Which side are you on? Insider-Outsider classification in conspiracy-theoretic social media

Pavan Holur[1], **Tianyi Wang**[1], **Shadi Shahsavari**[1],
Timothy Tangherlini[2], and **Vwani Roychowdhury**[1]
[1] Department of Electrical and Computer Engineering, UCLA
[2] Department of Scandinavian, UC Berkeley
{pholur,tianyiw,shadihpp,vwani}@ucla.edu, tango@berkeley.edu

Abstract

Social media is a breeding ground for threat narratives and related conspiracy theories. In these, an *outside* group threatens the integrity of an *inside* group, leading to the emergence of sharply defined group identities: *Insider*s – agents with whom the authors identify and *Outsider*s – agents who threaten the insiders. Inferring the members of these groups constitutes a challenging new NLP task: (i) Information is distributed over many poorly-constructed posts; (ii) Threats and threat agents are highly contextual, with the same post potentially having multiple agents assigned to membership in either group; (iii) An agent's identity is often implicit and transitive; and (iv) Phrases used to imply *Outsider* status often do not follow common negative sentiment patterns. To address these challenges, we define a novel *Insider-Outsider* classification task. Because we are not aware of any appropriate existing datasets or attendant models, we introduce a labeled dataset (CT5K) and design a model (NP2IO) to address this task. NP2IO leverages pretrained language modeling to classify *Insider*s and *Outsider*s. NP2IO is shown to be robust, generalizing to noun phrases not seen during training, and exceeding the performance of non-trivial baseline models by 20%.

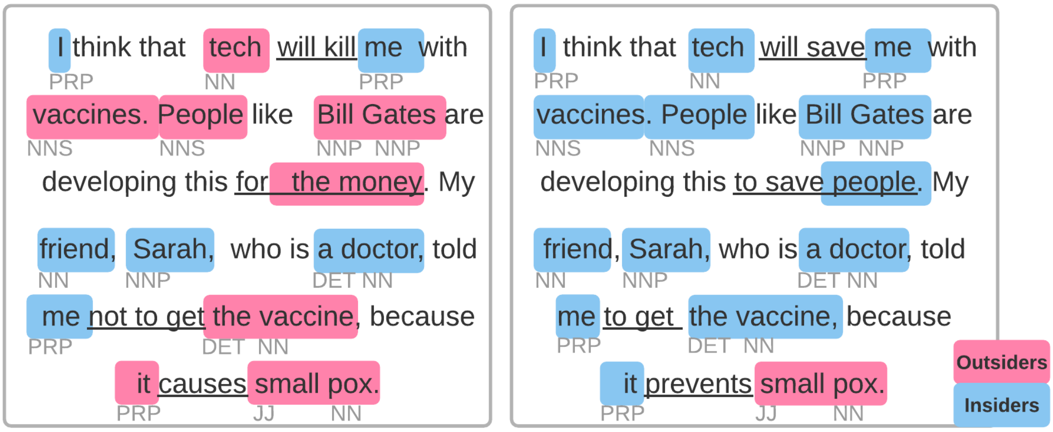
1 Background and Motivation

Narrative models – often succinctly represented as a network of characters, their roles, their interactions (*syuzhet*) and associated time-sequencing information (*fabula*) – have been a subject of considerable interest in computational linguistics and narrative theory. Stories rest on the generative backbone of narrative frameworks (Bailey, 1999; Beatty, 2016). While the details might vary from one story to another, this variation can be compressed into a limited set of domain-dependent narrative roles and functions (Dundes, 1962).

Social narratives that both directly and indirectly contribute to the construction of individual and group identities are an emergent phenomenon resulting from distributed social discourse. Currently, this phenomenon is most readily apparent on social media platforms, with their large piazzas and niche enclaves. Here, multiple threat-centric narratives emerge and, often, over time are linked together into complex conspiracy theories (Tangherlini et al., 2020). Conspiracy theories, and their constituent threat narratives (legend, rumor, personal experience narrative) share a signature semantic structure: an implicitly accepted *Insider* group; a diverse group of threatening *Outsider*s; specific threats from the *Outsider* directed at the *Insider*s; details of how and why *Outsider*s are threatening; and a set of strategies proposed for the *Insider*s to counter these threats (Tangherlini, 2018). Indeed, the *Insider*/*Outsider* groups are fundamental in most studies of belief narrative, and have been exhaustively studied in social theory and more specifically, in the context of conspiracy theories (Bodner et al., 2020; Barkun, 2013). On social media, these narratives are negotiated one post at a time, expressing only short pieces of the "immanent narrative whole" (Clover, 1986). This gives rise to a new type of computational linguistic problem: *Given a large enough corpus of social media text data, can one automatically distill semantically-labeled narratives (potentially several overlapping ones) that underlie the fragmentary conversational threads?*

Recent work (Shahsavari et al., 2020b; Tangherlini et al., 2020; Shahsavari et al., 2020a; Holur et al., 2021) has shown considerable promise that such scalable automated algorithms can be designed. An automated pipeline of interlocking machine learning modules decomposes the posts into actors, actants and their inter-actant relationships to create narrative networks via aggregation. *These network representations are interpretable on inspection*, allowing for the easy identification of the various signature semantic structures: *Insider*s,

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4975 - 4987
May 22-27, 2022 ©2022 Association for Computational Linguistics



I think that tech will kill me with vaccines. People like Bill Gates are developing this for the money. My friend, Sarah, who is a doctor, told me not to get the vaccine, because it causes small pox.

I think that tech will save me with vaccines. People like Bill Gates are developing this to save people. My friend, Sarah, who is a doctor, told me to get the vaccine, because it prevents small pox.

Figure 1: **A pair of inferred text segments labeled by NP2IO showing *Insider-Outsider* context-sensitivity:** Colored spans are used to highlight noun phrases that are inferred (red for *Outsider*s; blue for *Insider*s). POS tags are shown along with the noun phrases to illustrate an example of syntactic and semantic hints used by NP2IO to generate the inferred labels. Note that, based solely on context, the same agents ("tech", "vaccines", "People", "Bill Gates" and "the vaccine") switch *Insider-Outsider* label. *Even though the training data is highly biased in terms of the identities of the Insiders/Outsiders, the pretrained language model used in our classifier allows NP2IO to learn to infer using the context phrases and* not *by memorizing the labels.*

*Outsider*s, strategies for dealing with *Outsider*s and their attendant threats and, in the case of conspiracy theories, causal chains of events that support that theory.

By itself, this unsupervised platform does not "understand" the different narrative parts. Since the submodules are not trained to look for specific semantic abstractions inherent in conspiracy theories, the platform cannot automatically generate a semantically tagged narrative for downstream NLP tasks. It cannot, for example, generate a list across narratives of the various outside threats and attendant inside strategies being recommended on a social media forum, nor can it address why these threats and strategies are being discussed.

2 The Novel Insider vs. Outsider Classification Problem

As a fundamental first step bringing in supervised information to enable automated narrative structure discovery, we introduce the *Insider-Outsider* classification task: To classify the noun phrases in a post as *Insider*, *Outsider* or *neither*.

A working conceptualization of what we consider *Insider*s and *Outsider*s is provided in the following insets. As with most NLP tasks, we do not provide formal definitions of and rules to determine these groups. Instead we let a deep learning model learn the representations needed to capture these notions computationally by training on data annotated with human-generated labels.

Conceptualization of *Insider*s and *Outsider*s

Insiders: Some combination of actors and their associated pronouns, who display full agency (people, organizations, government), partial agency (policies, laws, rules, current events) or no agency (things, places, circumstances), with whom the author identifies (including themselves). These are often ascribed beneficial status;

Outsiders: A set of actors whom the author opposes and, in many cases, perceives as threatening the author and the insiders with disruption or harm. For our purposes, *these agents need not have full agency*: Diseases and natural disasters, for example, would be universal outsiders, and any man-made object/policy that works against the *Insider*s would be included in this group.

The partitioning of actors from a post into these different categories is inspired by social categorization, identification and comparison in the well-established Social Identity Theory (SIT) (Tajfel et al., 1979; Tajfel, 1974) and rests on established perspectives from Narrative Theory (Dundes, 1962; Labov and Waletzky, 1967; Nicolaisen, 1987).

Following are some of the reasons why this classification task is challenging and why the concepts of *Insiders*/*Outsiders* are not sufficiently captured by existing labeled datasets used in Sentiment

Analysis (SA) (discussed in more detail in Section 3):

1. **Commonly-held Beliefs and Worldviews:** Comprehensively incorporating shared values, crucial to the classification of *Insider*s and *Outsider*s, is a task with varied complexity. Some beliefs are easily enumerated: most humans share a perception of a nearly universal set of threats (virus, bomb, cancer, dictatorship) or threatening actions ("kills millions of people", "tries to mind-control everyone") or benevolent actions ("donating to a charitable cause", "curing disease", "freeing people"). Similarly, humans perceive themselves and their close family units as close, homogeneous groups with shared values, and therefore "I", "us", "my children" and "my family" are usually *Insider*s. In contrast, "they" and "them" are most often *Outsider*s.

Abstract beliefs pose a greater challenge as the actions that encode them can be varied and subtle. For example, in the post: "The microchips in vaccines track us", the noun phrase "microchips" is in the *Outsider* category as it violates the *Insider*s' right to privacy by "track[ing] us". Thus, greater attention needs to be paid in labeling datasets, highlighting ideas such as the right to freedom, religious beliefs, and notions of equality.

2. **Contextuality and Transitivity:** People express their opinions of *Insider/Outsider* affiliation by adding *contextual* clues that are embedded in the language of social media posts. For example, a post "We should build cell phone towers" suggests that "cell phone towers" are helpful to *Insider*s, whereas a post "We should build cell phone towers and show people how it fries their brains" suggests, in contrast, that "cell phone towers" are harmful to *Insiders* and belong, therefore, to the class of *Outsider*s. *Insider/Outsider* affiliations are also implied in a *transitive* fashion within a post. For example, consider two posts: (i) "Bill Gates is developing a vaccine. Vaccines *kill* people." and (ii) "Bill Gates is developing a vaccine. Vaccines *can eradicate* the pandemic." In the first case, the vaccine's toxic quality and attendant *Outsider* status would transfer to Bill Gates, making him an *Outsider* as well; in the second post, vaccine's beneficial qualities would transfer to him, now making "Bill Gates" an *Insider*.

3. **Model Requirement under Biased Data Conditions:** Designing effective classifiers that do not inherit bias from the training data – especially data in which particular groups or individuals are derided or dehumanized – is a challenging but necessary task. Because conspiracy theories evolve, building on earlier versions, and result in certain communities and individuals being "othered", our models *must* learn the phrases, contexts, and transitivity used to ascribe group membership, here either *Insider*s or *Outsider*s and not memorize the communities and/or individuals being targeted. Figure 1 illustrates an example where we probed our model to explore whether such a requirement is indeed satisfied. The first text conforms to the bias in our data, where "tech", "Bill Gates", and "vaccines" are primarily *Outsiders*. The second text switches the context by changing the phrases. Our classifier is able to correctly label these same entities, now presented in a different context, as *Insider*s! We believe that such subtle learning is possible because of the use of pretrained language models. We provide several such examples in Table 3 and Figure 3 and also evaluate our model for Zero-shot learning in Table 1 and Figure 6.

3 Our Framework and Related Work

Recent NLP efforts have examined the effectiveness of using pretrained Language Models (LM) such as BERT, DistilBERT, RoBERTa, and XLM to address downstream classification tasks through fine-tuning (Sanh et al., 2020; Liu et al., 2019; Lample and Conneau, 2019). Pretraining establishes the contextual dependencies of language prior to addressing a more specialized task, enabling rapid and efficient transfer learning. A crucial benefit of pretraining is that, in comparison to training a model from scratch, fewer labeled samples are necessary. By fine-tuning a pretrained LM, one can subsequently achieve competitive or better performance on an NLP task. As discussed in Section 2, since our model is required to be *contextual* and *transitive*, both of which are qualities that rely on the context embedded in language, we utilize a similar architecture.

In recent work involving span-based classification tasks, token-classification heads have proven to be very useful for tasks such as, Parts-of-Speech (POS) Tagging, Named Entity Recognition (NER) and variations of Sentiment Analysis (SA) (Yang et al., 2019; Vlad et al., 2019; Yin et al., 2020). Since the *Insider-Outsider* classification task is also set up as a noun phrase labeling task, our architecture uses a similar token-classification head on top of the pretrained LM backbone.

Current SA datasets' definitions of positive negative and neutral sentiments can be thought of as a "particularized" form of the *Insider-Outsider* classification task. For example, among the popular datasets used for SA, Rotten Tomatoes, Yelp reviews (Socher et al., 2013) and others (Dong et al., 2014; Pontiki et al., 2014) implicitly associate a sentiment's origin to the post's author (source) (a single *Insider*) and its intended target to a movie or restaurant (a single *Outsider* if the sentiment is *negative* or an *Insider* if *positive*). The post itself generally contains information about the target and particular aspects that the *Insider* found necessary to highlight.

In more recent SA work, such as Aspect-Based Sentiment Analysis (ABSA) (Gao et al., 2021; Li et al., 2019; Wang et al., 2021; Dai et al., 2021), researchers have developed models to extract sentiments – positive, negative, neutral – associated with particular aspects of a target *entity*. One of the subtasks of ABSA, aspect-level sentiment classification (ALSC), has a form that is particularly close to the *Insider-Outsider* classification. Interpreted in the context of our task, the author of the post is an *Insider* although now there can potentially be multiple targets or "aspects" that need to be classified as *Insider*s and *Outsider*s. Still, the constructed tasks in ABSA appear to not align well with the goal of *Insider-Outsider* classification: 1) Datasets are not *transitive*: Individual posts appear to have only one agent that needs classification, or a set of agents, each with their own separate sets of descriptors; 2) The ALSC data is often at the sentence-level as opposed to post-level, limiting the context-space for inference. Despite these obvious differences, we quantitatively verify our intuitions in Section 7.1, and show that ABSA models do not generalize to our dataset.

Closely related to ABSA is Stance Classification (SC) (also known as Stance Detection / Identification), the task of identifying the stance of the text author (`in favor of`, `against` or `neutral`) toward a target (an entity, concept, event, idea, opinion, claim, topic, etc.)(Walker et al., 2012; Zhang et al., 2017; Küçük and Can, 2021). Unlike ABSA, the target in SC does not need to be embedded as a span within the context. For example, a perfect SC model given an input for classification of context: *This house would abolish the monarchy.* and target: *Hereditary succession*, would predict the *Negative* label (Bar-Haim et al., 2017; Du et al., 2017). While SC appears to require a higher level of abstraction and, as a result, a model of higher complexity and better generalization power than those typically used for ABSA, current implementations of SC are limited by the finite set of queried targets; in other words, SC models currently do not generalize to unseen abstract targets. Yet, in real-time social media, potential targets and agents exhibit a continuous process of emergence, combination and dissipation. We seek to classify these shifting targets using the transitive property of language, and would like the language to provide clues about the class of one span *relative* to another. Ultimately, while SC models are a valuable step in the direction of better semantic understanding, they are ill-suited to our task.

Parallel to this work in SA, there are complementary efforts in consensus threat detection on social media (Wester et al., 2016; Kandias et al., 2013; Park et al., 2018), a task that broadly attempts to classify longer segments of text – such as comments on YouTube or tweets on Twitter – as more general "threats". The nuanced instruction to the labelers of the data is to *identify whether the author of the post is an Outsider from the labeler's perspective as an Insider*. Once again, we observe that this task aligns with the *Insider-Outsider* paradigm, but does not exhaust it, and the underlying models cannot accomplish our task.

The sets of *Insider*s and *Outsider*s comprise a higher-order belief system that cannot be adequately captured with the current working definitions of sentiment nor the currently available datasets. This problem presents *a primary motivation for creating a new dataset*. For example, the post: "Microchips are telling the government where we are", does not directly feature a form of prototypical sentiment associated with "microchips", "the government" and "we", yet clearly insinuates an invasion on our right to privacy making clear the *Insider*s ("we") and *Outsider*s ("microchips", "the government") in the post.

4 Data Collection

To construct our novel dataset – **C**onspiracy **T**heory-5000 (**CT5K**) – we designed crawlers to extract a corpus of social media posts generated by the underlying narrative framework of vaccine hesitancy (Details of the crawlers are documented in Appendix A.1). Vaccine hesitancy is a remarkably resilient belief fueled by conspiracy theories that overlaps with multiple other narratives including ones addressing "depopulation", "government overreach and the deep state", "limits on freedom of choice" and "Satanism". The belief's evolution on social media has already enabled researchers to take the first steps in modeling critical parts of the underlying generative models that drive anti-vaccination conversations on the internet (Tangherlini et al., 2016; Bandari et al., 2017). Moreover, vaccine hesitancy is especially relevant in the context of the ongoing COVID-19 pandemic (Burki, 2020).

On the crawled corpus, we extract the noun-chunks from each post using SpaCy's noun chunk extraction module and dependency parsers (Honnibal and Johnson, 2015). A noun chunk is a subtree of the dependency parse tree, the headword of which is a noun. The result is a set of post-phrase pairs, $(\mathtt{p}, \mathtt{n})$, where $\mathtt{p}$ is a post and $\mathtt{n}$ is one of the noun phrases extracted from the post.

Amazon Mechanical Turk (AMT) (see Appendix A.2 for labeler instructions) was used to label the post-phrase pairs. For each pair, the labeler was asked, *given the context*, whether the writer of the post $\mathtt{p}$ perceives the noun phrase $\mathtt{n}$ to be an *Insider*, *Outsider* or *neither* (N/A). The labeler then provides a label $c \in \mathcal{C}$, where $\mathcal{C} = \{\textit{Insider}, \textit{Outsider}, \textit{N/A}\}$ (hence $|\mathcal{C}| = 3$). The triplets of post-phrase pairs along with their labels form the dataset $\mathcal{D} = \left\{\left((\mathtt{p}_i, \mathtt{n}_i), c_i\right)\right\}_{i=1}^{|\mathcal{D}|}$. Note that a single post can appear in multiple triplets, because multiple different noun phrases can be extracted and labeled from a single post. The overall class distribution and a few conditional class distributions across the labeled samples for several particular noun phrases are provided in Figure 5 in the Appendix B.

Manual inspection of the labeled samples $((\mathtt{p}, \mathtt{n}), c)$ suggests that the quality of the dataset is good ($< 10\%$ misclassified by random sampling). The now-labeled CT5K dataset (Holur et al., 2022)[1] ($|\mathcal{D}| = 5000$ samples) is split into training (90%), and 10% testing sets. 10% of the training set is held out for validation. The final training set is 20-fold augmented by BERT-driven multi-token insertion (Ma, 2019).

5 Methodology and Pipeline

The **N**oun-**P**hrase-to-***I**nsider*-***O**utsider* (NP2IO) model [2] adopts a token classification architecture comprising a BERT-like pre-trained backbone and a softmax classifier on top of the backbone. Token-level labels are induced from the span-level labels for the fine-tuning over CT5K, and the span-level labeling of noun phrases is done through majority vote during inference.

5.1 Fine-tuning Details

An outline of the fine-tuning pipeline is provided in Figure 2.

Given a labeled example $((\mathtt{p}, \mathtt{n}), c)$, the model labels each token t_i in the post $\mathtt{p} = [t_1, \ldots, t_N]$, where N is the number of tokens in the post $\mathtt{p}$. The BERT-like backbone embeds each token t_i into a contextual representation $\Phi_i \in \mathbb{R}^d$ (for example, $d = 768$ for BERT-base or RoBERTa-base). The embedding is then passed to the softmax classification layer

$$\boldsymbol{\pi}_i \triangleq \text{Softmax}(\mathbf{W}^T \Phi_i + \mathbf{b}) \qquad (1)$$

where $\boldsymbol{\pi}_i \in \Delta^{|\mathcal{C}|}$ is the *Insider-Outsider* classification prediction probability vector of the i^{th} token, and $\mathbf{W} \in \mathbb{R}^{d \times |\mathcal{C}|}$ and $\mathbf{b} \in \mathbb{R}^{|\mathcal{C}|}$ are the parameters of the classifier.

The ground truth class label c accounts for all occurrences of the noun phrase $\mathtt{n}$ in the post $\mathtt{p}$. We use this span-level label to induce the token-level label and facilitate the computation of the fine-tuning loss.

Concretely, consider the spans where the noun phrase $\mathtt{n}$ occurs in the post $\mathtt{p}$: $S_{\mathtt{n}} = \{s_1, \ldots, s_M\}$, where $s_j \in S_{\mathtt{n}}$ denotes the span of the j^{th} occurrence of $\mathtt{n}$, and M is the number of occurrences of $\mathtt{n}$ in $\mathtt{p}$. Each span is a sequence of one or more tokens. The set of tokens appearing in one of these labeled spans is:

$$T_{\mathtt{n}} = \{t \in \mathtt{p} \mid \exists s \in S_{\mathtt{n}} \text{ s.t. } t \in s\}. \qquad (2)$$

We define the fine-tuning loss $\mathcal{L}$ of the labeled example $((\mathtt{p}, \mathtt{n}), c)$ as the cross-entropy (CE) loss computed over $T_{\mathtt{n}}$ using c as the label for each token in it,

$$\mathcal{L}(\mathtt{p}, \mathtt{n}, c) = \sum_{i : t_i \in T_{\mathtt{n}}} - \log\left((\boldsymbol{\pi}_i)_c\right) \qquad (3)$$

where $(\boldsymbol{\pi}_i)_c$ denotes the prediction probability for the class $c \in \mathcal{C}$ of the i^{th} token.

The fine-tuning is done with mini-batch gradient descent for the classification layer and a number of self-attention layers in the backbone. The number of fine-tuned self-attention layers is a hyperparameter. The scope of hyperparameter tuning is provided in Table 4.

[1] See: Data and Model Checkpoints

[2] Code Repository: NP2IO

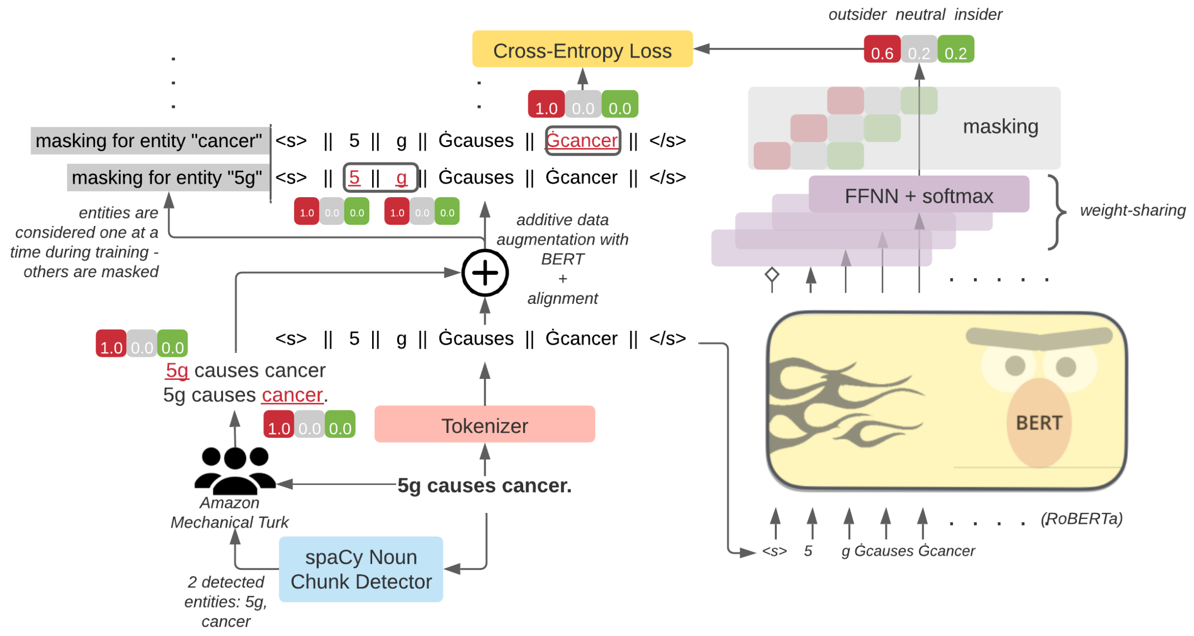

outsider neutral insider
0.6 0.2 0.2
Cross-Entropy Loss
1.0 0.0 0.0
masking
FFNN + softmax
weight-sharing
masking for entity "cancer"
<s> || 5 || g || Ġcauses || Ġcancer || </s>
masking for entity "5g"
<s> || 5 || g || Ġcauses || Ġcancer || </s>
1.0 0.0 0.0 1.0 0.0 0.0
entities are considered one at a time during training - others are masked
additive data augmentation with BERT + alignment
1.0 0.0 0.0
<s> || 5 || g || Ġcauses || Ġcancer || </s>
5g causes cancer
5g causes cancer.
1.0 0.0 0.0
Tokenizer
BERT
Amazon Mechanical Turk
5g causes cancer.
(RoBERTa)
<s> 5 g Ġcauses Ġcancer
spaCy Noun Chunk Detector
2 detected entities: 5g, cancer

Figure 2: **NP2IO - An Outline of the Fine-tuning Pipeline:** A post is tokenized and aligned to noun chunks that are independently identified from the post with a pre-trained SpaCy parser. The BERT model is fine-tuned to identify the labels of each token in context of a post based on AMT labels of the higher-order noun phrases. Loss is Cross-Entropy (CE) loss computed on only tokens relevant for detection post SpaCy-noun phrase identification.

5.2 Real-time Inference and Accuracy Measurement

During fine-tuning, we extend the label of a noun phrase to all of its constituent tokens; during inference, conversely, we summarize constituent token labels to classify the noun phrases by a majority vote. For a pair of post and noun-phrase $(\mathtt{p}, \mathtt{n})$, assuming the definition of $\{t_i\}_{i=1}^{N}$, $\{\boldsymbol{\pi}_i\}_{i=1}^{N}$ and $T_{\mathtt{n}}$ from the Section 5.1, the *Insider-Outsider* label prediction $\hat{c}$ is given by

$$\hat{c} = \arg\max_{k} \sum_{i:t_i \in T_{\mathtt{n}}} \mathbb{1}_{\{k=(\arg\max_{\kappa}(\boldsymbol{\pi}_i)_{\kappa})\}}. \qquad (4)$$

Now c can be compared to $\hat{c}$ with a number of classification evaluation metrics. Visual display of individual inference results such as those in Figure 1 are supported by displaCy (Honnibal and Montani, 2017).

6 Baseline Models

In this section, we list baselines that we compare to our model's performance ordered by increasing parameter complexity.

- **Random Model (RND):** Given a sample from the testing set $\{\mathtt{p}, \mathtt{n}\}$, $\hat{c}$ is randomly selected with uniform distribution from $\mathcal{C} = \{\textit{Insider}, \textit{Outsider}, \textit{N/A}\}$.
- **Deterministic Model (DET - I/O/NA):** For any post-phrase pair $(\mathtt{p}, \mathtt{n})$, give a fixed classification prediction: $\hat{c} = \textit{Insider}$ (DET-I), $\hat{c} = \textit{Outsider}$ (DET-O) or $\hat{c} = \textit{N/A}$ (DET-NA).
- **Naïve Bayes Model (NB / NB-L):** Given a training set, the naïve Bayes classifier estimates the likelihood of each class conditioned on a noun chunk $\mathcal{P}_{\mathcal{C},\mathcal{N}}(c|\mathtt{n})$ assuming its independence w.r.t. the surrounding context. That is, a noun phrase predicted more frequently in the training-set as an *Insider* will be predicted as an *Insider* during the inference, regardless of the context. For noun phrases not encountered during training, the uniform prior distribution over $\mathcal{C}$ is used for the prediction. The noun chunk may be lemmatized (by word) during training and testing to shrink the conditioned event space. We abbreviate the naïve Bayes model without lemmatization as NB, and the one with lemmatization as NB-L.
- **GloVe+CBOW+XGBoost (CBOW - 1/2/5):** This baseline takes into account the context of a post but uses global word embeddings, instead of contextual-embeddings. A window length w is fixed such that for each noun phrase, we extract the w words before and w words after the noun phrase, creating a set of context words, $\mathcal{S}_w$. Stopwords are filtered, and the remaining con-

text words are lemmatized and encoded via 300-dimensional GloVe (Pennington et al., 2014). The Continuous Bag of Words (CBOW) model (Mikolov et al., 2013) averages the representative GloVe vectors in $\mathcal{S}_w$ to create an aggregate contextual vector for the noun phrase. XGBoost (Chen and Guestrin, 2016) is used to classify the aggregated contextual vector. The same model is applied on the test set to generate labels. We consider window lengths of 1, 2 and 5 (CBOW-1, CBOW-2 and CBOW-5 respectively).

7 Results and Evaluation

Comparison of NP2IO to baselines is provided in Table 1. The random (RND) and deterministic (DET-I, DET-O, DET-NA) models perform poorly. We present these results to get a better sense of the unbalanced nature of the labels in the CT5K dataset (see Figure 5). The naïve Bayes model (NB) and its lemmatized form (NB-L) outperform the trivial baselines. However, they perform *worse* than the two contextual models, GloVe+CBOW+XGBoost and NP2IO. This fact validates a crucial property of our dataset: *Despite the bias in the gold standard labels for particular noun phrases such as "I", "they" and "microchip" – see Figure 5 in Appendix B – context dependence plays a crucial role in Insider-Outsider classification.* Furthermore, NP2IO outperforms GloVe+CBOW+XGBoost (CBOW-1, CBOW-2, CBOW-5) summarily. While both types of models employ context-dependence to classify noun phrases, NP2IO does so more effectively. The fine-tuning loss convergence plot for the optimal performing NP2IO model is presented in Figure 4 in Appendix B and model checkpoints are uploaded in the data repository.

7.1 Does CT5K really differ from prior ABSA datasets?

Given the limitations of current ABSA datasets for our task (see Section 2 and Section 3), we computationally show that CT5K is indeed a different dataset, particularly in comparison to other classical ones in Table 2. For this experiment, we train near-state-of-the-art ABSA models with RoBERTa-base backbone (Dai et al., 2021) on three popular ABSA datasets – Laptop reviews and Restaurant reviews from SemEval 2014 task 4 (Pontiki et al., 2014), and Tweets (Dong et al., 2014). Each trained model is then evaluated on all three datasets *as well as* the test set of CT5K. The *Insider* class in CT5K is mapped to the *positive* sentiment and the *Outsider* class to the *negative* sentiment. The F1-macro scores of the models trained and tested among the three ABSA datasets are much higher than the scores when testing on the CT5K dataset. *Clearly, models that are successful with typical ABSA datasets do not effectively generalize to CT5K, suggesting that our dataset is different.*

7.2 Classifying Noun Phrases at Zero-shot

A challenge for any model, such as NP2IO, is zero-shot performance, when it encounters noun phrases never tagged during training. Answering this question offers a means for validating the context-dependence requirement, mentioned in Section 2. This evaluation is conducted on a subset of the entire testing set: A sample of the subset $\{\mathtt{p}, \mathtt{n}\}$ is such that the word-lemmatized, stopword-removed form of $\mathtt{n}$ does not exist in the set of word-lemmatized, stopword-removed noun phrases seen during training. We extract 30% of test samples to be in this set. The results are presented in Table 1. As expected, the performance of the naïve Bayes models (NB, NB-L) degrades severely to random. The performance of the contextual models CBOW-1/2/5, and NP2IO stay strong, suggesting effective context sensitivity in inferring the correct labels for these models. A visualization of the zero-shot capabilities of NP2IO on unseen noun phrases is presented in Figure 6 in Appendix B.

7.3 Does NP2IO Memorize? An Adversarial Experiment

We construct a set of adversarial samples to evaluate the extent to which NP2IO accurately classifies a noun phrase that has a highly-biased label distribution in CT5K. We consider 3 noun phrases in particular: "microchip", "government", and "chemical". Each of these has been largely labeled as *Outsiders*. The adversarial samples for each phrase, in contrast, are manually aggregated (5 seed posts augmented 20 times each) to suggest that the phrase is an *Insider* (see Table 5 in Appendix B for the seed posts). We compute the recall of NP2IO in detecting these *Insider* labels (results in Table 3). NP2IO is moderately robust against adversarial attacks: *In other words, highly-skewed distributions of labels for noun phrases in our dataset do not appear to imbue a similar drastic bias into our model.*

	Performance on the Test Set					Performance in Zero-Shot				
Model	**Acc.**	**P**	**R**	**F1**	**F1(w)**	**Acc.**	**P**	**R**	**F1**	**F1(w)**
RND	0.334	0.343	0.334	0.321	0.350	0.280	0.273	0.241	0.239	0.316
DET-I	0.312	0.104	0.333	0.159	0.148	0.280	0.093	0.333	0.146	0.123
DET-O	0.504	0.168	0.333	0.223	0.338	0.593	0.198	0.333	0.248	0.442
DET-NA	0.184	0.061	0.333	0.104	0.057	0.127	0.042	0.333	0.075	0.028
NB	0.520	0.473	0.478	0.474	0.523	0.333	0.341	0.310	0.295	0.369
NB-L	0.468	0.397	0.387	0.386	0.453	0.360	0.389	0.434	0.356	0.373
CBOW-1	0.490	0.419	0.383	0.373	0.448	0.527	0.408	0.361	0.360	0.489
CBOW-2	0.520	0.462	0.415	0.410	0.484	0.553	0.441	0.375	0.368	0.509
CBOW-5	0.526	0.459	0.419	0.414	0.489	0.553	0.393	0.375	0.369	0.514
NP2IO	**0.650**	**0.629**	**0.546**	**0.534**	**0.619**	**0.693**	**0.682**	**0.536**	**0.543**	**0.671**

Table 1: **Performance of NP2IO versus multiple baselines on the test set:** Our model (in bold) performs competitively and outperforms Naïve Bayes, CBOW models across metrics. Furthermore, it retains its performance to classify noun phrases unseen (post-lemmatization and stopword removal) during training. Predictably, the performance of the Naïve Bayes classifier in this zero-shot setting drops drastically to near random.

Test Dataset	**Train Dataset**		
	Laptop	Restaurants	Tweets
Laptop	0.804	0.768	0.658
Restaurants	0.754	0.825	0.657
Tweets	0.526	0.546	0.745
CT5K	**0.347**	**0.424**	**0.412**

Table 2: **F1-macro scores for the ABSA model trained on conventional SA datasets from SemEval 2014 task 4:** All models perform poorly in testing on the CT5K dataset while performing well in testing on ABSA datasets. This suggests that the CT5K dataset is indeed differentiated from the ABSA datasets.

Noun Phrases	**CT5K *Outsider* Labels (%)**	***Insider* Recall in adversarial text**
microchip	100%	**80%**
government	80%	**89%**
chemical	100%	**62%**

Table 3: **Adversarial inferencing tasks for the trained NP2IO model:** Three noun phrases with very high *Outsider* status (100%, 80%, 100%, respectively) in the CT5K training set are used to construct posts where their contextual role is beneficial, and hence, should be labeled as *Insider* (see Section 2). The results show that NP2IO largely learned to use the contextual information for its inference logic, and did not memorize the agent bias in CT5K. We speculate that the exhibited bias towards "Chemicals" is due to the large body of text documents that discusses the adverse effects of chemicals, and hence is encoded in the embedding structure of pretrained LM models that NP2IO cannot always overrule; at least yet.

8 Concluding Remarks

We presented a challenging *Insider-Outsider* classification task, a novel framework necessary for addressing burgeoning misinformation and the proliferation of threat narratives on social media. We compiled a labeled CT5K dataset of conspiracy-theoretic posts from multiple social media platforms and presented a competitive NP2IO model that outperforms non-trivial baselines. We have demonstrated that NP2IO is contextual and transitive via its zero-shot performance, adversarial studies and qualitative studies. We have also shown that the CT5K dataset consists of underlying information that is different from existing ABSA datasets.

Given NP2IO's ability to identify *Insider*s and *Outsider*s in a text segment, we can extend the inference engine to an entire set of interrelated samples in order to extract, visualize and *interpret* the underlying narrative (see Figure 3). This marks a first and significant step in teasing out narratives from fragmentary social media records, with many of its essential semantic parts – such as, *Insider/Outsider* – tagged in an automated fashion. As extensive evaluations of the NP2IO model show, our engine has learned the causal phrases used to designate the labels. We believe an immediate future work can identify such causal phrases, yet another step toward semantic understanding of the parts of a narrative. Broadly, work similar to this promises to expedite the development of models that rely on a computational foundation of structured information, and that are better at *explaining* causal chains of inference, a particularly important feature in the tackling of misinforma-

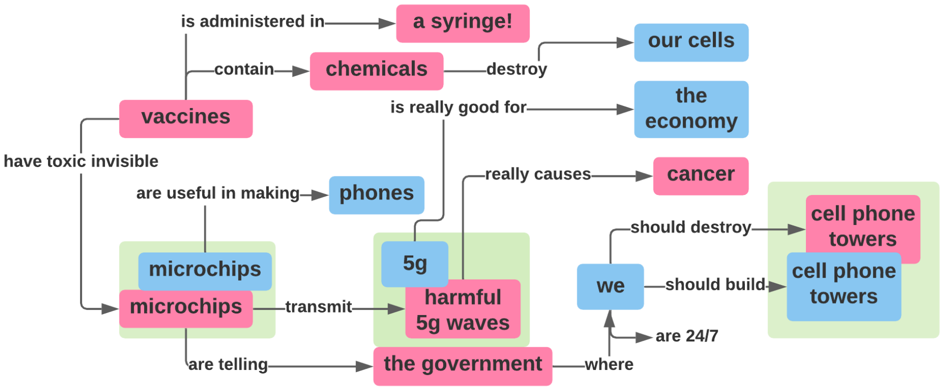


Figure 3: **An actor-actant subnarrative network constructed from social media posts:** Selected posts from anti-vaccination forums such as *qresearch* on 4chan were decomposed into relationship tuples using a state-of-the-art relationship extraction pipeline from previous work (Tangherlini et al., 2020) and these relationships are overlayed with the inferences from NP2IO. This results in a network where the nodes are the noun phrases and the edges are the verb phrases, with each edge representing an extracted relationship from a post. In this network, a connected component emerged capturing a major sub-theory in vaccine hesitancy. This highlights NP2IO's ability at inferring both the threat-centric orientation of the narrative space as well as the negotiation dynamics in play, thereby providing qualitative insight into how NP2IO may be used in future work to extract large-scale relationship networks that are interpretable. The green boxes highlight the noun phrases that have contradictory membership in the *Insider*s and the *Outsider*s classes as their affiliations are deliberated.

tion. Indeed, NP2IO's success has answered the question: "Which side are you on?" What remains to be synthesized from language is: "Why?"

References

Paul Bailey. 1999. Searching for storiness: Story-generation from a reader's perspective. In *Working notes of the Narrative Intelligence Symposium*, pages 157–164.

Roja Bandari, Zicong Zhou, Hai Qian, Timothy R. Tangherlini, and Vwani P. Roychowdhury. 2017. A resistant strain: Revealing the online grassroots rise of the antivaccination movement. *Computer*, 50(11):60–67.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain. Association for Computational Linguistics.

Michael Barkun. 2013. *A Culture of Conspiracy: Apocalyptic Visions in Contemporary America*. University of California Press.

John Beatty. 2016. What are narratives good for? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 58:33–40.

John Bodner, Wendy Welch, and Ian Brodie. 2020. *COVID-19 conspiracy theories: QAnon, 5G, the New World Order and other viral ideas*. McFarland.

Talha Burki. 2020. The online anti-vaccine movement in the age of covid-19. *The Lancet Digital Health*, 2(10):e504–e505.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

David Chong, Erl Lee, Matthew Fan, Pavan Holur, Shadi Shahsavari, Timothy Tangherlini, and Vwani Roychowdhury. 2021. A real-time platform for contextualized conspiracy theory analysis. In *2021 International Conference on Data Mining Workshops (ICDMW) (forthcoming)*. IEEE.

Carol J Clover. 1986. The long prose form. *Arkiv för nordisk filologi*, 101:10–39.

Junqi Dai, Hang Yan, Tianxiang Sun, Pengfei Liu, and Xipeng Qiu. 2021. Does syntax matter? A strong baseline for aspect-based sentiment analysis with roberta. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1816–1829. Association for Computational Linguistics.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54, Baltimore, Maryland. Association for Computational Linguistics.

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3988–3994.

Alan Dundes. 1962. *The Morphology of North American Indian Folktales*. Indiana University.

Lei Gao, Yulong Wang, Tongcun Liu, Jingyu Wang, Lei Zhang, and Jianxin Liao. 2021. Question-driven span labeling model for aspect–opinion pair extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12875–12883.

Pavan Holur, David Chong, Erl Lee, Matthew Fan, Shadi Shahsavari, Tianyi Wang, Timothy R Tangherlini, and Vwani Roychowdhury. 2022. Acl 2022 - supplementary data files.

Pavan Holur, Shadi Shahsavari, Ehsan Ebrahimzadeh, Timothy R. Tangherlini, and Vwani Roychowdhury. 2021. Modelling social readers: novel tools for addressing reception from online book reviews. *Royal Society Open Science*, 8(12):210797.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Miltiadis Kandias, Vasilis Stavrou, Nick Bozovic, and Dimitris Gritzalis. 2013. Proactive insider threat detection through social media: The youtube case. In *Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society*, pages 261–266.

Dilek Küçük and Fazli Can. 2021. Stance Detection: A Survey. *ACM Computing Surveys*, 53(1):1–37.

William Labov and Joshua Waletzky. 1967. Narrative analysis. inj. helm (ed.), essays on the verbal and visual arts (pp. 12–44).

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting BERT for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 34–41, Hong Kong, China. 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.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Wilhelm FH Nicolaisen. 1987. The linguistic structure of legends. *Perspectives on Contemporary Legend*, 2(1):61–67.

Won Park, Youngin You, and Kyungho Lee. 2018. Detecting potential insider threat: Analyzing insiders' sentiment exposed in social media. *Security and Communication Networks*, 2018.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Shadi Shahsavari, Ehsan Ebrahimzadeh, Behnam Shahbazi, Misagh Falahi, Pavan Holur, Roja Bandari, Timothy R. Tangherlini, and Vwani Roychowdhury. 2020a. An automated pipeline for character and relationship extraction from readers literary book reviews on goodreads.com. In *12th ACM Conference on Web Science*, WebSci '20, page 277–286, New York, NY, USA. Association for Computing Machinery.

Shadi Shahsavari, Pavan Holur, Tianyi Wang, Timothy R Tangherlini, and Vwani Roychowdhury. 2020b. Conspiracy in the time of corona: automatic detection of emerging covid-19 conspiracy theories in social media and the news. *Journal of computational social science*, 3(2):279–317.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Henri Tajfel. 1974. Social identity and intergroup behaviour. *Social science information*, 13(2):65–93.

Henri Tajfel, John C Turner, William G Austin, and Stephen Worchel. 1979. An integrative theory of intergroup conflict. *Organizational identity: A reader*, 56(65):9780203505984–16.

Timothy R Tangherlini. 2018. Toward a generative model of legend: Pizzas, bridges, vaccines, and witches. *Humanities*, 7(1):1.

Timothy R Tangherlini, Vwani Roychowdhury, Beth Glenn, Catherine M Crespi, Roja Bandari, Akshay Wadia, Misagh Falahi, Ehsan Ebrahimzadeh, and Roshan Bastani. 2016. "mommy blogs" and the vaccination exemption narrative: Results from a machine-learning approach for story aggregation on parenting social media sites. *JMIR Public Health Surveill*, 2(2):e166.

Timothy R Tangherlini, Shadi Shahsavari, Behnam Shahbazi, Ehsan Ebrahimzadeh, and Vwani Roychowdhury. 2020. An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, pizzagate and storytelling on the web. *PloS one*, 15(6):e0233879.

Maxim Tkachenko, Mikhail Malyuk, Nikita Shevchenko, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2021. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

George-Alexandru Vlad, Mircea-Adrian Tanase, Cristian Onose, and Dumitru-Clementin Cercel. 2019. Sentence-level propaganda detection in news articles with transfer learning and BERT-BiLSTM-capsule model. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 148–154, Hong Kong, China. Association for Computational Linguistics.

Marilyn Walker, Pranav Anand, Rob Abbott, and Ricky Grant. 2012. Stance Classification using Dialogic Properties of Persuasion. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 592–596, Montréal, Canada. Association for Computational Linguistics.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Automated Concatenation of Embeddings for Structured Prediction. In *the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (**ACL-IJCNLP 2021**)*. Association for Computational Linguistics.

Aksel Wester, Lilja Øvrelid, Erik Velldal, and Hugo Lewi Hammer. 2016. Threat detection in online discussions. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 66–71.

Kisu Yang, Dongyub Lee, Taesun Whang, Seolhwa Lee, and Heuiseok Lim. 2019. Emotionx-ku: Bert-max based contextual emotion classifier.

Da Yin, Tao Meng, and Kai-Wei Chang. 2020. SentiBERT: A transferable transformer-based architecture for compositional sentiment semantics. In *Proceedings of the 58th Conference of the Association for Computational Linguistics, ACL 2020, Seattle, USA*.

Shaodian Zhang, Lin Qiu, Frank Chen, Weinan Zhang, Yong Yu, and Noémie Elhadad. 2017. We Make Choices We Think are Going to Save Us: Debate and Stance Identification for Online Breast Cancer CAM Discussions. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 1073–1081, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Appendices

A Data Collection

A.1 Automated Crawling of Social Media

A daily data collection method (Chong et al., 2021) aggregates heterogeneous data from various social media platforms including Reddit, YouTube, 4chan and 8kun. Our implementation of this pipeline has extracted potentially conspiracy theoretic posts between March 2020 and June 2021. We select a subset of these posts that are relevant to vaccine hesitancy and that include (a) at least one of the words in ['vaccine', 'mrna', 'pfizer', 'moderna', 'j&j', 'johnson', 'chip', 'pharm'] and (b) between 150 to 700 characters. The end-to-end data processing pipeline is *uncased*.

A.2 Instructions to AMT Labelers

Amazon Mechanical Turk Labelers were required to be at the Masters' level (exceeding a trust baseline provided by Amazon), were required to speak English, and were required to be residing in the United States. No personally identifying information was collected. Users were asked to create an account on a LabelStudio (Tkachenko et al., 2020-2021) platform to answer a set of 60-80 questions or 2 hours worth of questions. Each question included (a) A real anonymized social media post with a highlighted sentence, (b) The sentence highlighted in (a) but with the noun phrase of interest highlighted. The question prompt read: *Please let us know whether the entity highlighted in bold AS PERCEIVED BY THE WRITER is a good/bad or neutral entity.*

Labelers were reminded several times via popups and other means that the labels were to be chosen with respect to the author of the post and not the labeler's inherent biases and/or political preferences.

A.3 Ethics statement about the collected social media data

Data was collected using verified research Application Programming Interfaces (API) provided by the social media companies for non-commercial study. In order to explore data on fringe platforms such as 4chan and 8kun where standard APIs are not available, the data was scraped using a Selenium-based crawler. All the retrieved samples were ensured to be public: the posts could be accessed by anyone on the internet without requiring explicit consent by the authors. Furthermore, we made sure to avoid using Personal Identifiable Information (PII) such as the user location, time of posting and other metadata: indeed, we hid even the specific social media platform from which a particular post was mined. The extracted text was cleaned by fixing capitalization, filtering special characters, adjusting inter-word spacing and correcting punctuation, all of which further obfuscated the identity of the author of a particular post.

B Supplementary Figures and Tables

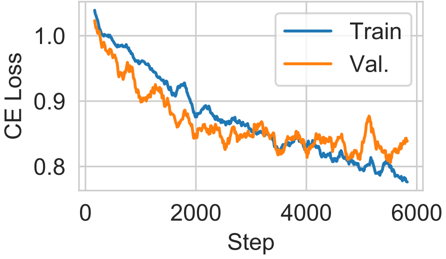


Figure 4: **Convergence plot for NP2IO:** Shown above is the training and validation CE loss with optimal parameters (in bold) from Table 4. Model checkpoints are in data repository.

Parameters	Values
Batch Size	32, **64**, 128
Trainable Layers	0, 1, **2**, 5
LR	1E-7, **1E-6**, 1E-5, 1E-4
Pretrained Backbone	BERT-base, **DistilBERT-base**, RoBERTa-base, RoBERTa-large

Table 4: **A summary of the parameters considered for finetuning:** NP2IO's best-performing (by validation loss) parameters are in bold.

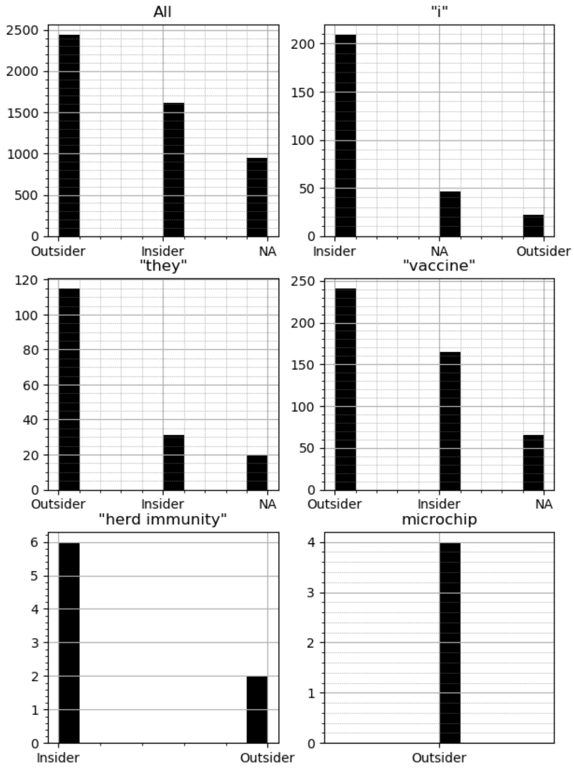


Figure 5: **Histograms that show the distributions of labels in CT5K**: The plot for "All" represents the full 3-category label distribution across all entities, for "I" the bias toward *Insider*s is evident, "They" are mostly outsiders, and there is no clear consensus label for "Vaccine" and "Herd Immunity". Microchips are always tagged as *Outsider*s.

NP	Seed Posts (augmented to 100 posts per NP)
microchip	"I love **microchip**s.", "I feel that **microchip**s are great.", "**microchip**s are lovely and extremely useful.", "I believe **microchip**s are useful in making phones.", "**Microchip**s have made me a lot of money."
government	"The **government** helps keep me safe.","The **government** does a good job.","I think that without the **government**, we would be worse off.","The **government** keeps us safe.", "A **government** is important to keep our society stable."
chemical	"**Chemical**s save us.","**Chemical**s can cure cancer.","I think **chemical**s can help elongate our lives.","I think **chemical**s are great and helps keep us healthy.","**Chemical**s can help remove ringworms."

Table 5: **The set of** 5 ***Insider*-oriented core posts per noun phrase (in bold) that have a high skew toward *Outsider* labels in CT5K:** Each seed post is augmented 20 times to create a set of 100 adversarial posts per phrase. NP2IO infers the label for the key noun phrase across these samples. The *adversarial* recall is presented in Table 3.

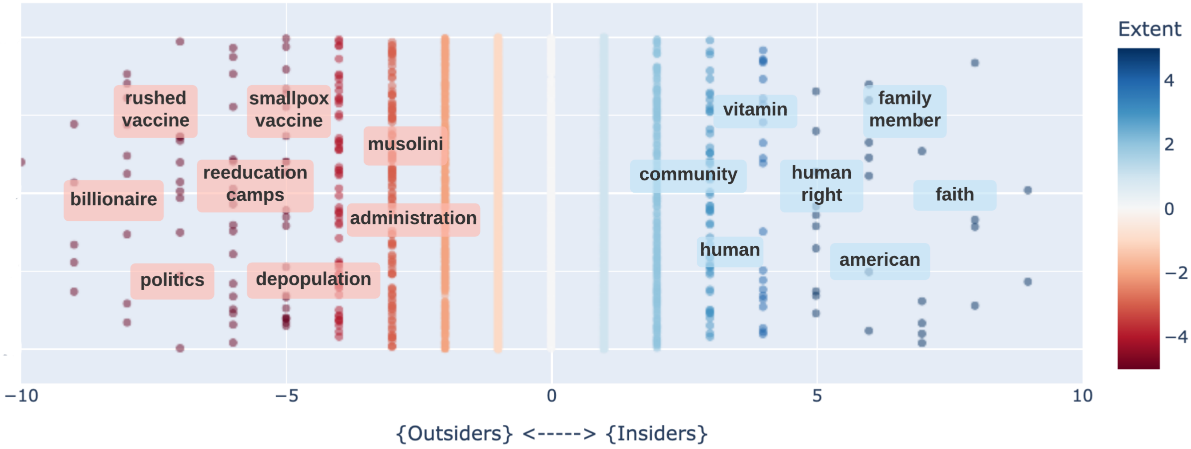


Figure 6: **Zero-shot *Insider-Outsider* Classification Profile:** This figure shows the consensus vote for noun phrases that do not occur in the training set. The x-axis represents the consensus-vote-count and the y-axis, the indices of the noun phrases. The consensus vote is computed for each noun phrase $\mathtt{n}$ by passing all the posts that include $\mathtt{n}$ through NP2IO. Each *Insider* vote is $+1$ and *Outsider* vote is -1. The consensus-vote-count is also color-coded for better visualization. The zero-shot classification is qualitatively observed to correctly classify popular noun phrases such as “reeducation camps”,“depopulation” as *Outsider*s and “american” and “faith” as *Insider*s in the subnarrative of the anti-vaccination movement.