 2021). The field of natural language processing (NLP) .also experienced these effects, such as when Google fired Timnit Gebru over her co-authorship a paper critical of large language models (LLMs) (Metz and Wakabayashi, 2020).

To investigate these power dynamics further, we analyze a modified version of the ACL OCL dataset to identify patterns in funding sources and espoused values (including fairness, novelty, reproducibility, and performance) in NLP papers. Specifically, we look at a subset of the dataset containing papers that introduce new tasks and datasets, as those are explicitly intended to be reused in future work and thus represent research priorities in the field. We apply three overarching methods to identify funding sources from the article text: regular expressions; fine-tuned RoBERTa models; and LLMs. We then analyze the values in the papers by applying an LLM in two distinct ways and running topic modeling on text expressing each value to characterize how it is being applied.

We find that using an optimized prompting technique on Mistral 7B Instruct v.0.2 gives an accuracy of 95% and loss of 0.11 in identifying funding sources, outperforming the other models. Second, we devise a LLM-based approach to paper value identification that overcomes hallucination by filtering out responses that are not in the original text. Finally, we present an analysis of the results: that more influential NLP papers tend to disclose corporate and defense funding sources more, that the values of performance and easy implementation drop in prevalence over time while reproducibility increases, and that our corpus shows minor differences in the values of papers funded by the different sources. Ultimately, we hope to encourage greater self-reflection within the NLP research community about where funding comes from and how it affects the values motivating research.

2 Related Work

The task of funding source classification is a form of named-entity recognition (NER). Prior researchers have used ensemble approaches to accomplish this, such as the $FundingFinder$ pipeline, which filters relevant subsections of text and applies sequential learning (Kayal et al., 2019). Transformer-based methods have great potential to be applied to this task. BERT and RoBERTa showed unprecedented performance on closed-form tasks (Devlin et al., 2019; Liu et al., 2019). Transformers are also used in LLMs such as Mistral and Gemini, which are able to do a variety of tasks that they were not explicitly trained on (Team et al., 2023; Jiang et al., 2023). LLMs have been shown to perform well on open-ended social science tasks (Ziems et al., 2024).

Several researchers have identified values prevalent in computing research. Birhane et al. (2021) qualitatively coded a corpus of machine learning papers to identify values that underlay arguments authors were making. Blodgett et al. (2020) did so on papers on NLP "bias" to identify implicit norms conceptualized by researchers. Such work has previously been done using qualitative methods, but computational social scientists have adopted methods like framing (Ali and Hassan, 2022). Topic modeling is a common unsupervised approach to this, and Latent Dirichlet Allocation (LDA) and BERTopic are two common algorithms (Jelodar et al., 2019; Grootendorst, 2022).

3 Data[1]

3.1 NLP Papers

We use the dataset constructed by Held et al. (2023), which builds off the ACL OCL Corpus. ACL OCL contains $73,285$ papers published at a variety of venues and publications affiliated with the Association for Computational Linguistics (ACL) between 1965 and 2022 (Rohatgi et al., 2023). The dataset also includes geographic, language, and citation information. Prior to this quarter, we filtered it dataset to include only papers that introduce tasks and datasets. The final dataset has $29,151$ rows. We sampled $1,000$ papers from the overall dataset for this project.

To build our models, we created a training set and a test set from this dataset. We hand-annotated 60 and 30 rows, respectively, for different funding sources for the classification task described below. For the models trained with curriculum learning, we augmented the training set by replacing randomly-selected sentences in paper chunks with new ones and modifying the labels accordingly. To get these new sentences, we sampled positive and negative sentences related to funding from GPT-4.The augmented training set was only used for the naive RoBERTa model.

3.2 Research Values

Since a key component of our project is to detect the values encoded in the papers in the corpus, we draw on the values identified by Birhane et al. (2021). Their annotation pipeline and results are available on GitHub, which we used to design our methods to capture the presence of values in research papers. Using the links in their dataset, we scraped each of the 100 papers they annotated and used `pdftotext` to parse the text. Since their annotations are at a sentence-level granularity and we are interested in a paper-level granularity, we say that a paper references a value if the number of sentences referencing that value is greater than 0.

Birhane et al. (2021) had two people annotate each paper in the dataset. In our analysis, we only used papers where both annotators agreed on the value expressed because instances where the value is more obvious may be easier to computationally identify.

[1] We elected to break from the provided template in structuring our paper but include all required information.

4 Modeling Methods

4.1 Funding Classification

The funding sources for a paper are generally listed in the “Acknowledgements” section, which is often found at the end of the paper before references. Research in NLP is usually funded by one or more of the following sources: defense and security-focused government departments, scientific innovation-focused government agencies, corporations, and independent grants from academic institutions and philanthropic foundations. There are also papers which do not receive institutional funding.

To identify this information about the papers in our dataset, we took two overarching approaches to multi-label classification of the funding sources of papers: RoBERTa classification and language model question-answering. We compared both to a baseline model we built using regular expressions.

For the Regex model, we compiled a list of known funding organizations and their categorizations based on the training data. At inference, we identified the start of the acknowledgements section of a paper by searching for the Regex pattern “`[aA]cknowledg[e]?ments?`” in the paper. Then, we used keyword searching “`re.search(rf"\b{source}\b", paper))`” to check if any words or phrases in that excerpt was a `source` included in our known set.

4.1.1 RoBERTa Models

We fine-tuned RoBERTa to classify papers based on their funding sources Liu et al. (2019). However, many papers exceed RoBERTa’s context window, so we broke each paper into chunks and ran each chunk through RoBERTa. Depending on the paper, this resulted in up to 50 vectors. Then, we experimented with two ways of combining the results into predictions for the overall paper.

Our first “naive” approach makes predictions on each chunk of a paper and combines the probabilities of each chunk. Our second attention-based approach combines RoBERTa’s outputs using learnable coefficients. For each of these, we tried two versions of training to address GPU constraints: freezing RoBERTa and updating some parameters using LoRA (Hu et al., 2021).

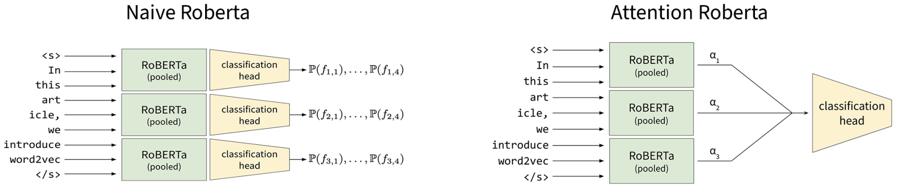

Figure 1: The naive and attention-based RoBERTa models.

Naive RoBERTa Model Suppose a given paper is broken into K chunks, each of which fit into RoBERTa’s context window. Let $f_{k,j}$ be a random variable that represents whether there is evidence of funding source j in chunk k, taking on the value of 1 if there is evidence and 0 otherwise. Similarly, let f_j be 1 if the entire paper was funded by funding source j and 0 otherwise. We learned a classifier to predict $\mathbb{P}(f_{k,1}), \ldots, \mathbb{P}(f_{k,4})$ for each chunk and combined them to get probabilities $\mathbb{P}(f_1), \ldots, \mathbb{P}(f_4)$ for the full paper.

We call this model, illustrated in Figure 1, “naive” because it assumes that the variables $f_{k,j}$ are independent. Then,

$$\mathbb{P}(f_j = 1) = 1 - \mathbb{P}(f_j = 0) = 1 - \prod_{k=1}^{K} \mathbb{P}(f_{k,j} = 0) = 1 - \prod_{k=1}^{K} \left(1 - \mathbb{P}(f_{k,j} = 1)\right),$$

so the overall probability can be represented using the probabilities of the chunks. In practice, we do this with the sum of logs for stability and backpropagate loss at the chunk level. We used our augmented training data for this model.

Attention RoBERTa Model Four our second approach, we combined the outputs of the RoBERTa model using learnable coefficients. We drew on the attention mechanism to combine the vectors, as shown in Figure 1. We had the model learn 50 parameters, $\alpha_1, \ldots, \alpha_{50}$, which were used to create a weighted sum of the RoBERTa outputs. In cases where a paper had less than 50 chunks, we divided the attention weights accordingly. For example, for a paper with two chunks, we split the weights in half: $\widetilde{\alpha}_1 = \alpha_1 + \cdots + \alpha_{25}$ and $\widetilde{\alpha}_2 = \alpha_{26} + \cdots + \alpha_{50}$. This allowed us to learn a combination of the RoBERTa outputs that was sensitive to the location in the paper where funding information appeared. The resulting 768-dimensional vector was passed through a multilayer perceptron.

4.1.2 Language Model Question-Answering

Given the high performance of instruction-tuned language models on a variety of tasks, we also examined their performance on the funding source labeling task using two prompting approaches. We used the 4-bit quantized, instruction-tuned, Mistral 7B Instruct v.0.2 model (Jiang et al., 2023).

Zero-shot Prompting First, we prompted Mistral with the basic version of our task. The prompt contained the paper to label with an instruction to respond “yes” or “no” to whether the paper was sponsored by each funding source or no funding sources. This approach was designed to minimize environmental impact and enable faster labeling since classifying a paper requires five tokens in total and the queries are parallelizable.

Then, we identified a collection of words $\mathcal{Y}$ that mean “yes” and a collection of words $\mathcal{N}$ that mean “no.” We looked at the model’s predicted distribution for the next token w and computed $p_{\text{yes}} = \mathbb{P}(w \in \mathcal{Y})$ and $p_{\text{no}} = \mathbb{P}(w \in \mathcal{N})$. Finally, we rescaled the probabilities so they add up to one. Practically, $\{$“Yes”, “yes”$\}$ carries all the support for $\mathcal{Y}$ and similarly for $\mathcal{N}$.

Optimized Prompting We used a handful of strategies to optimize the prompt beyond our zero-shot approach: beam search for examples, reasoning traces, and sequential classification.

First, we provided examples of organizations that fell into each type of funding source, optimized using beam search. We sampled examples of funding sources from GPT-4 at a high temperature and evaluated each collection of examples based on how well the model performed when prompted with them. GPT-4 received feedback on the surviving example collections, so it could perturb them towards higher performance. At the end, we had a high-performing collection of examples that we used in the final prompt.

We also allowed the model to produce a reasoning trace rather than a single “yes” or “no” token. This resulted in a major computational decline since we had to sample and generate tokens sequentially, but it enabled the model to extract the funding sources as tokens before categorizing them.

Finally, we had the model generate the answers sequentially, predicting “yes” or “no” for each funding source on a new line. This reflects the idea that some combinations of funding sources are impossible (e.g., a paper can’t have been funded by “Defense” and “None”).

5 Value Characterization Methods

5.1 Value Identification

Since the expression of values is often subtle, we used a large language model, Gemini 1.0 Pro, to classify whether a value is in a paper and identify the relevant sentences. We ran both of these approaches on the papers in the Birhane et al. (2021) dataset as well as our dataset. We focused on six of the 67 values in the original paper that we thought could be more reliably identified by a language model: fairness, performance, ease of implementation, reproducibility, build on prior work, and novelty. We selected these based on the annotator agreement in existing annotations, the concreteness of the presence of each value in example sentences provided in the appendix of the ArXiv version of Birhane et al. (2021), whether we felt they would appear in a structured format in the text, and our assessment of how effectively we would be able to identify the values ourselves.

Classification Prompting First, we prompted Gemini to output a binary classification. The prompt contained the paper to label with an instruction to respond “yes” or “no” to whether the paper mentioned a particular value. We retained these responses in binary form to compare them to Birhane et al. (2021)'s annotations.

Generation Prompting Next, we prompted Gemini to output quotes from the paper. The prompt contained the paper to take quotes from with an instruction to return excerpts of the paper relevant to a particular value, including a definition of the value. We crafted the definitions of the value from our background knowledge and from the examples provided in the ArXiv paper. We parsed sentences in the responses to identify the quotes and check that they actually appear in the paper before combining them into a single string.

5.2 Topic Modeling

We used topic modeling to analyze the excerpts. First, we filtered out sentences that do not appear in the paper text to ensure that we did not model hallucinations. Then, we implemented LDA and BERTopic to cluster the paper excerpts into topics Grootendorst (2022).

Latent Dirichlet Allocation As a baseline, we implemented LDA using the `gensim` package (Rehurek and Sojka, 2011). This involved preprocessing and vectorizing the documents, selecting an optimal number of topics, training the LDA model based on the document-term matrix, and assigning topic labels to each document. In the preprocessing step, we made all the text lowercase, filtered out non-word characters using regular expressions, word-tokenized the text, and removed English stopwords. We optimized the number of topics by maximizing the `u_mass` topic coherence score, which measures the semantic similarity between high-scoring words within each topic.

BERTopic We primarily followed the BERTopic procedure in Grootendorst (2022). To do so, we ran the value excerpts through BERT to get a 768-dimensional vector $v_{a,t}$ for every (article, value) pair and reduced the dimensionality of the BERT embeddings. Then, for each value, we clustered the embeddings and used TF-IDF to pick words in each cluster as labels. We had one deviation from Grootendorst (2022) in our preprocessing: we followed the same steps as described in Section 5.2 and concatenated the resulting tokens together with a space. This prevented BERTopic from producing topics that index highly on stopwords and enabled us to compare the two approaches.

6 Evaluation Results

Before examining patterns in funding sources and values in NLP, we assess the performance of our approaches to extract that information.

6.1 Funding

Model	**Test Accuracy**	**Loss**
Regex Parsing	0.79	N/A
Attention RoBERTa model (frozen base)	0.72	1.71
Attention RoBERTa model (LoRA)	0.72	1.33
Naive RoBERTa model (frozen base)	**0.73**	1.16
Naive RoBERTa model (LoRA)	**0.73**	**1.11**
Zero-Shot Prompting Mistral 7B	0.86	0.50
Optimized Prompting Mistral 7B	**0.95**	**0.11**

Table 1: Funding classification performance results.

Following the methodology described in Section 4.1, we implemented seven techniques for identifying funding sources in papers. Table 1 lists the accuracy and loss of each technique, computed as follows. Let $y_{a,c}$ be 1 if article a was funded by funding source c and 0 otherwise, and let $m(a) \in [0,1]^5$ be the model's predictions for article a.

$$\mathcal{A} = \frac{1}{5N} \sum_{a=1}^{N} \sum_{c=1}^{5} 1\{\text{round}(m(a)_c) = y_{a,c}\} \qquad \mathcal{L} = -\frac{1}{5N} \sum_{a=1}^{N} \sum_{c=1}^{5} y_{a,c} \log m(a)_c$$

The optimized prompting on Mistral had the highest accuracy and lowest loss, and the naive RoBERTa model had the best loss among that type of classifier.

Further qualitative evaluation showed us a few features of the approaches. With optimized prompting, Mistral often assigned a high probability to the incorrect label, which is consistent with the literature on hallucinations (Huang et al., 2023). For a few documents in our data, we found that the acknowledgements information was included in the original paper but did not appear in the parsed version in the ACL OCL dataset. This reflects the challenge with applying datasets to purposes for which they were not explicitly designed.

We also saw that our RoBERTa models had lower accuracy compared to our Regex baseline model. We had to train naive RoBERTa on the chunk level due to context constraints but can only evaluate on the document level. We further note the limitation of our small test set and training set, which did not capture all of the possible funding organizations.

6.2 Value Identification

Metric	Novel	Performance	Reproduc.	Fair	Easy Implem.	Recent W.	Classic W.
FPR	1.00	0.54	0.56	0.00	0.43	0.93	0.45
FNR	0.04	0.14	0.00	N/A	0.20	0.05	0.52
F1	0.84	0.87	0.05	1.00	0.22	0.86	0.38
Comp. F1	0.70	0.78				0.30	0.00

Table 2: False positive rate, false negative rate, and F1 score of LLM value identification.

To test the ability of an LLM to identify values, we queried Gemini 1.0 Pro and binarized model responses and Birhane et al. (2021)'s responses as described in Section 5.1 for evaluation. The results for the classification task are listed in Table 2, alongside an evaluation by Birhane et al. (2021) on a simple logistic regression model. In our model, novelty, performance, and building on recent work are identified with high F1 scores, but the false positive rates are also high, indicating strong recall because the number of true positives of the model balances the number of false positives. For novelty and performance, our F1 scores are similar to Birhane et al. (2021). Our qualitative evaluation explains some of these results: an analysis of the makeup of the dataset highlights significant class imbalance, which can cause high F1 and high FPR. We also found that text within each paper rarely contains explicit keywords for each value, which means that the model lacks a clear signal.

When analyzing the excerpts produced by Gemini, we found that 51% of papers had hallucinated results. We qualitatively analyzed a sample of non-hallucinated sentences and found varying degrees of expression of the value in every excerpt that we read. This corroborates the observed high false positive rate because it suggests that Gemini picked up on even slight expressions of a value.

6.3 Topic Modeling

We ran topic modeling on non-hallucinated sentences as described in Section 5.2. First, we computed C_{UMass} topic coherence for topics produced through both topic modeling algorithms. Topic coherence is a measure of the quality of topics produced by topic modeling algorithms, focusing on interpretability and semantic consistency. C_{UMass} is calculated based on document co-occurrence statistics, focusing on log probabilities of top words in each topic. That is, $C_{\text{UMass}} \propto \log(p(w_i, w_j)/p(w_j))$, where w_i and w_j are words in the topic. We selected this metric because ACL OCL is a highly specialized corpus and external word distributions would be different. Second, we qualitatively evaluated the quality of topics on a scale from 1 to 5 based on the interpretability of topic words, where 1 is poor quality and 5 is high quality. The results of these evaluations are shown in Table 3.

Our results are quite mixed. Topic coherence scores show that LDA is a superior algorithm for having high coherence in all values but novelty. However, we explicitly optimized for this when selecting the number of topics in LDA, which we did not for BERTopic. That makes this measure less valid for comparing algorithms. For human-evaluated quality, the results are very mixed, with each algorithm doing better for half of the values. While the topics were somewhat interpretable, they did not reflect the kinds of changes we are looking to track over time. For example, for performance we are interested in what makes a "good" model. Researchers may prioritize qualitative evaluation,

Algorithm	Evaluation	Novel	Performance	Reproduc.	Fair	Easy Implem.	Past W.
LDA	*# Topics*	2	2	2	4	30	3
	Coherence	-1.50	**-1.32**	**-1.69**	**-5.52**	**-6.84**	**-1.33**
	Quality	**3**	3	2	**2**	**4**	3
BERTopic	*# Topics*	2	10	4	2	2	12
	Coherence	**-1.37**	-3.26	-3.64	-14.29	-7.83	-2.60
	Quality	2	**5**	**3**	1	2	**5**

Table 3: Number of topics, topic coherence scores, and quality scores of LDA and BERTopic topics.

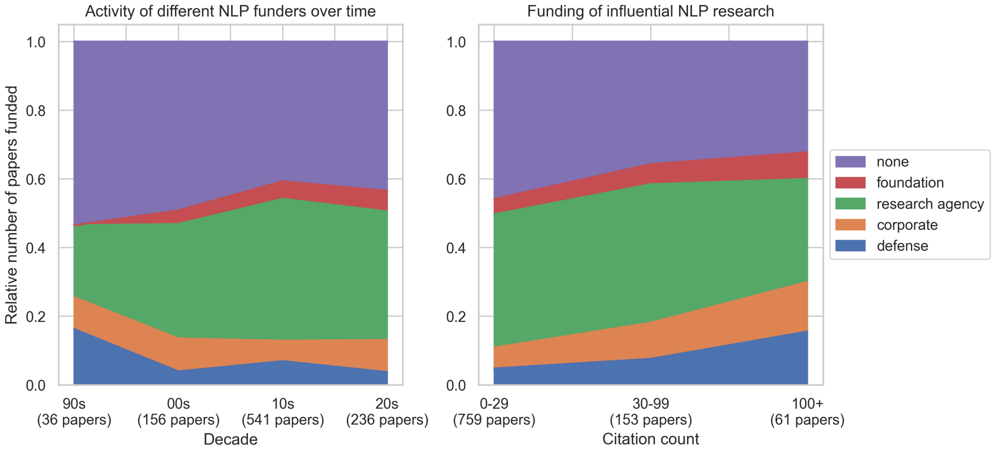

Figure 2: Funding over time, and based on citation count, of the papers we sampled from the corpus.

benchmark scores, or statistical metrics among other types of performance. BERTopic had the highest quality topics for that value, but they only reflected differences in areas of NLP ("parsing" vs. "bleu" vs. "asr"), which is not helpful for our analysis. Our interest is in frames rather than literal topics.

7 Funding and Values in NLP

Here, we apply our models to analyze the relationship between funding sources and values in NLP. We took the best method for each task and ran them on a random sample of $1,000$ papers from the ACL OCL dataset to label funding sources and values.

First, we examine funding. Figure 2 shows how funding of NLP papers have changed over time and based on the influence of papers. Our y-axis represents the level of activity of a funding source in a particular decade or citation bucket. The graph illustrates a decline in evidence of defense and security funding between the 1990s and 2000s, corresponding with an increase in disclosures of funding received from scientific research agencies. This uptake aligns with the end of one of the so-called "AI winters." Foundation funding activity also rose in that time period. We found more evidence of funding in influential papers than in less influential papers, with influence approximated by the number of citations. And, papers with the highest citations tended to report defense and corporate funding more than papers with the lowest citations.

Next, we examine values. Figure 3 shows the same breakdown of time and influence for the values espoused by papers. The time-based trends reflect typical patterns for a growing field: emphasis on performance and ease of implementation has decreased as hardware and software capabilities have increased; reproducibility and building on past work has become more emphasized as there is more past work to build on; and novelty has remained a fairly consistent value. There is little observable difference in values across papers with different levels of influence.

Finally, Figure 4 shows the value distributions of papers in our dataset funded by different sources. We observe that corporate-funded papers and papers without a listed funding source are more likely

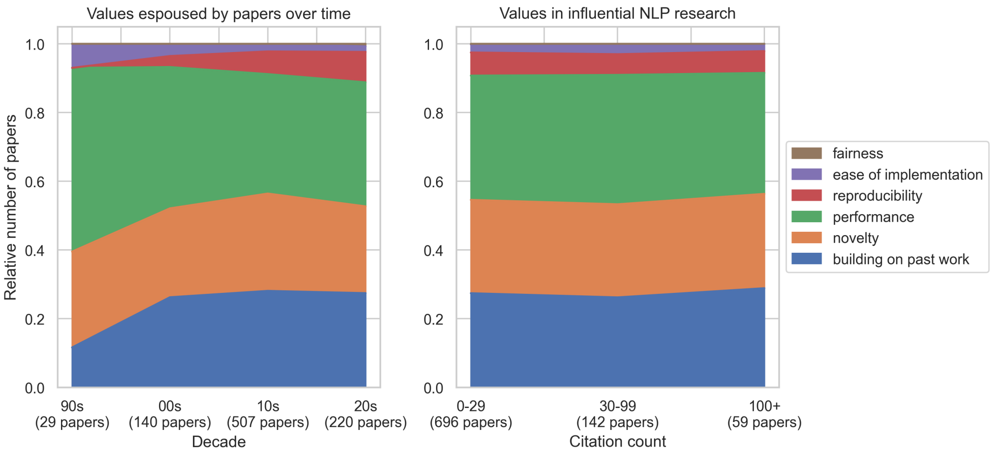

Figure 3: Proportion of papers over time and in each influence category that espouse the values we searched for. If Gemini did not identify any value excerpts, we excluded the paper from this figure.

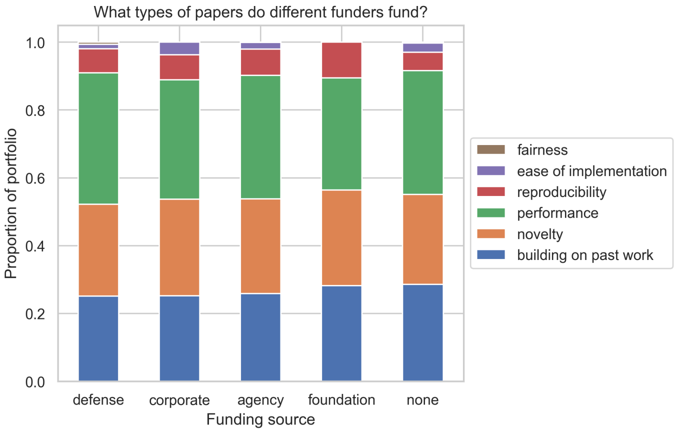

Figure 4: The different funding sources largely fund the same papers, in terms of values espoused.

to discuss ease of implementation, while foundation-funded papers have a slightly lower focus on performance. However, the distributions seem are largely identical, and our visualization does not capture differences over time or value-normalized patterns.

8 Conclusion

To identify funding sources in NLP research papers, we experimented with seven techniques of three overarching types. The large language prompting models had the highest accuracy and lowest loss, followed by the Regex Model, and the RoBERTa models had the lowest accuracy. While we found topic modeling insufficient to characterize the frames of values, we identified values in papers by selecting excerpts. This method has proven to be qualitatively successful and resistant to hallucination due to a filtration step. Our results also demonstrate several patterns in the distribution of funding sources and values over time as well as the relationship of those factors to each other.

Future work could expand on our analysis of funding and values in several ways. We found that topic modeling was inadequate to capture framing in an unsupervised fashion, and there is a need for better methods. Such framing should be explored more deeply with respect to how values in NLP research are expressed in different ways over time and by researchers with different funding sources and sub-disciplines. We hope that this work spurs interest in critically analyzing how the values of research are influenced by the power structures in which it is implicated.

References

Mohammad Ali and Naeemul Hassan. 2022. A survey of computational framing analysis approaches. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9335–9348.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2021. The Values Encoded in Machine Learning Research. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, 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.

Maarten R. Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

William Held, Camille Harris, Michael Best, and Diyi Yang. 2023. A Material Lens on Coloniality in NLP.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *ArXiv*, abs/2106.09685.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ArXiv*, abs/2311.05232.

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia tools and applications*, 78:15169–15211.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Subhradeep Kayal, Zubair Afzal, George Tsatsaronis, Marius Doornenbal, Sophia Katrenko, and Michelle Gregory. 2019. A framework to automatically extract funding information from text. In *Machine Learning, Optimization, and Data Science: 4th International Conference, LOD 2018, Volterra, Italy, September 13-16, 2018, Revised Selected Papers 4*, pages 317–328. Springer.

George Jr. Lardner. 1992. Army Accuses SDI Critic of Falsifying Credentials, Scientist's Security Clearance is Suspended. *The Washington Post*.

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. *arXiv preprint arXiv:1907.11692*.

Cade Metz and Daisuke Wakabayashi. 2020. Google Researcher Says She Was Fired Over Paper Highlighting Bias in A.I. *The New York Times*.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. 2023. The ACL OCL Corpus: Advancing Open Science in Computational Linguistics. *arXiv preprint arXiv:2305.14996*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Meredith Whittaker. 2021. The steep cost of capture. *Interactions*, 28:50 – 55.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, pages 1–55.