

“We Demand Justice!”: Towards Grounding Political Text in Social Context

Rajkumar Pujari
Purdue University, USA
rpujari@purdue.edu

Chengfei Wu
Purdue University, USA
wu1491@purdue.edu

Dan Goldwasser
Purdue University, USA
dgoldwas@purdue.edu

Abstract

Social media discourse from US politicians frequently consists of ‘seemingly similar language used by opposing sides of the political spectrum’. But often, it translates to starkly contrasting real-world actions. For instance, “We need to keep our students safe from mass shootings” may signal either “arming teachers to stop the shooter” or “banning guns to reduce mass shootings” depending on who says it and their political stance on the issue. In this paper, we define and characterize the context that is required to fully understand such ambiguous statements in a computational setting and ground them in real-world entities, actions, and attitudes. To that end, we propose two challenging datasets that require an understanding of the real-world context of the text to be solved effectively. We benchmark these datasets against baselines built upon large pre-trained models such as BERT, RoBERTa, GPT-3, etc.. Additionally, we develop and benchmark more structured baselines building upon existing *Discourse Contextualization Framework* and *Political Actor Representation* models. We perform analysis of the datasets and baseline predictions to obtain further insights into the pragmatic language understanding challenges posed by the proposed social grounding tasks.

1 Introduction

In the past decade or so, micro-blogging websites, especially Twitter have become the primary channel of political engagement between politicians and the general population in the US. As a result, discourse from politicians became more succinct. Often, politicians from the same party coordinate their responses to developing events to amplify the impact of their intended message (Weber and Neumann, 2021; Vaes et al., 2011). Hence, repetitive phrases such as ‘Thoughts and Prayers’ are extensively used to signal more nuanced stances across several events.



Figure 1: An example of varied *intended meanings* behind the same political message depending on the Author and Event in context

Further, these platforms also allow for interactivity among politicians from opposing parties. This often results in messages that are phrased similarly but intended to signal starkly contrasting real-world actions. ‘Thoughts and Prayers’ is often used to express condolences by many Republicans in mass shooting events. In contrast, the same phrase is used in an angry or sarcastic tone by Democrats signaling a call for action demanding ‘tighter gun control measures’. The example in Figure 1, “We need to keep our teachers safe!” shows various manifestations of a common political message that signals different interpretations depending on the context and the author.

Humans familiar with the historical stances of a politician and, possessing specific knowledge about the event from the news, are able to understand the *intended meaning* of such messages. But, computationally understanding the correct meaning of such language is challenging. Our main question in this paper is - *Can an NLP model find the right meaning? Social Context Understanding*, needed for building successful models for such tasks, can

Tweet Target Entity and Sentiment		Vague Text Disambiguation	
Tweet: I believe any credible and thorough FBI investigation should include interviews with the two key witnesses – Judge Kavanaugh and Dr. Ford. That did not happen. Event: Kavanaugh Supreme Court Nomination Author: Doris Matusi (Democratic Politician) Targets: FBI (negative), Kavanaugh (negative), Christine Ford (positive), Donald Trump (negative)		Vague Text: First, but not the last. Event: US withdraws from Paris climate agreement that enforces environmental targets after three years Author Party: Republican Disambiguation: The withdrawal from the Paris climate agreement is the first step of many to come for the Trump administration. It will not be the last, as more positive changes are sure to follow.	
Target Task Data Statistics		Vague Text Data Statistics	
Unique Tweets	865	Unique Vague Texts	93
Positive Targets	1513	Positive Examples	739
Negative Targets	1085	Negative Examples	2217
Neutral Targets	784	Total Examples	2956
Non-Targets	2509	Number of Events	9
Total Data Examples	5891		
Number of Events	3		

Incorrect Disambiguations:
1) Joe Biden’s inauguration marks the first day of a new era of progress and prosperity, more lasting positive changes are coming. (Incorrect Event)
2) The Paris climate agreement withdrawal is the first of many backward steps this Trump administration is sure to take in destroying our environment. (Incorrect Stance)
3) This is the time for America to move forward and make progress without being held back by a global agreement that doesn’t serve our interests. (Doesn’t match the vague text)

Table 1: Examples of Annotated Datasets and their statistics

come from a wide variety of sources. The politician’s party affiliation, historical stances on the issue, social relationships; knowledge about the involved entities, and related prior and upcoming events, etc, are a few examples of the relevant social context.

In the example in figure 1, for event #1, we could augment the information that Kamala Harris almost always expresses negative sentiment towards the entity guns in her discourse and Mike Pence often expresses positive sentiment. Jointly modeling this information along with the event description could allow us to easily disambiguate the intended meaning. This example motivates the need for tasks and models that jointly model social context. Most of the existing works study the impact of modeling social context either: 1) as the impact of individual aspects of social context (Yang et al., 2016) or 2) in learning representations of specific contextual features from the task objective itself (Karimi et al., 2020; Ansolabehere et al., 2001; Mehta and Goldwasser, 2021). Recently, two models that aim to capture a unified view of the social context were proposed (Pujari and Goldwasser, 2021; Feng et al., 2022).

In this paper, we develop tasks that break down ‘Social Context Grounding’ into three foundation components: *target entities*, *sentiment towards the targets*, and *intended meaning*. To that end, we present two tasks, namely, ‘Target Entity and Sentiment Detection’ and ‘Vague Text Disambiguation’. In ‘Target Entity and Sentiment Detection’, the goal is to predict: 1) whether a given entity is the *intended target* of the given tweet from a known politician and 2) the sentiment towards the intended targets. The data consists of tweets that *don’t always mention the targets* in their text. In ‘Vague

Text Disambiguation’, given an ambiguous message such as “*We demand justice*”, an associated event, and the author’s party affiliation, the task is to identify a *plausible* unambiguous explanation of the message. We show examples of each task in table 1. We describe the datasets and data collection process in detail in §3.

We use the annotations from ‘Target Entity and Sentiment Detection’ dataset to visualize the discourse in the context of a recent US political event, namely *Brett Kavanaugh Supreme Court Nomination* in §6.3. We perform a *human study* on a subset of ‘Vague Text Disambiguation’ dataset. Humans achieve 94.85% accuracy on the task, demonstrating that the task is solvable for humans with reasonable knowledge about the events. We also perform a human evaluation of LLM (GPT-3, GPT-NeoX) generations for *vague text disambiguation*. GPT-3 is able to generate reasonable explanations 73.26% of times compared to GPT-NeoX (20.04%) and human workers (79.80%). We discuss these experiments further in §6.

Further, we evaluate the performance of *four* types of models for the tasks:

1. Pretrained Language Model (PLM) baselines
2. Static contextualized embedding baselines
3. Large Language Model (GPT-3) based in-context learning
4. Discourse contextualization models

Our empirical results demonstrate that discourse contextualization models outperform the other models in all the tasks. We also find that static contextualized embeddings work better than PLM-based embeddings and GPT-3 performs better than text baselines but worse than discourse contextualization models. We present an error analysis of the outputs on each of the tasks to gain further in-

sights. We describe the models in §4 and present the results in §5.

Our contributions can be summarized as:

1. Defining and operationalizing the ‘Social Context Grounding’ task in political discourse
2. Evaluating various state-of-the-art context representation models on the task. Adopting existing discourse contextualization framework for the proposed tasks and evaluating GPT-3 in-context learning performance on the tasks
3. Performing human studies to benchmark the dataset difficulty and GPT-3 generation performance comparison with human workers¹

2 Related Work

Modeling social context is necessary to achieve human-level performance in natural language understanding (Hovy and Yang, 2021). In political discourse, messages are often targeted at the voter base who are aware of the political context (Weber and Neumann, 2021; Vaes et al., 2011). Hence, they are succinct and vague by design. This increases the importance of explicit social context modeling even more.

Previous works tried to model social context in the form of social representation for entity linking (Yang et al., 2016), social media connections relationship for fake news detection (Baly et al., 2018; Mehta et al., 2022) and political bias detection (Li and Goldwasser, 2019; Baly et al., 2020). But, all these works model specific aspects of social context that are relevant to the task in focus.

Recently, two discourse contextualization models that aim to capture social context as a whole were proposed (Pujari and Goldwasser, 2021; Feng et al., 2022). Feng et al. (2022) evaluate their model on bias detection and hyper-partisan argument detection in news articles. Pujari and Goldwasser (2021) evaluate their model on grade prediction and roll call vote prediction.

Although, evaluation tasks presented in Pujari and Goldwasser (2021) and Feng et al. (2022) show interesting social context understanding, these tasks are not fully representative of the challenges posed by ‘Social Context Grounding’. There is a need for evaluation tasks that benchmark political social context understanding of NLP models. Zhan et al. (2023) is one such dataset for dialogue understanding. They address general social com-

monsense understanding. In our work, we target, political domain understanding.

3 Datasets

We design and collect two datasets for ‘Social Context Grounding’ evaluation. As discussed in §1, we focus on mainly on three components of ‘Social Context’: *target entities*, *sentiment towards the targets*, and *intended meaning of vague text with context*. *Tweet Target Entity and Sentiment* task requires identifying *targets* and *sentiment towards them*. *Vague Text Disambiguation Task* evaluates identifying the correct *intended meaning*.

In the *Tweet Target Entity and Sentiment* dataset, we collect annotations of real opinionated tweets from known politicians for their intended targets and sentiment towards them. We focus on *three* political events for this task. The dataset and its collection process are described in detail below in §3.1.

In the *Vague Text Disambiguation Task*, we collect plausible explanations for a set of vague texts, given its context. The context consists of *author party affiliation* and *specific event*. We focus on *eight* political events. This dataset is detailed in §3.2. Examples and data statistics are shown in table 1.

3.1 Tweet Target Entity and Sentiment Task

In this task, given a tweet T , its context, and an entity E , the objective is to predict whether or not E is an intended target of T and the sentiment towards E . Political discourse often contains opinionated discourse about world events and social issues. Understanding the entities which are the focal points of the discussion and the sentiment expressed toward them forms a major part of understanding the discussion. We collect tweets that don’t directly mention the target entities. Thus, understanding the event details and the author’s general stances about the entities involved in the event is necessary to solve this task effectively.

We focus our tweet collection on three recent divisive events: *George Floyd Protests*, *2021 US Capitol Attacks*, and *Brett Kavanaugh’s Supreme Court Nomination*. We pick the events such that they have clearly contrasting talking points from the left and right-leaning politicians in the US. We identify relevant participating entities for each of the three events. Examples of focal talking points and involved entities for the event *George Floyd*

¹We submitted our code. Our data and code for baselines will be publicly released upon acceptance

Protests are listed below. Details for other events are included in the appendix.

- **Left:** police brutality, systemic racism against black people, need for action to cleanse the police structures
- **Right:** support for the police organization while condemning the specific incident, general lawlessness, and violence of the protests, corruption among BLM organization

Involved Entities: *George Floyd, United States Police, Derek Chauvin, Donald Trump, Joe Biden, United States Congress, Black people, Democratic Party, Republican Party, BLM, Antifa*

3.1.1 Data Collection Process

We filter 3,454 tweets for the *three* events using hashtags, keyword-based querying, and the dates of the event-based filtering from the Congress Tweets repository corpus². We collect a subset of 1,779 tweets that contain media (images/video) to increase the chances of the tweet text not containing the target entity mentions. Then, we use 6 in-house human annotators and Amazon Mechanical Turk (AMT) workers who are familiar with the event context for annotation. We ask them to annotate the targeted entities and sentiments towards the targets. The authors of this paper also participated in the annotation process. We provide them with entity options based on the event in the focus of the tweet. Annotators are allowed to add additional options if needed. We also ask the annotators to mark non-targets for each tweet. We instruct them to keep the non-targets as relevant to the event as possible to create harder negative examples. Each tweet is annotated by three annotators. We filter 865 unique tweets with 5,891 annotations, with majority agreement on each tweet. All the AMT annotations were additionally verified by in-house annotators for correctness. AMT workers were paid USD 1 per tweet. It took 3 minutes on average for each assignment, resulting in an hourly pay of USD 20. We include screenshots of the collection task GUIs in the appendix.

We split the train, and test sets by events, authors, and targets to make test the general social grounding capabilities of the models. The test set also consists of authors, targets, and events not seen in the training set. We use *Capitol Riots* event for the test set of *Target Entity and Sentiment Task*. We split the examples into 4,370 train, 511 develop-

ment, and 1,009 test examples.

3.2 Vague Text Disambiguation Task

The task of *Vague Text Disambiguation* evaluates the model’s ability to identify a plausible explanation of an ambiguous quote given the event context and author affiliation. The rationale behind this task is that “*ambiguous language could be assigned grounded meaning when we know who is saying it and in which context*”. For instance “*protecting our children from mass shootings*” could easily be disambiguated as either “*ban guns*” or “*arm teachers*” when we know the stance of the politician on the issue of ‘*gun rights*’.

We collect ambiguous quote candidates that have clear opposing interpretations in the context of the given event. We use AMT workers to write the initial explanations for the quotes. We validate and filter the data using expert annotation. We use the validated data as in-context examples for large language models to generate more candidates. Human annotators filter good generations while also analyzing the performance of LLMs on this task. These data points are further added to the dataset.

3.2.1 Data Collection

We focus our vague text collection on the tweets posted by politicians (i.e. senators and representatives) during the years 2019 to 2021 from Congress Tweets corpus. The tweets are collected automatically on a daily basis. We identify a list of 9 well-known events that happened during that period (including related entities) and use the number of news reports related to each event to determine the duration of each event. We then get the tweets posted during the time frame we predict. By using a pre-trained named entity recognition (NER) model based on BERT (Devlin et al., 2019), we collect tweets that do not contain entity mentions to identify potential candidates for vague texts that can be interpreted in opposing senses. We manually inspect and identify such examples.

For the total of 93 ambiguous tweets that we found, we match them with different events that we find them suit to and use both Democrat and Republican as the author party affiliation. Using the above method, we are able to generate a total of 600 examples from AMT. For each tweet, we ask two questions to AMT workers: 1) we ask the worker to identify the sentiment towards the three most relevant entities in the event and 2) write a

²<https://github.com/alexlitel/congresstweets>

detailed explanation of the tweet given the event and author’s party affiliation. After this step, each example was manually screened by in-house annotators to verify correctness. We then ask three domain experts to vote on whether the annotation correctly reflects its given context. Using manual inspection we are able to create a total of 374 good examples.

Using these good examples, we are able to generate more examples using large language models (LLMs) by using these examples as few-shot examples in the prompt. We use both GPT-NeoX (Black et al., 2022) and GPT-3 (Brown et al., 2020a) for candidate generation. For each generated answer, a manual inspection is also held to ensure the quality. Among all the 919 examples generated by GPT-NeoX and GPT-3, we are able to obtain a total of 650 good generations. After removing redundant samples, we obtain 365 examples. Hence, we obtain a total of 739 annotations. For each of the examples, we ask in-house annotators to select 3 relevant negative options. We instruct them to pick hard examples that might contain similar entities as the correct interpretation. We discuss the results of human validation in §6.

Similar to the previous task, we split the train, test sets by events, and vague text to test the general social understanding capabilities of the model. We reserve *Donald Trump’s second impeachment verdict* event for the test set. We also reserve Democratic examples of 2 events and Republican examples of 2 events exclusively for the test set. We split the dataset into 1,908 train examples, 460 development examples, and 580 test examples.

4 Baselines

As mentioned in §1, we perform baseline experiments in *four* stages: Pretrained Language Model (PLM) baselines, GPT-3 baselines, static contextualized embedding baselines and discourse contextualization framework baselines. The main objective of our experimental design is to evaluate various types of context information inclusion into the models on the two datasets. We use BERT (base and large) (Devlin et al., 2019), RoBERTa (base and large) (Liu et al., 2019) PLMs for our experiments. We report the results of all our baseline experiments in tables 2 and 3.

In the first phase, we evaluate the performance of the fine-tuned pre-trained language models (PLM) on both tasks. We experiment with three types

of contextual inclusion in this phase. First, we don’t include any contextual information. Second, we include the author’s Twitter bio information. Finally, we evaluate the information from author, event, and target Wikipedia page embeddings.

In the second phase, we evaluate GPT-3 in zero-shot and four-shot in-context learning modes on both datasets. We provide contextual information in the prompt in the form of short event descriptions and authors’ party affiliations. As GPT-3 is trained on news data until September 2021, GPT-3 is trained on data about most of the events in our dataset.

In the third phase, we use frozen embeddings from Political Actor Representation (PAR) (Feng et al., 2022) and Discourse Contextualization Framework (DCF) (Pujari and Goldwasser, 2021) models. We use the PAR author embeddings available on their GitHub repository. For DCF, we obtain the models from their repository and generate author, event, text, and target embeddings using available model parameters.

In the third phase, we use the tweets of politicians from other events and build discourse contextualization graphs for our data as proposed in Pujari et al. (2022). We use Wikipedia pages of authors, events, and targets to add contextual information to the graph. We train the DCF model for each of the three tasks.

In the second phase, we use politician embeddings generated using the Discourse Contextualization Framework (DCF) (Pujari and Goldwasser, 2021) and Political Actor Representation (PAR) framework (Feng et al., 2022) for the classification and multiple-choice selection variants on both tasks.

In the third phase of our experiments, we back-propagate to the Discourse Contextualization Framework (DCF) and evaluate the models on the classification and multiple-choice selection variants of both tasks.

Details of our baseline experimental setup are discussed in §4.1. Results of our baseline experiments are discussed in section §5.

4.1 Experimental Setup

Target Entity identification is designed as a binary classification task. Inputs are a tuple of (*author, event, tweet, entity*). Output prediction is whether or not the entity is an intended target of the tweet. Sentiment identification is designed as a 4- class

Model	Target Identification				Sentiment Identification			
	Prec	Rec	Macro-F1	Acc	Prec	Rec	Macro-F1	Acc
PLM Baselines - No Context								
BERT-large	69.09	72.35	68.83	70.56	58.74	60.17	58.95	58.37
RoBERTa-base	66.58	69.54	65.14	66.40	61.68	61.27	61.36	60.65
PLM Baselines + Twitter Bio Context								
BERT-large + user-bio	69.03	71.86	69.34	71.66	60.02	60.44	60.13	59.86
RoBERTa-base + user-bio	65.83	68.65	64.79	66.30	60.06	59.91	59.94	59.46
PLM Baseline + Wikipedia Context								
BERT-large + wiki	63.58	65.78	60.33	61.05	53.48	56.44	53.9	53.32
RoBERTa-base + wiki	69.02	72.32	68.62	70.27	57.62	59.1	58.07	58.28
LLM Baseline								
GPT-3 0-shot	69.25	70.58	69.77	73.78	56.2	55.04	54.18	56.80
GPT-3 4-shot	69.81	72.99	66.45	67.03	58.12	57.10	55.00	57.51
Static Contextualized Embeddings								
BERT-large + PAR Embs	65.4	67.33	60.25	60.56	55.24	57.54	55.89	55.80
RoBERTa-base + PAR Embs	68.38	71.63	67.67	69.18	55.01	56.89	55.51	55.40
BERT-large + DCF Embs	68.76	72.02	68.32	69.97	61.59	63.25	61.22	60.75
RoBERTa-base + DCF Embs	72.89	75.95	73.56	75.82	63.05	63.52	62.90	63.03
Discourse Contextualization Model								
BERT-large + DCF	71.12	74.61	71.17	72.94	65.81	65.25	65.34	65.31
RoBERTa-base + DCF	70.44	73.86	70.39	72.15	63.45	63.34	63.37	63.23

Table 2: Results of baseline experiments on *Target Entity* and *Sentiment* test sets. *Target Identification* is a binary-classification task. *Sentiment Identification* is a 4-class classification task. We report macro-averaged Precision, macro-averaged Recall, macro-averaged F1, and Accuracy metrics.

Model	Vague Text Disambiguation			
	Prec	Rec	Macro-F1	Acc
PLM Baselines - No Context				
BERT-large	52.24	55.58	50.28	53.75
RoBERTa-base	55.3	51.82	54.53	56.08
PLM Baseline + Wikipedia Context				
BERT-large + wiki	52.31	46.9	66.87	76.03
BERT-base + wiki	51.85	38.62	64.36	75.69
LLM Baseline				
GPT-3 0-shot	63.10	62.92	62.58	63.5
GPT-3 4-shot	62.05	62.29	61.86	62.04
Static Contextualized Embeddings				
BERT-large + PAR	47.68	49.66	65.53	73.79
BERT-base + PAR	45.93	54.48	65.49	72.59
BERT-large + DCF Embs	47.18	63.45	67.55	73.10
BERT-base + DCF Embs	56.58	59.31	71.71	78.45
Discourse Contextualization Model				
BERT-large + DCF	52.76	59.31	69.94	76.55
BERT-base + DCF	52.73	60.00	70.06	76.55

Table 3: Results of baseline experiments on *Vague Text Disambiguation* dataset test split. This task is designed as a binary classification task. We report macro-averaged Precision, macro-averaged Recall, macro-averaged F1, and Accuracy metrics

classification task. Inputs are the same as the target detection task, predictions are one of [*positive*, *neutral*, *negative*, *non-target*]. Vague text identification is designed as a binary task. Inputs are (*party*, *vague text*, *event*, *explanation*). Output prediction is whether or not the explanation matches

the given context.

In the first phase, for no context experiments, we use the author name, event name, tweet text, and target name embeddings obtained using PLMs as inputs to an MLP classifier. We concatenate the embeddings as input. For Twitter bio context experiments, we replace author embedding with the PLM embeddings of their Twitter bio. For Wikipedia context experiments, we replace author, event, and target embeddings by the PLM embeddings of their respective Wikipedia pages. The wiki-context baseline contains all the information sufficient for humans to solve both tasks.

In the second phase, we use train split examples as in-context few-shots to obtain GPT-3 results. We provide a task description and event description summaries in the prompt.

In the third phase, for PAR experiments we replace the author embedding with the PAR embeddings released on their GitHub repository. We replace missing author embeddings with their Wikipedia embeddings. For the *Vague Text* task, we average PAR embeddings for all politicians of the party to obtain party embeddings. For DCF embeddings, we are able to generate embeddings for the author, event, tweet, entity, vague text, and explanation using context graphs we build using author

tweets from other events. We build the graph using author, event, tweet, other tweets and target entity nodes for the first task. We use similar graphs for the second task as well.

In the third phase, we back-propagate the loss from the task training to the DCF model parameters. This allows us to train the DCF representations for our tasks.

We use the HuggingFace Transformers (Wolf et al., 2020) library implementations of PLMs. We use GPT-NeoX (Black et al., 2022) and GPT-3 (Brown et al., 2020b) via OpenAI API for our LLM-based experiments. We run 100 epochs for all experiments. We use 10 NVIDIA GeForce 1080i GPUs for our experiments. We use the train, development, and test splits detailed in §3 for our experiments. We use the development macro-F1 score as our early stopping criterion.

5 Results

We show the results of our baseline experiments in tables 2 and 3. We evaluate our models using macro-averaged precision, macro-averaged-recall, macro-F1, and accuracy scores. As the classes are not balanced, macro-F1 is the main metric we use to benchmark the results.

In the Target Entity task, we observe that RoBERTa-base + DCF embeddings obtain the best performance of F1 73.56 compared to 68.83 F1 for the best no-context baselines. Twitter bio context and wiki-context hardly improve upon Target Entity and Sentiment tasks. This shows the importance of modeling contextual information explicitly. The wiki-context model has all the information necessary to solve the tasks but as the information is not explicitly modeled, the performance doesn't improve. No context performance well above the random performance of 50 indicates the bias in the target entity distribution among classes. We discuss this in §6.3. In the Sentiment detection task, we see that BERT-large + DCF back-propagation model outperforms all the other models with an F1 score of 65.34. Wiki-context models perform worse than no-context baselines indicating that PLM embeddings might be adding noise to the task.

In the *Vague Text Disambiguation* task results in table 3, we see that DCF models outperform other models significantly. An F1 score of 71.71 is obtained for BERT-base + DCF embeddings. BERT-base performance well as opposed to bigger models might be due to the DCF model's original learn-

ing tasks being trained on BERT-base embeddings. For this task, we initiate DCF experiments with the trained MLP layers of the wiki-context model. We observe that this initialization gives a performance increase of around 0.85 F1 points on the best model.

Overall, the results indicate that explicitly modeling social context helps with these tasks. DCF model mainly represents the context in the form of text documents for all nodes. Further symbolic addition of other types of context such as social relationships among politicians and relationships between various nodes could further help in achieving better performance on these tasks.

6 Analysis and Discussion

6.1 Vague Text LLM Generation Quality

We look into the generation quality of our LLM-generated disambiguation texts. While GPT-NeoX (Black et al., 2022) produced only 98 good examples out of the 498 generated instances with the rest being redundant, GPT-3 (Brown et al., 2020a) performed much. Among the 430 instances that are generated, 315 of them are annotated as good which converts to an acceptance rate of 20.04% for GPT-NeoX and 73.26% for GPT-3 respectively. In-house annotators evaluated the quality of the generated responses for how well they aligned with the contextual information. They rejected examples that were either too vague, align with the wrong ideology, or were irrelevant. In the prompt, we condition the input examples in all the few shots to the same event and affiliation as the input vague text. In comparison, the validation of AMT annotations for the same task yielded 79.8% good examples even after extensive training and qualification tests. Most of the rejections from AMT were attributed to careless annotations.

6.2 Vague Text Human Performance

We look into how humans perform on the vague text disambiguation dataset. To do so, we randomly sampled 97 questions from the dataset and asked in-house annotators to answer them as multiple-choice questions. Each vague text, context pair was given 4 choices out of which only one was correct. We provide a brief event description along with all other available metadata to the annotator. Each question was answered by 3 annotators. Among the 97 questions that were answered, we observe the accuracy to be 94.85%, which shows this is

Democrat Only Entities		Common Entities					Republican Only Entities	
Target	Sentiment	Agreed-Upon Entities		Divisive Entities			Target	Sentiment
		Target	Sentiment	Sentiment (D)	Target	Sentiment (R)		
Anita Hill Patty Murray Merrick Garland Jeff Flake	Positive Positive Positive Negative	US Supreme Court US Senate FBI Judiciary Committee	Neutral Neutral Neutral Neutral	Positive Positive Positive Negative Negative Negative	Christine Blasey Ford Deborah Ramirez Julie Swetnick Brett Kavanaugh Donald Trump Mitch McConnell	Negative Negative Negative Positive Positive Positive	Susan Collins Chuck Grassley Diane Feinstein Chuck Schumer Sean Hannity	Positive Positive Negative Negative Neutral

Table 4: Target Entity-Sentiment centric view of *Kavanaugh Supreme Court Nomination* discourse

solvable for humans who understand the context.

6.3 Target Entity Visualization

In table 4, we aim to study an event using entity target sentiment annotations. We study the event *Kavanaugh Supreme Court Nomination* using the annotated data from Target Entity and Sentiment task. We identify entities that are discussed by both parties. We further separate them into divisive and agreed-upon entities based on expressed sentiments. We also show partisan discussed entities for the event. This analysis paints a very accurate picture of the discussed event. We observe that the main entities of Trump, Dr. Ford, Kavanaugh, Senate majority leader McConnell, and other accusers/survivors emerge as divisive entities. Entities such as Susan Collins and Anita Hill who were vocal mouthpieces of the respective party stances but didn’t directly participate in the event emerge as partisan entities. Supreme Court, FBI, and other entities occur in the discourse but only as neutral entities. Thus, this analysis shows the usefulness of the *Target Entity and Sentiment* identification in accurately summarizing events. Successful models on these tasks could greatly help in understanding political events better.

6.4 DCF Context Understanding

We look into the cases from the data that are incorrectly predicted when using Wikipedia pages but correctly predicted when using the DCF model. We report some of the examples in table 5 in the appendix. In examples 1 and 2 of target entity task, we can see that when the entity is not explicitly mentioned in the tweet, the Wiki-Context model fails to identify them as the target entities. We posit that while the Wikipedia page of each relevant event will contain these names, explicit modeling of entities in DCF model allows these examples to be correctly classified.

In examples 1 – 3 of vague text disambiguation task, we can see that when there are no clear terms indicating the sentiment towards a view, the Wiki-

Context model fails to disambiguate the tweet text. We posit that while the Wikipedia page of each relevant event will contain enough information, it will not tell the model what each party will say about the issue, explicit modeling of the authors in the DCF model allows these tweet texts to be correctly disambiguated.

7 Conclusion and Future Work

In this paper, we motivate, define, and operationalize “*Social Context Grounding*” for political discourse. We build two datasets that are useful to evaluate social context grounding in NLP models. We experiment with many types of context inclusion in NLP models and benchmark existing SOTA models. We show that explicit modeling of social context even in a small model outperforms simple encoding of context and GPT-3 as well on social context grounding.

Future work includes building datasets for other components of Social Context Grounding and building models that account for context in different forms to the DCF model as social relationships and symbolic contextual information are not included in the DCF model.

Limitations

Our work only addresses English language text in US political domain. We also build upon large language models and large PLMs which are trained upon huge amounts of uncured data. Although we employed human validation at each stage, biases could creep into the datasets. We also don’t account for the completeness of our datasets as it is a pioneering work on a new problem. Social context is vast and could have myriad of components. We only take a step in the direction of social context grounding in this work. The performance on these datasets might not indicate full social context understanding but they should help in sparking research in the direction of models that explicitly model such context. Although we tuned our prompts a

lot, it is possible that better prompts and evolving models might produce better results on the LLM baselines. Our qualitative analysis is predicated on a handful of examples. They are attempts to interpret the results of large neural models and hence don't carry as much confidence as our empirical observations.

Ethics Statement

In this work, our data collecting process consist of using both AMT and GPT-3. For the Target Entity and Sentiment Task, we are paying the worker \$1 for each HIT and expect an average work time of 3 minutes. This can be convert into hourly rate of \$20 which is well above the minimum wage set by the federal government. For the Vague Text Disambiguation Task, we are paying the worker \$1.10 for each HIT and expect an average work time of 3 minutes. This can be convert into hourly rate of \$22.

We recognize collecting political views from AMT and GPT-3 may come with bias or explicit results and have implemented gatekeepers to filter out unqualified workers and remove explicit results from the dataset.

References


- Stephen Ansolabehere, James M. Snyder, and Charles Stewart. 2001. [The effects of party and preferences on congressional roll-call voting](#). *Legislative Studies Quarterly*, 26(4):533–572.
- Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991.
- Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. [Predicting factuality of reporting and bias of news media sources](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, Brussels, Belgium. Association for Computational Linguistics.
- Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. [GPT-NeoX-20B: An open-source autoregressive language model](#). In *Proceedings of the ACL Workshop on Challenges & Perspectives in Creating Large Language Models*.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. [Language models are few-shot learners](#).
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. [Language models are few-shot learners](#). *CoRR*, abs/2005.14165.
- 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.
- Shangbin Feng, Zhaoxuan Tan, Zilong Chen, Ningnan Wang, Peisheng Yu, Qinghua Zheng, Xiaojun Chang, and Minnan Luo. 2022. Par: Political actor representation learning with social context and expert knowledge. *arXiv preprint arXiv:2210.08362*.
- Dirk Hovy and Diyi Yang. 2021. [The importance of modeling social factors of language: Theory and practice](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.
- Hamid Karimi, Tyler Derr, Aaron Brookhouse, and Jiliang Tang. 2020. [Multi-factor congressional vote prediction](#). In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM '19*, page 266–273, New York, NY, USA. Association for Computing Machinery.
- Chang Li and Dan Goldwasser. 2019. [Encoding social information with graph convolutional networks for Political perspective detection in news media](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2594–2604, Florence, Italy. Association for Computational Linguistics.

- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.
- Nikhil Mehta and Dan Goldwasser. 2021. [Tackling fake news detection by interactively learning representations using graph neural networks](#). In *Proceedings of the First Workshop on Interactive Learning for Natural Language Processing*, pages 46–53, Online. Association for Computational Linguistics.
- Nikhil Mehta, María Leonor Pacheco, and Dan Goldwasser. 2022. Tackling fake news detection by continually improving social context representations using graph neural networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1363–1380.
- Rajkumar Pujari and Dan Goldwasser. 2021. [Understanding politics via contextualized discourse processing](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1353–1367, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Rajkumar Pujari, Erik Oveson, Priyanka Kulkarni, and Elnaz Nouri. 2022. [Reinforcement guided multi-task learning framework for low-resource stereotype detection](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6703–6712, Dublin, Ireland. Association for Computational Linguistics.
- Jeroen Vaes, Maria Paola Paladino, and Chiara Magagnotti. 2011. [The human message in politics: The impact of emotional slogans on subtle conformity](#). *The Journal of Social Psychology*, 151(2):162–179. PMID: 21476460.
- Derek Weber and Frank Neumann. 2021. Amplifying influence through coordinated behaviour in social networks. *Social Network Analysis and Mining*, 11.
- 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.
- Yi Yang, Ming-Wei Chang, and Jacob Eisenstein. 2016. [Toward socially-infused information extraction: Embedding authors, mentions, and entities](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1452–1461, Austin, Texas. Association for Computational Linguistics.
- Haolan Zhan, Zhuang Li, Yufei Wang, Linhao Luo, Tao Feng, Xiaoxi Kang, Yuncheng Hua, Lizhen Qu, Lay-Ki Soon, Suraj Sharma, Ingrid Zukerman, Zhaleh Semnani-Azad, and Gholamreza Haffari. 2023. [Socialdial: A benchmark for socially-aware dialogue systems](#).

A Appendix


Target Entity and Sentiment Task	Vague Text Disambiguation Task
<p>Tweet: Republicans held Justice Scalia's seat open for more than 400 days. Justice Kennedy's seat has been vacant for less than two months. It's more important to investigate a serious allegation of sexual assault than to rush Kavanaugh onto the Supreme Court for a lifetime appointment. Author: Adam Schiff (Democrat) Event: Brett Kavanaugh Supreme Court nomination</p> <p>Entity: Christine Blasey Ford</p> <p>Wiki-Context Prediction: Not Target DCF Prediction: Target (correct)</p>	<p>Tweet: Thanks for this.</p> <p>Affiliation: Democrat Event: United States withdrawal from the Paris Agreement Paraphrase: There's nothing surprising in withdrawing from the Paris agreement. Thanks for not caring our environment and future generations. Wiki-Context Prediction: No DCF Prediction: Yes (correct)</p>
<p>Tweet: We will not be intimidated. Democracy will not be intimidated. We must hold the individuals responsible for the Jan. 6th attack on the U.S. Capitol responsible. Thank you @RepAOC for tonight's Special Order Hour and we will continue our efforts to #HoldThemAllAccountable. Author: Adriano Espaillat (Democrat) Event: January 6 United States Capitol attack</p> <p>Entity: Donald Trump</p> <p>Wiki-Context Predicted: Not Target DCF Prediction: Target (correct)</p>	<p>Tweet: Let us say enough. Enough.</p> <p>Affiliation: Democrat Event: Second impeachment of Donald Trump ended with not guilty Paraphrase: The failure of the Democrats to impeach Donald Trump is a strong moment for our legislature which can get back to its work helping the American people. Today we've been able to tell the American people what we have known all along, that Donald Trump was not guilty of these charges. Wiki-Context Predicted: Yes DCF Prediction: No (correct)</p>
<p>Tweet: #GeorgeFloyd #BlackLivesMatter #justiceinpolicing QT @OmarJimenez Former Minneapolis police officer Derek Chauvin is in the process of being released from the Hennepin County correctional facility his attorney tells us. He is one of the four officers charged in the death of George Floyd. He faces murder and manslaughter charges. Author: Adriano Espaillat (Democrat) Event: George Floyd protests Entity: Derek Chauvin</p> <p>Wiki-Context Predicted Sentiment: Positive DCF Prediction: Negative (correct)</p>	<p>Tweet: Lots of honking and screaming from balconies. Something must be going on.</p> <p>Affiliation: Democrat Event: Presidential election of 2020 Paraphrase: I'm sure that the people are celebrating the election results. Wiki-Context Prediction: No DCF Prediction: Yes (correct)</p>

Table 5: Examples where baseline model fails but DCF works



Sheila Jackson...
@Jackson... · Follow

In case you're wondering, that's Deborah Ramirez on the left (the woman who claimed Brett Kavanaugh sexually assaulted her) and that's @realDonaldTrump's choice to sit on the #SCOTUS on the right. Who says further investigation won't reveal more information?? #StopKavanaugh



6:02 PM · Oct 1, 2018

459 · Reply · C...

InstructionsShortcuts

Potential Targets:

NEGATIVE × NON-TARGET × POSITIVE ×

['Brett Kavanaugh', 'Christine Blasey Ford', 'Deborah Ramirez', 'Julie Swetnick', 'Dianne Feinstein', 'Rachel Mitchell', 'Supreme Court of the United States', 'Democratic Party (United States)', 'Republican Party (United States)', 'United States Department of Justice', 'Federal Bureau of Investigation', '@realDonaldTrump', 'None of the above (Carefully considered all above options)']

Tweet Text:

In case you're wondering that's Deborah Ramirez on the left (the woman who claimed Brett Kavanaugh sexually assaulted her) and that's @realDonaldTrump's choice to sit on the

Labels

Positive Sentiment 1

Negative Sentiment 2

Neutral Sentiment 3

Non-Target 4

☐ No entities to label

Figure 2: An example of *Tweet Target Entity and Sentiment Annotation* GUI

Task

Example 1

Example 2

Instructions

Background

General Context

Date Published (MM/DD/YYYY): 11/04/2020
Author: **Republican** - Anonymous
Event Happened: United States withdrawal from the Paris Agreement

Short Event Description

After a three-year delay, the US has become the first nation in the world to formally withdraw from the Paris climate agreement.

Tweet

Just the beginning...

⚠

Please read the example, instruction and background information carefully before proceed.
Even if you have seen the tweet before, the context has changed. Not reading it carefully may result in rejection.

×

Sentiment Analysis

Entities may not be present in the tweet text, please try your best to inference from the context.

Sentiment towards Donald Trump

☐ Positive
☐ Neutral
☐ Negative

Sentiment towards Republicans

☐ Positive
☐ Neutral
☐ Negative

Sentiment towards Democrats

☐ Positive
☐ Neutral
☐ Negative

Paraphrase

Provide an paraphrase for the given tweet and context

Refer to the examples given and try to think about what the author want to say explicitly.

Write the paraphrase here as if you are the original author...

Submit

Figure 3: An example of *Vague Text Disambiguation* GUI