# *"We Demand Justice!"*: Towards Social Context Grounding of Political Texts

**Anonymous ACL submission**

## Abstract

Political discourse on social media often contains similar language used with opposing intended meanings. For example, the phrase *thoughts and prayers*, is used to express sympathy for mass shooting victims, as well as satirically criticize the lack of legislative action on gun control. Fully understanding such discourse just by reading the text is difficult. However, knowledge of the social context information makes it easier.

We characterize the social context required to fully understand such ambiguous discourse, by grounding the text in real-world entities, actions, and attitudes. We propose two datasets that require an understanding of social context and benchmark them using large pre-trained language models, and several novel structured models. We show that structured models, explicitly modeling social context, outperform larger models on both tasks, but still lag significantly behind human performance. Finally, we perform an extensive analysis, to obtain further insights into the language understanding challenges posed by our social grounding tasks.

## 1 Introduction

Over the past decade, micro-blogging websites have become the primary medium for US politicians to interact with general citizens and influence their stances for gaining support. As a result, politicians from the same party often coordinate the phrasing of their social messaging, to amplify their impact (Vaes et al., 2011; Weber and Neumann, 2021). Hence, repetitive, succinct phrases, such as *"Thoughts and Prayers"*, are extensively used, although they signal more nuanced stances. Moreover, the interaction among politicians from opposing parties often leads to messaging phrased similarly, but signaling opposing real-world actions. For example, *'Thoughts and Prayers'*, when used by Republicans, expresses condolences in mass shooting events, but when used by Democrats conveys an *angry* or *sarcastic* tone as a call for action demanding *"tighter gun control measures"*. Similarly, fig. 1 shows contrasting interpretations of the phrase "*We need to keep our teachers safe*!" depending on different speakers and in the context of different events.


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

Humans familiar with the stances of a politician and, possessing knowledge about the event from the news, can easily understand the *intended meaning* of political phrases. However, automatically understanding such language is challenging. Our main question in this paper is - ***Can an NLP model find the right meaning?*** From a linguistic perspective, we follow the distinction (Bach, 2008) between *semantic* interpretation (i.e., meaning encoded directly in the utterance and does not change based on its external context), and *pragmatic* interpretation (that depends on extra-linguistic information). The latter has gathered significant interest in the NLP community recently (Bender and Koller, 2020; Bisk et al., 2020), focusing on language understanding, when grounded in an external context (Fried et al., 2023). To a large extent,

| Tweet Target Entity and Sentiment | | | | Vague Text Disambiguation |
|---|---|---|---|---|
| **Tweet:** As if we needed more evidence. #kavanaugh<br>**Event:** Kavanaugh Supreme Court Nomination<br>**Author:** Earl Blumenauer (Democrat Politician)<br>**Targets:** Brett Kavanaugh (negative), Julie Swetnick (positive) Christine Ford (positive), Deborah Ramirez (positive) | | | | **Vague Text:** First, but not the last.<br>**Event:** US withdraws from Paris climate agreement that enforces environmental targets after three years<br>**Author Party:** Republican<br>**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.<br>**Incorrect Disambiguations:** |
| **Target Task Data Statistics** | | **Vague Text Data Statistics** | | 1) Joe Biden's inauguration marks the first day of a new era of progress and prosperity, lasting positive changes are coming. (Incorrect Event) |
| Unique Tweets | 865 | Unique Vague Texts | 93 | 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) |
| Positive Targets | 1513 | Positive Examples | 739 | 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) |
| Negative Targets | 1085 | Negative Examples | 2217 | |
| Neutral Targets | 784 | Total Examples | 2956 | |
| Non-Targets | 2509 | Number of Events | 9 | |
| Total Data Examples | 5891 | Hard Test Examples | 180 | |
| Number of Events | 3 | | | |

Table 1: Examples of Annotated Datasets and their statistics

the focus of such studies has been on grounding language in a perceptual environment (e.g., image captioning (Andreas and Klein, 2016; Sharma et al., 2018; Alikhani et al., 2020), instruction following (Wang et al., 2016; Suhr et al., 2019; Lachmy et al., 2022), and game playing (Potts, 2012; Udagawa and Aizawa, 2019) tasks). Unlike these works, in this paper, we focus on *grounding language in a social context*, i.e., modeling the common ground (Clark and Brennan, 1991; Traum, 1994; Stalnaker, 2002) between the author and their social media followers, that enables understanding an otherwise highly ambiguous utterances. The *Social Context Understanding*, needed for building successful models for such tasks, can come from a wide variety of sources. The politician's affiliation and historical stances on the issue provide can capture crucial social context. Social relationships, knowledge about the involved entities, and related prior and upcoming events form important part of the puzzle as well. In fig. 1 event #1, combining the event information (*school shooting*) with the speakers' gun control stances, would facilitate understanding the *intended meaning* of the text.

**The main motivation of this paper work is to operationalize the *'Social Context Grounding'* problem as a pragmatic understanding task**. From a practical perspective, this would enable the creation of better NLP-CSS models that can process social media text in settings that require contextualized understanding. We suggest several datasets, designed to evaluate this ability in computational models. These task capture the intended meaning at different level of granularity. At the most basic level, providing the social context can help identify the entities targeted, and the sentiment towards them. In fig. 1, the social context ⟨event#1, Harris⟩and the text *"we need to keep our teachers safe"* ⇒ *"negative attitude towards guns"*. A more nuanced account of meaning, which we formulate as a separate task, captures the specific means in which the negative attitude is expressed (the Interpretation in fig. 1). We additionally present two datasets corresponding to these tasks, namely, *'Target Entity and Sentiment Detection'* and *'Vague Text Disambiguation'*. In the first, the goal is to predict: 1) whether a given entity is the *intended target* of a politician's tweet and 2) the sentiment towards the intended targets. We explicitly focus on tweets that *do not always mention the targets* in their text to incentivize modeling the pragmatic communicative intent of the text. In the second task, given an ambiguous political message such as *"We demand justice"* and its social context (*associated event*, & the *author's party affiliation*), the task is to identify a *plausible* unambiguous explanation of the message. Note that the ground truth for all these tasks is based on human pragmatic interpretation, i.e., "*guns*" is a negative target of "*we need to keep our teachers safe*", despite not being mentioned in the text, since it was perceived in this way by a team of human annotators reading the tweet and knowing social context. We show examples of each task in table 1. We describe the datasets in detail in section 3.

We evaluate the performance of various models, as a way to test the need for social context and compare different approaches for modeling it. These include pre-trained LM-based classifiers, and LLM in-context learning (Brown et al., 2020a; Black et al., 2022), which use a textual representation of the social context. We also adopt an existing graph-based discourse contextualization framework (Pujari and Goldwasser, 2021; Feng et al., 2022), to

explicitly model the social context needed to solve the proposed tasks. Our results demonstrate that the discourse contextualization models outperform other models on both tasks. We present an error analysis to gain further insights. We describe the models in section 4 and the results in section 5.

We also present a qualitative visualization of a political event, *Brett Kavanaugh Supreme Court Nomination* (section 6.4), from target entity-sentiment perspective. It showcases a unique summary of the event discourse. We perform human evaluation on our *'Vague Text Disambiguation'* dataset, and observe that humans find this task much easier than the evaluated models. We also present observations of human vs. LLM errors in disambiguation. In summary, our contributions are:

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. We adopt existing discourse contextualization framework for the proposed tasks, and evaluate GPT-3's in-context learning performance, as well.
3. Performing human studies to benchmark the dataset difficulty and GPT-3 generation performance, when compared to human workers.[1]

## 2 Related Work

**Pragmatic Language Grounding** gained significant focus recently (Bender and Koller, 2020; Bisk et al., 2020) following the rise of Pretrained Language Models (Devlin et al., 2019; Liu et al., 2019; Brown et al., 2020a) as unified NLP models. Most grounding tasks address multi-modal or physical environment descriptions (Barnard et al., 2003; Vogel and Jurafsky, 2010; Chen and Mooney, 2011; Tellex et al., 2011; Mitchell et al., 2012; Anderson et al., 2018). We refer the reader to (Fried et al., 2023) for a thorough overview. In contrast, we focus on grounding language in a *social* context.

**Social Context Modeling** Hovy and Yang (2021) show that modeling social context is necessary for human-level NLU. As political messages are often targeted at the voter base aware of the political context (Weber and Neumann, 2021; Vaes et al., 2011), they are vague by design. Several previous works model social context 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). These works model partial aspects of social context, relevant to their tasks. Two recent frameworks aim to capture social context holistically (Pujari and Goldwasser, 2021; Feng et al., 2022). Evaluation tasks presented in both works show interesting social context understanding but are not fully representative of the challenges of *Social Context Grounding*. Zhan et al. (2023) propose a dataset for dialogue understanding addressing general social commonsense.

**Related Semantic and Pragmatic tasks** closest to our *Target Entity Sentiment Identification* task is *Stance Detection* in social media (Mohammad et al., 2016; AlDayel and Magdy, 2020). To clarify our contribution, Mohammad et al. (2016), a popular SemEval task, looks at sentiment towards 5 targets, while our data has 362 unique targets. Allaway and McKeown (2020) and Zhang et al. (2022) also propose stance datasets on tweets. But, they focus mainly on semantic understanding of text that allows them to predict agreement or disagreement with well-defined statements. Our *Vague Text Disambiguation* task is related to recent works that study implicit inferences (Hoyle et al., 2023), and pragmatic understanding (Hu et al., 2023). However, our tasks evaluate pragmatic understanding using an explicit context, absent in those tasks.

## 3 Social Context Grounding Tasks

We design and collect two datasets for *Social Context Grounding* evaluation, and define three pragmatic interpretation tasks. In the *Tweet Target Entity and Sentiment* dataset, we collect annotations of opinionated tweets from known politicians for their intended targets and sentiments towards them. We focus on three political events for this task. The dataset and its collection are described below in section 3.1. In the *Vague Text Disambiguation Task*, we collect plausible explanations of vague texts, given the social context, consisting of *author affiliation* and *specific event*. We focus on eight political events. This dataset is detailed in section 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 a target of $T$ and the sentiment towards $E$. Political discourse often contains opinionated discourse about world events and social issues. We collect tweets that don't directly mention the target entities.

[1] Our data and code will be released under MIT license

Thus, connecting the text with the event details and the author's general perspectives is necessary to solve this task effectively. We pick the focal entities for the given event and let human annotators expand on that initial set, based on their interpretation of the contextualized text. A *target* entity is conceptualized as an entity present in the full intended interpretation of the tweet.

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 identify relevant participating entities for each of the three events. Examples of the involved entities for the event *George Floyd Protests* were *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 Target-Sentiment Data Collection

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 corpus[2]. 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 incentivize testing 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, $511$ development, and $1,009$ test examples. We compute the mean Cohen's kappa score for annotations and report inter-annotator agreement for annotated targets (0.47) and sentiment (0.73)

[2] https://github.com/alexlitel/congresstweets

### 3.2 Vague Text Disambiguation Task

The task of *Vague Text Disambiguation* is designed to capture pragmatic interpretation at a finer-grained level. It can be viewed as a variant of the well known paraphrase task, adapted for the social context settings. The model is evaluated on its ability to identify plausible interpretations (i.e., a sentence explicitly describing the author's intent) of an ambiguous quote given the event context and author's affiliation. E.g., "*protect our children from mass shootings*" could easily be disambiguated as either "*ban guns*" or "*arm teachers*" when the author's stance on the issue of '*gun rights*' is known.

Our data collection effort is designed to capture different aspects of social context grounding and facilitate detailed error analysis. Defined as a binary classification task over tuples ⟨`Party, Event, Vague text, Explicit text`⟩, we create negative examples by flipping tuple elements values of positive examples. This allows us to evaluate whether models can capture event relevance, political stance, or constrain the interpretation based on the vague text. For example, in the context of Event #1 in fig. 1, we can test if models simply capture the correlation between Democrats and negative stance towards guns access by replacing the vague text to *"let your voice be heard"*, which would make the interpretation in fig. 1 implausible despite being consistent with that stance, while other consistent interpretations would be plausible (e.g., *"go outside and join the march for our lives"*).

#### 3.2.1 Vague Text Data Collection

Data collection was done in several steps. (1)**Vague Texts Collection**. We collected vague text candidates from tweets by US politicians (i.e. senators and representatives) between the years 2019 to 2021 from Congress Tweets corpus. We identified a list of 9 well-known events from that period and identified event-related tweets using their time frame and relevant hashtags. We used a pre-trained BERT-based (Devlin et al., 2019) NER model to

collect tweets that do not contain any entity mentions to identify potential candidates for vague texts. We manually identified examples that could have contrasting senses by flipping their social context. We obtain 93 vague text candidates via this process.
(2)**In-Context Plausible Meaning Annotation**. We match the 93 ambiguous tweets with different events that fit them. We use both Democrat and Republican as the author party affiliation. We obtain 600 *context-tweet* pairs for AMT annotation. For each tweet, we ask AMT workers to annotate the following two aspects: 1) sentiment towards the three most relevant entities in the event (sanity check) and 2) a detailed explanation of the *intended meaning* given the event and author's party affiliation. We obtain 469 reasonable annotations. After this step, each annotation was screened by in-house annotators. We ask three in-house annotators to vote on the *correctness*, *appropriateness*, and *plausibility* of the annotation given the context. Thus, we create a total of 374 examples.
(3)**LLM-based Data Expansion**. Using these examples, we further generate candidates for the task using LLM few-shot prompting. We use the examples from the previous step as in-context few-shot examples in the prompt. We use GPT-NeoX (Black et al., 2022) and GPT-3 (Brown et al., 2020a) for candidate generation. For each generated answer, manual inspection by three in-house annotators is performed to ensure data quality. We generate 928 candidates using GPT-NeoX and GPT-3. Human expert filtering results in 650 generations that pass the quality check. After removing redundant samples, we obtain 365 examples. Thus, we obtain a total of 739 annotations for this task. Then, for each of the 739 examples, we ask in-house annotators to select 3 relevant negative options from the pool of explanations. We instruct them to pick hard examples that might contain similar entities as the correct interpretation. This results in $2,956$ binary classification data samples. We analyze and discuss the results of human validation of large LM generations in section 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 $1,916$ train, 460 development, and 580 test examples. 180 of the test examples are from events/party contexts unseen in train data.

## 4 Modeling Social Context

The key technical question this paper puts forward is how to model the social context, such that the above tasks can be solved with high accuracy. We observe that humans can perform this task well (section 6.3), and evaluate different context modeling approaches in terms of their ability to replicate human judgments. These correspond to `No Context`, `Text-based` context representation (e.g., Twitter Bio, relevant Wikipedia articles), and `Graph-based` context representation, simulating the social media information that human users are exposed to when reading the vague texts.

We report the results of all our baseline experiments in table 2 and table 3. The first set of results evaluate fine-tuned pre-trained language models (PLM), namely BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), with three stages of modeling context. Firstly, we evaluate no contextual information setting. Second, we include the authors' Twitter bios as context. Finally, we evaluate the information from the author, event, and target entity Wikipedia pages as context (models denoted `PLM Baselines {No, Twitter Bio, Wikipedia} Context`, respectively).

We evaluate GPT-3[3] in *zero-shot* and *four-shot* in-context learning paradigm on both tasks. We provide contextual information in the prompt as short event descriptions and authors' affiliation descriptions. Note that GPT-3 is trained on news data until Sep. 2021 which includes the events in our data (models denoted `LLM Baseline`).

We evaluate the performance of politician embeddings from Political Actor Representation (PAR) (Feng et al., 2022) and Discourse Contextualization Framework (DCF) (Pujari and Goldwasser, 2021) models. (models denoted `Static Contextutalized Embeddings`). We use PAR embeddings available on their GitHub repository[4]. For DCF model, we use released pre-trained models from GitHub repository[5] to generate author, event, text, and target entity embeddings. We evaluate the embeddings on both tasks. We briefly review these models in section 4.1 & section 4.2.

Finally, we use tweets of politicians from related previous events and build context graphs for each

[3] gpt-3.5-turbo-1106 via OpenAI API
[4] https://github.com/BunsenFeng/PAR
[5] https://github.com/pujari-rajkumar/compositional_learner

| Model | | Target Identification | | | | Sentiment Identification | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Prec | Rec | Macro-F1 | Acc | Prec | Rec | Macro-F1 | Acc |
| No Context Baselines | 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 |
| PLMs +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 |
| PLMs +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 | 59.10 | 58.07 | 58.28 |
| LLMs | GPT-3 0-shot | 69.25 | 70.58 | 69.77 | 73.78 | 56.20 | 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 Contextutalized Embedding Models | RoBERTa-base + PAR Embs | 68.38 | 71.63 | 67.67 | 69.18 | 55.01 | 56.89 | 55.51 | 55.40 |
| | BERT-large + PAR Embs | 65.40 | 67.33 | 60.25 | 60.56 | 55.24 | 57.54 | 55.89 | 55.80 |
| | RoBERTa-base + DCF Embs | **72.89** | **75.95** | **73.56** | **75.82** | 63.05 | 63.52 | 62.90 | 63.03 |
| | BERT-large + DCF Embs | 68.76 | 72.02 | 68.32 | 69.97 | 61.59 | 63.25 | 61.22 | 60.75 |
| Discourse Contextualized Models | 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* (binary task) and *Sentiment* (4-classes) test sets. We report macro-averaged Precision, macro-averaged Recall, macro-averaged F1, and Accuracy metrics.

| Model | Vague Text Disambiguation | | | |
|---|---|---|---|---|
| | Prec | Rec | Macro-F1 | Acc |
| **No Context Baselines** | | | | |
| BERT-large | 52.24 | 55.58 | 50.28 | 53.75 |
| RoBERTa-base | 55.3 | 51.82 | 54.53 | 56.08 |
| **PLMs + Wikipedia Context** | | | | |
| BERT-large + wiki | 52.31 | 46.90 | 66.87 | 76.03 |
| BERT-base + wiki | 51.85 | 38.62 | 64.36 | 75.69 |
| **LLMs** | | | | |
| 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 Contextutalized Embedding Models** | | | | |
| 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 Models** | | | | |
| 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, a binary classification task. We report macro-averaged Precision, macro-averaged Recall, macro-averaged F1, and Acc. metrics

data example as proposed in Pujari and Goldwasser (2021). We use Wikipedia pages of authors, events, and target entities to add social context information to the graph. Then, we train the Discourse Contextualization Framework (DCF) for each task and evaluate its performance on both tasks (models denoted `Discourse Contextualization Model`). Further details of our baseline experiments are presented in subsection section 4.3. Results of our baseline experiments are discussed in section 5.

### 4.1 Discourse Contextualization Framework

Discourse Contextualization Framework (DCF) (Pujari and Goldwasser, 2021) leverages relations among social context components to learn contextualized representations for text, politicians, events, and issues. It consists of *encoder* and *composer* modules that compute holistic representations of the context graph. The encoder creates an initial representation of nodes. Composer propagates the information within the graph to update node representations. They define link prediction learning tasks over context graphs to train the model. They show that their representations significantly outperform several PLM-based baselines trained using the same learning tasks.

### 4.2 Political Actor Representation

Feng et al. (2022) propose the *Political Actor Representation* (PAR) framework, a graph-based approach to learn more effective politician embeddings. They propose three learning tasks, namely, 1) Expert Knowledge Alignment 2) Stance Consistency training & 3) Echo chamber simulation, to infuse social context into the politician representations. They show that PAR representations outperform SOTA models on *Roll Call Vote Prediction* and *Political Perspective Detection*.

### 4.3 Experimental Setup

*Target Entity Detection* is binary classification with ⟨*author, event, tweet, target-entity*⟩as input and *target/non-target* label as output. *Sentiment Detection* is set up as 4-way classification. Input is the same as the target task and output is one of: {*positive, neutral, negative, non-target*}. *Vague Text Disambiguation* is a binary classification task with ⟨*party-affiliation, event, vague-text, explanation-text*⟩and a *match/no-match* label as output.

In phase 1 no-context baselines, we use the author, event, tweet, and target embeddings generated by PLMs. We concatenate them for input. In Twitter-bio models, we use the author's Twitter bio embeddings to represent them. Wiki context models receive Wikipedia page embeddings of author,

event, and target embeddings. *It is interesting to note that the Wikipedia context models get all the information needed to solve the tasks.*. In phase 2 LLM experiments, we use train samples as in-context demonstrations. We provide task and event descriptions in the prompt. In phase 3 PAR models, we use politician embeddings released on the PAR GitHub repository to represent authors. We replace missing authors with their wiki embeddings. For the *Vague Text* task, we average PAR embeddings for all politicians of the party to obtain party embeddings. For DCF embedding models, we generate representations for all the inputs using context graphs. We also use authors' tweets from relevant past events. We build graphs using author, event, tweet, relevant tweets, and target entity as nodes and edges as defined in the original DCF paper. In phase 4, we use the same setup as the DCF embedding model and additionally back-propagate to DCF parameters. This allows us to fine-tune the DCF context graph representation for our tasks.

## 5 Results

The results of our baseline experiments are described in Tab. 2 and 3. We evaluate our models using macro-averaged precision, recall, F1, and accuracy metrics (due to class imbalance, we focus on macro-F1). Several patterns, consistent across all tasks, emerge. **First**, *modeling social context is still an open problem*. None of our models were able to perform close to human level. **Second**, *adding context can help performance*, compared to the No-Context baselines, models incorporating context performed better, with very few exceptions. **Third**, *LLMs are not the panacea for social-context pragmatic tasks*. Despite having access to a textual context representation as part of the prompt, and having access to relevant event-related documents during their training phase, these models underperform compared to much simpler models that were fine-tuned for this task. **Finally**, *explicit context modeling using the DCF model consistently leads to the best performance*. The DCF model mainly represents the social 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. In the *Target Entity* task, RoBERTa-base + DCF embeddings obtain 73.56 F1 vs. 68.83 for the best no-context baseline. Twitter bio and wiki-context hardly improve, demonstrating the effectiveness of modeling contextual information explicitly vs. concatenating context as text documents. No context performance well above the random performance of 50 F1 indicates the bias in the target entity distribution among classes. We discuss this in section 6.4. In *Sentiment Identification* task, we see that BERT-large + DCF back-propagation outperforms all other models. *Vague Text Disambiguation* task results in table 3 show that DCF models outperform other models significantly. 71.71 F1 is obtained by BERT-base + DCF embeddings. BERT-base performing better than bigger PLMs might be due to DCF model's learning tasks being trained using BERT-base embeddings.

## 6 Analysis and Discussion

### 6.1 Ablation Analysis on Vague Text Task

We report ablation studies in table 5 on the Vague Text task test set. We consider 5 splits: **(1)** Unseen Party: ⟨*party, event*⟩not in the train set but ⟨*opposing-party, event*⟩is present, **(2)** Unseen Event: ⟨*party* ⟩not in train set, **(3)** Flip Event: negative samples with corresponding 'event flipped-party/vague tweet matched' positive samples in train set and analogous **(4)** Flip Party and **(5)** Flip Tweet splits. We observe the best model in each category. They obtain weaker performance on unseen splits, as expected, unseen events being the hardest. Contextualized models achieve higher margins. DCF gains 7.6(13.2%) and DCF embeddings attain 8.12(20.42%) macro-F1 improvement over BERT-base+wiki compared to respective margins of 8.86% and 11.42% on the full test set. In the flip splits with only negative examples, accuracy gain over random baseline for all splits is seen. This indicates that models learn to jointly condition on context information rather than learn spurious correlations over particular aspects of the context. Specifically, flip-tweet split results indicate that models don't just learn party-explanation mapping.

### 6.2 Vague Text LLM Generation Quality

We look into the 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 better. Among the 430 generated instances, 315 were annotated as good which converts to an accep-

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

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

| Data Split | Unseen Party | Unseen Event | Flip Tweet | Flip Event | Flip Party |
|---|---|---|---|---|---|
| | Ma-F1 | Ma-F1 | Acc | Acc | Acc |
| Random | 44.70 | 29.69 | 75 | 75 | 75 |
| BERT-base+wiki | 57.58 | 39.76 | 88.14 | 89.77 | 87.77 |
| BERT-base +DCF Embs | 61.79 | 47.88 | 86.10 | 93.18 | 84.57 |
| BERT-base+DCF | 65.18 | 45.65 | 82.03 | 89.77 | 84.04 |

Table 5: Ablation Study Results on Vague Text Task

tance 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.

### 6.3 Vague Text Human Performance

We look into how humans perform on the *Vague Text Disambiguation* task. We randomly sample $97$ questions and ask 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 the metadata available to the annotator. Each question was answered by $3$ annotators. Among the $97$ answered questions, the accuracy was $94.85\%$, which shows this task is easy for humans who understand the context. Respective performance of best models on this subset of data for BERT-base+wiki ($54.89\%$), BERT-base+DCF-embs ($63.38\%$), BERT-base+DCF ($64.79\%$) is much lower than human performance.

### 6.4 Target Entity Visualization

The main goal of this analysis is to demonstrate the usefulness and inspire modeling research in the direction of entity-sentiment-centric view of political events. table 4 visualizes one component of how partisan discourse is structured in these events. We study *Kavanaugh Supreme Court Nomination*. We identify discussed entities and separate them into divisive and agreed-upon entities. This analysis paints an accurate picture of the discussed event. We observe that the main entities of Trump, Dr. Ford, Kavanaugh, Sen. 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 but only as neutral entities.

### 6.5 DCF Context Understanding

We look into examples that are incorrectly predicted using Wikipedia pages but correctly predicted by the DCF model in the appendix (table 6). In examples $1$ & $2$ of *Target Entity-Sentiment* task, when the entity is not explicitly mentioned in the tweet, the Wiki-Context model fails to identify them as the targets. We posit that while the Wikipedia page of each relevant event will contain these names, explicit modeling of entities in the DCF model allows correct classification. Examples $1-3$ of *Vague Text Disambiguation* task show that when no clear terms indicate the sentiment towards a view, the Wiki-Context model fails to disambiguate the tweet text. Explicit modeling of politician nodes seems to help the DCF model.

## 7 Conclusion and Future Work

In this paper, we motivate, define, and operationalize '*Social Context Grounding*' for political text. We build two novel datasets to evaluate social context grounding in NLP models that 'are easy for humans' when the relevant social context is provided. We experiment with many types of contextual models. We show that explicit modeling of social context outperforms other models while lacking behind humans.

## 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 uncurated 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 a 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, 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. We believe the insights from our findings will encourage more research in this area. For example, the development of discourse contextualized models that aim to model human-style understanding of background knowledge, emotional intelligence, and societal context understanding is a natural next step of our research.

## Ethics Statement

In this work, our data collection process consists of using both AMT and GPT-3. For the *Target Entity and Sentiment* task, we pay AMT workers $1 per HIT and expect an average work time of 3 minutes. This translates to an hourly rate of $20 which is above the federal minimum wage. For the *Vague Text Disambiguation* task, we pay AMT workers $1.10 per HIT and expect an average work time of 3 minutes. This translated to an hourly rate of $22.

We recognize collecting political views from AMT and GPT-3 may come with bias or explicit results and employ expert gatekeepers to filter out unqualified workers and remove explicit results from the dataset. Domain experts used for annotation are chosen to ensure that they are fully familiar with the events in focus. Domain experts were provided with the context related to the events via their Wikipedia pages, background on the general issue in focus, fully contextualized quotes, and authors' historical discourse obtained from ontheissues.org. We have an annotation quid-pro-quo system in our lab which allows us to have a network of in-house annotators. In-house domain experts are researchers in the CSS area with familiarity with a range of issues and stances in the US political scene. They are given the information necessary to understand the events in focus in the form of Wikipedia articles, quotes from the politicians in focus obtained from ontheissues.org, and news articles related to the event. We make the annotation process as unambiguous as possible. In our annotation exercise, we ask the annotators to mark only high-confidence annotations that can be clearly explained. We use a majority vote from 3 annotators to validate the annotations for the target entity task.

Our task is aimed at understanding and grounding polarized text in its intended meaning. We take examples where the intended meaning is clearly backed by several existing real-world quotes. We do not manufacture the meaning to the vague statements, we only write down unambiguous explanations where context clearly dictates the provided meaning. Applications of our research as we envision would be adding necessary context to short texts by being able to identify past discourse from the authors that are relevant to the particular text in its context. It would also be able to ground the text in news articles that expand upon the short texts to provide full context.

## References


Abeer AlDayel and Walid Magdy. 2020. Stance detection on social media: State of the art and trends. *CoRR*, abs/2006.03644.

Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. 2020. Cross-modal coherence modeling for caption generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6535, Online. Association for Computational Linguistics.

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on*

*computer vision and pattern recognition*, pages 3674–3683.

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Austin, Texas. Association for Computational Linguistics.

Kent Bach. 2008. *Pragmatics and the Philosophy of Language*, pages 463 – 487. Wiley Online Library.

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.

Kobus Barnard, Pinar Duygulu, David Forsyth, Nando De Freitas, David M Blei, and Michael I Jordan. 2003. Matching words and pictures. *The Journal of Machine Learning Research*, 3:1107–1135.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. 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.

David Chen and Raymond Mooney. 2011. Learning to interpret natural language navigation instructions from observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 859–865.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. *Perspectives on Socially Shared Cognition*, pages 127 – 149.

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*.

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2023. Pragmatics in language grounding: Phenomena, tasks, and modeling approaches. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12619–12640, Singapore. Association for Computational Linguistics.

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.

Alexander Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. 2023. Natural language decompositions of implicit content enable better text representations. In *Proceedings of the 2023 Conference*

*on Empirical Methods in Natural Language Processing*, pages 13188–13214, Singapore. Association for Computational Linguistics.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Royi Lachmy, Valentina Pyatkin, Avshalom Manevich, and Reut Tsarfaty. 2022. Draw me a flower: Processing and grounding abstraction in natural language. *Transactions of the Association for Computational Linguistics*, 10:1341–1356.

Chang Li and Dan Goldwasser. 2019. Encoding social information with graph convolutional networks forPolitical 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, 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.

Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa Mensch, Alexander Berg, Tamara Berg, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2016. Stance and sentiment in tweets. *CoRR*, abs/1605.01655.

Christopher Potts. 2012. Goal-driven answers in the cards dialogue corpus. In *Proceedings of the 30th west coast conference on formal linguistics*, pages 1–20. Cascadilla Proceedings Project Somerville, MA.

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.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5/6):701–721.

Alane Suhr, Claudia Yan, Jack Schluger, Stanley Yu, Hadi Khader, Marwa Mouallem, Iris Zhang, and Yoav Artzi. 2019. Executing instructions in situated collaborative interactions. 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 2119–2130, Hong Kong, China. Association for Computational Linguistics.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew Walter, Ashis Banerjee, Seth Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 1507–1514.

David Traum. 1994. A computational theory of grounding in natural language conversation.

Takuma Udagawa and Akiko Aizawa. 2019. A natural language corpus of common grounding under continuous and partially-observable context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7120–7127.

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.

Adam Vogel and Dan Jurafsky. 2010. Learning to follow navigational directions. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 806–814.

Sida I. Wang, Percy Liang, and Christopher D. Manning. 2016. Learning language games through interaction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2368–2378, Berlin, Germany. Association for Computational Linguistics.

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.

Xinliang Frederick Zhang, Nick Beauchamp, and Lu Wang. 2022. Generative entity-to-entity stance detection with knowledge graph augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9950–9969, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

## A Reproducibility

We use the HuggingFace Transformers (Wolf et al., 2020) library for PLMs. We use GPT-NeoX implementation by ElutherAI (Black et al., 2022) and GPT-3 (Brown et al., 2020b) via OpenAI API for our LLM baselines. 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 section 3 for our experiments. We use the development macro-F1 for early stopping. We run all our experiments using random seeds to ensure reproducibility. We experiment with a random seed value set to {13}. We report the results of the 3 fold cross-validation. We report only the mean on all the cross-fold validation results for clarity. All our code, datasets, and result logs will be released publicly upon acceptance. We experiment with 3, 5, and 10 fold cross-validation. As the results on the development data are almost identical, we report the results of 3 fold cross-validation in all our experiments.

## B Error Analysis

## C Annotation Interfaces

## D GPT Prompts

**GPT-3 prompts used for target-entity task:**

Event: <event>
Event background: <background-description>
Tweet: <tweet-text>
Author: <author-name>
Author Party: <party-affiliation>
Author background: <first two sentences of author-wiki-page>
Target Entity: <entity-name>
Entity background: <first two sentence of entity-wiki-page> Task: Identify if the given entity is a target of the tweet. A target entity is defined as an entity that would be present in the full unambiguous explanation of the tweet.
Is the given entity a target entity of the tweet? Answer yes or no.

**GPT-3 prompts used for target-sentiment task:**

Event: <event>
Event background: <background-description>
Tweet: <tweet-text>
Author: <author-name>
Author Party: <party-affiliation>
Author background: <first two sentences of author-wiki-page>
Target Entity: <entity-name>
Entity background: <first two sentence of entity-wiki-page> Task: Identify the sentiment of the tweet towards the given target entity. Consider that the tweet is ambiguous and the entity might be implied without being explicitly mentioned.
What is the sentiment of the tweet towards the target entity? Answer with positive, negative, or neutral.

**GPT-3 prompts used for vague text disambiguation task:**

Event: <event>
Event background: <background-description>
Vague message: <vague-text>
Author Party: <party-affiliation>
Author background: <first two sentences of party-wiki-page>
Task: Given the event, vague message, and party affiliation of the author, explain unambiguously the intended meaning of the vague message.
Generate an unambiguous explanation for the vague message given the party affiliation of the author and the event in context.

| Target Entity and Sentiment Task | Vague Text Disambiguation Task |
|---|---|
| **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.<br>**Author**: Adam Schiff (Democrat)<br>**Event**: Brett Kavanaugh Supreme Court nomination<br>**Entity**: Christine Blasey Ford<br>**Wiki-Context Prediction**: Not Target \| **DCF Prediction**: Target (correct) | **Tweet**: Thanks for this.<br>**Affiliation**: Democrat<br>**Event**: United States withdrawal from the Paris Agreement<br>**Paraphrase**: There's nothing surprising in withdrawing from the Paris agreement. Thanks for not caring our environment and future generations.<br>**Wiki-Context Prediction**: No \| **DCF Prediction**: Yes (correct) |
| **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.<br>**Author**: Adriano Espaillat (Democrat)<br>**Event**: January 6 United States Capitol attack<br>**Entity**: Donald Trump<br>**Wiki-Context Predicted**: Not Target \| **DCF Prediction**: Target (correct) | **Tweet**: Let us say enough. Enough.<br>**Affiliation**: Democrat<br>**Event**: Second impeachment of Donald Trump ended with not guilty<br>**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.<br>**Wiki-Context Predicted**: Yes \| **DCF Prediction**: No (correct) |
| **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.<br>**Author**: Adriano Espaillat (Democrat)<br>**Event**: George Floyd protests<br>**Entity**: Derek Chauvin<br>**Wiki-Context Predicted Sentiment**: Positive \| **DCF Prediction**: Negative (correct) | **Tweet**: Lots of honking and screaming from balconies. Something must be going on.<br>**Affiliation**: Democrat<br>**Event**: Presidential election of 2020<br>**Paraphrase**: I'm sure that the people are celebrating the election results.<br>**Wiki-Context Prediction**: No \| **DCF Prediction**: Yes (correct) |

Table 6: Examples where baseline model fails but DCF works

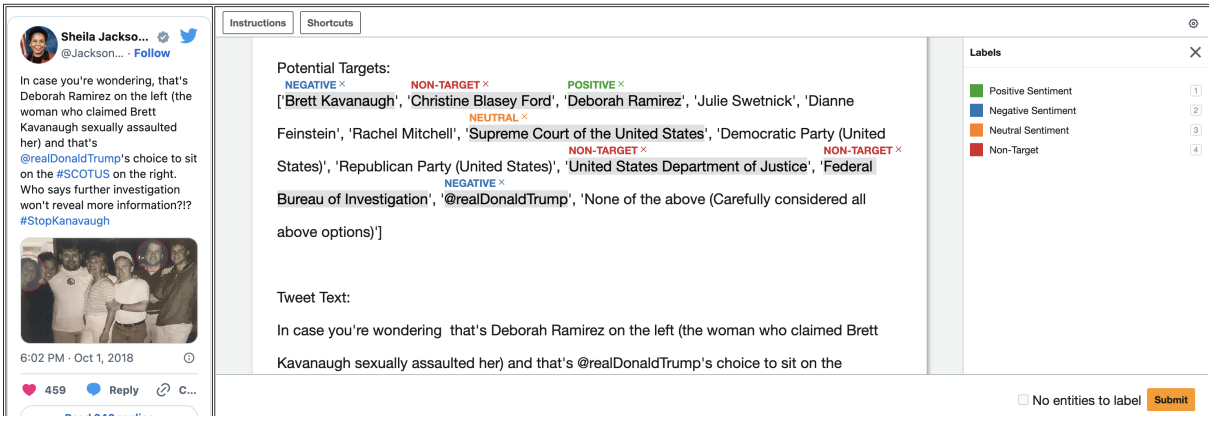


Figure 2: An example of *Tweet Target Entity and Sentiment Annotation* GUI

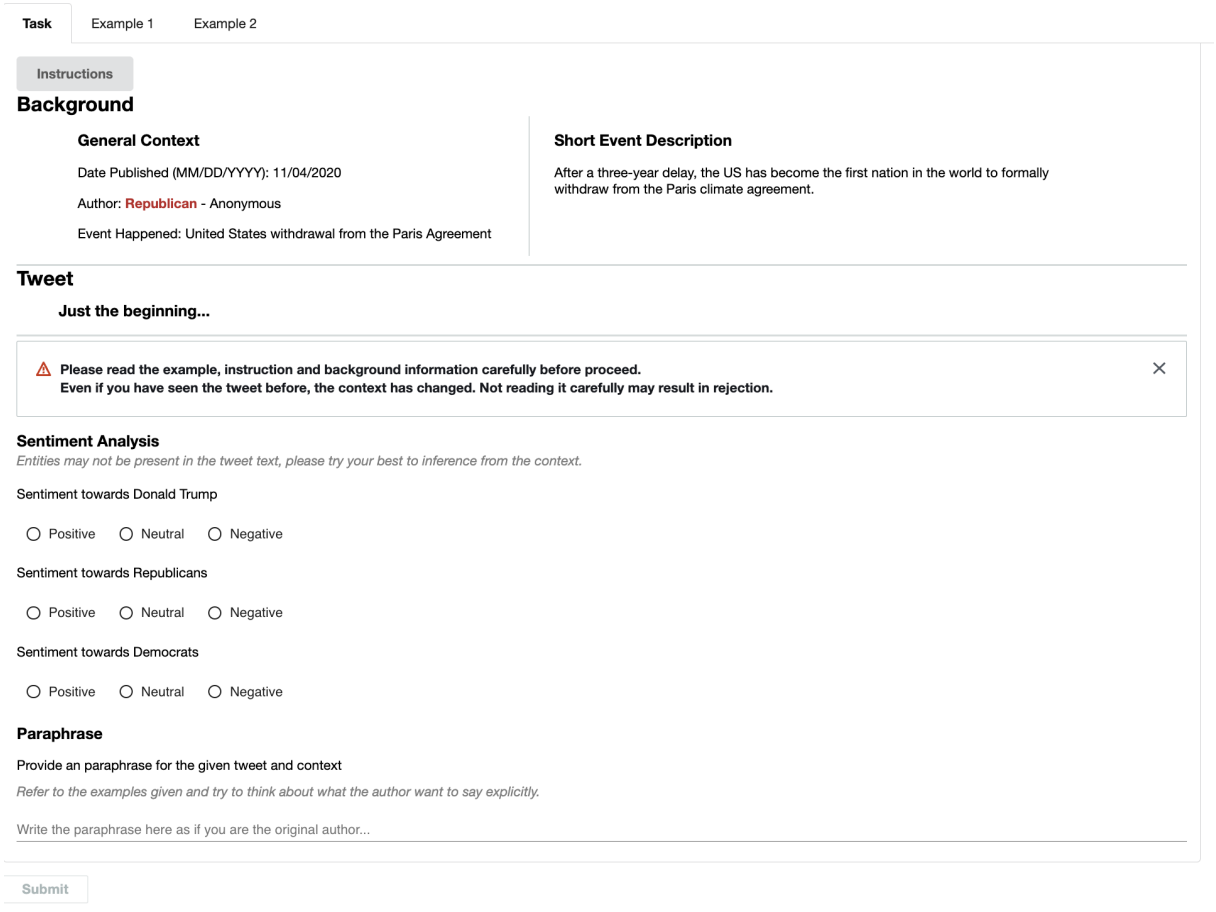


Figure 3: An example of *Vague Text Disambiguation* GUI