NEURAL-SYMBOLIC MODELING FOR NATURAL LANGUAGE DISCOURSE

by

Maria Leonor Pacheco

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Computer Science

West Lafayette, Indiana

May 2022

THE PURDUE UNIVERSITY GRADUATE SCHOOL STATEMENT OF COMMITTEE APPROVAL

Dr. Dan Goldwasser, Chair

Department of Computer Science

Dr. Jean Honorio

Department of Computer Science

Dr. Tiark Rompf

Department of Computer Science

Dr. Ming Yin

Department of Computer Science

Approved by:

Dr. Kihong Park

To my family and friends, scattered all over the world.

ACKNOWLEDGMENTS

I am immensely grateful to my advisor, Dan Goldwasser, for his guidance and support throughout my PhD studies. His contributions were key in the conception of the DRaiL framework in the Summer of 2016, and his ongoing support and encouragement proved invaluable to see this ambitious project through. I learned so much from Dan throughout these years, from how to ask the right questions to how to communicate my research. I hope to emulate his best advising qualities with my future students.

Many thanks goes to my committee members, Jean Honorio, Ming Yin and Tiark Rompf, for their insightful comments and feedback during my preliminary examination and defense. Jean and Ming's intellectual contributions were also key to develop the work presented in Chapters 4 and 6, respectively. A special thanks goes to Professor Cristina Nita-Rotaru. Our years-long collaboration taught me how to approach interdisciplinary work effectively. Working with Cristina has been almost like having a second advisor, and I am very grateful for having had the opportunity to learn from her. I would like to acknowledge all of my other research collaborators inside and outside Purdue, whose contributions made much of the work presented in this dissertation possible: Lyle Ungar, Manuel Widmoser, Ayush Jain, Tunazzina Islam, Shamik Roy, I-Ta Lee, Nikhil Mehta, Xiao Zhang, Chang Li, Max von Hippel, Ben Weintraub and Samuel Jero. I would also like to thank the rest of the members of Purdue NLP for their continuous support and feedback.

Another special thanks goes to Professors H.E. Dunsmore and Sunil Prabhakar for giving me the opportunity to serve as an instructor and head TA early in my graduate studies. This experience, together with their valuable mentorship, made me a better, more invested educator. I would also like to thank my collaborators and mentors at Microsoft: Bahar Sarrafzadeh, Sujay Jauhar, Julia Kiseleva and Brent Hecht, their guidance has been invaluable in the last stages of my PhD. I would again like to thank Professors Dan Goldwasser, Cristina Nita-Rotaru, Sunil Prabhakar, Ming Yin and Brent Hecht for their support and guidance during the academic job search.

I am very fortunate to have a loving family and friends who have encouraged me throughout the highs and lows of the PhD experience. I am very thankful to my parents, Rosa Amelia

Gonzalez and Luis Emilio Pacheco, my sister, Carolina Pacheco, and my friends, especially Glebys Gonzalez, Aldo Porco and Maru Cabrera. You all made this experience all that much more worthwhile. Lastly, I am forever grateful to my wonderful spouse, Chelsy Albertson, for all of her love, endless support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Language “in the wild” is complex and ambiguous and relies on a shared understanding of the world for its interpretation. Most current natural language processing methods represent language by learning word co-occurrence patterns from massive amounts of linguistic data. This representation can be very powerful, but it is insufficient to capture the meaning behind written and spoken communication.

In this dissertation, I will motivate neural-symbolic representations for dealing with these challenges. On the one hand, symbols have inherent explanatory power, and they can help us express contextual knowledge and enforce consistency across different decisions. On the other hand, neural networks allow us to learn expressive distributed representations and make sense of large amounts of linguistic data. I will introduce a holistic framework that covers all stages of the neural-symbolic pipeline: modeling, learning, inference, and its application for diverse discourse scenarios, such as analyzing online discussions, mining argumentative structures, and understanding public discourse at scale. I will show the advantages of neural-symbolic representations with respect to end-to-end neural approaches and traditional statistical relational learning methods.

In addition to this, I will demonstrate the advantages of neural-symbolic representations for learning in low-supervision settings, as well as their capabilities to decompose and explain high-level decision. Lastly, I will explore interactive protocols to help human experts in making sense of large repositories of textual data, and leverage neural-symbolic representations as the interface to inject expert human knowledge in the process of partitioning, classifying and organizing large language resources.

1. INTRODUCTION

Language "in the wild" is complex and ambiguous, and relies on a shared understanding of the world for its interpretation. Namely, a lot of the context needed to convey meaning is not explicit in the language used. As an example, we can look at two statements made by political opponents on the topic of immigration to the United States (Fig. 1.1). These two statements express opinions on the same topic, and use very similar wording to communicate very different ideas. Knowledge of what liberals and conservatives value is key to grasp these contrasting views.

Most current natural language processing methods represent language by learning word co-occurrence patterns from massive amounts of linguistic data. This representation can be very powerful, but it lacks the mechanism to represent real-world context. To address this challenge, we need to find a way to model the concepts and abstractions that allow us to characterize the information expressed in the text. For example, we could analyze the meaning of political statements by explicitly modeling the sentiments, beliefs and world views of their authors. In Figure 1.1, Argument 1 is emphasizing *fairness* towards *asylum seekers*, which is known to be a *liberal* talking point, while Argument 2 is emphasizing *fairness* towards *legal immigrants* within the US immigration system, which is known to be *conservative* talking point. Explicitly modeling and reasoning over these concepts lets us disambiguate the specific language used to express opinions, and allows us to create a model of the world that explains public discourse and human interactions.

Argument 1

If people are trying to flee the dangerous countries they are from, it is unjust to subject them to a grueling, long and demanding process to stay in the US.

Argument 2

Many legal immigrants to the US went through long and demanding procedures in order to gain their status. It is unjust to allow others to circumvent these rules.

Figure 1.1. Statements made by political opponents in the context of immigration to the US

Reasoning about abstract concepts requires making predictions over multiple, often interdependent, variables. Traditionally, this is done in one of the following ways: (1) *Structured learning*, and more broadly, *statistical relational learning*, in which symbolic abstractions are used to describe the relational properties of the domain and probabilistic graphical models are used to reason under uncertainty; and (2) *End-to-end deep learning* techniques, where the complex input is mapped to outputs directly using complex neural architectures, without explicitly decomposing the decision process into parts. In this case, dependencies are represented in a latent high-dimensional space. The two approaches model relational data in very different ways, resulting in models with complementary properties. Symbolic models are interpretable, and allow domain experts to directly inject their knowledge and constrain the learning problem. Neural models capture dependencies using the network architecture and are better equipped to deal with noisy data, such as text. However, they are often difficult to interpret and constrain according to domain knowledge.

This dissertation is motivated by the opportunities of combining the flexibility afforded by neural methods to identify patterns in large-scale language data, with a principled way to reason over higher-level patterns using a **combined neural-symbolic representation**. On the one hand, symbols have inherent explanatory power, and they can help us express domain knowledge and enforce consistency across different decisions. When analyzing public discourse, we could use symbols to explicitly represent conceptual frameworks studied in the social sciences, such as *social homophily theory* and *moral foundation theory*, and enforce consistency between them based on our understanding of the world. On the other hand, neural models allow us to learn expressive distributed representations that generalize across different textual inputs, helping us make sense of large amounts of linguistic data. Moreover, neural models can provide us with a shared high-dimensional space to ground abstract concepts in language and align different modalities.

Several frameworks have been proposed in the broader AI literature to take advantage of these complimentary strengths by combining particular properties of symbolic and neural approaches [1]–[4]. However, most of these frameworks were designed for classical relational learning tasks, and their applicability to natural language scenarios can be limited. Further, most recent neural-symbolic work coming from the NLP community focuses on applications such as word math problems, visual reasoning and question-answering. These domains are a natural fit for neural-symbolic representations, as they usually deal with symbolic inputs and outputs (e.g. knowledge graphs and mathematical symbols), or concepts that are very concrete (e.g. shapes and colors). While unexplored, discourse analysis is an excellent fit for neural-symbolic representations. On the one hand, discourse has inherent structure. Neural-symbolic representations allow us to explicitly model the interactions between participants in a conversation, or between characters in a story, using typed relations. On the other hand, neural-symbolic representations give us a common language to reason across modalities. For example, we could exploit the principle of social homophily, stating that people with strong social ties are likely to hold similar views, to reason about statements made by two individuals. Finally, we could use symbols to model high-level conceptual frameworks that can support our analysis, and learn expressive representations for them.

In this dissertation, we look at the challenge of combining these two modeling paradigms for discourse-level natural language tasks. We focus on four key challenges and opportunities in the space of neural-symbolic discourse analysis, and cover all stages of the pipeline: modeling, learning, inference and real-world applications.

1.1 Declarative Modeling

Statistical approaches for natural language processing typically rely on an iterative process of collecting and annotating data, engineering features, specifying predictive models and analyzing errors. Knowledge is communicated as a set of labeled input-output pairs, and the focus is placed either on coming up with an informative set of features, or in devising learning models that capture the information in a latent high dimensional space. Alternatively, high-level modeling allows us to decompose the decision into smaller parts, and express knowledge in a structured way. Declarative modeling in particular, allows us to shift our focus to what we want to achieve, rather than how to achieve it. This is a key advantage when collaborating with people outside of machine learning and NLP, as it gives them an interface to contribute their domain expertise, and focus on the aspects, variables and interactions that they want to model.

In this dissertation, we present DRaiL: a declarative modeling framework that uses a combined neural-symbolic representation for modeling the interaction between multiple decisions in structured and relational domains. Unlike end-to-end neural networks that work directly over feature vectors, users can explicitly model high-level concepts by defining a set of relevant entities and relations. Then, dependencies between different aspects can be expressed using first-order logic rules. DRaiL's language allows us to quickly prototype relational models in a principled way, and study the interaction between representation, inference and learning. The neural component of DRaiL embeds entities and relations in a shared distributed space, allowing us to learn representations that are relation specific (e.g. in Example 1, Arg. 1 and Arg. 2 can be similar with respect to their moral foundation -*fairness*-, but different with respect to their ideological messaging). The symbolic component of DRaiL allows us to express consistency and constraints over different decisions. In turn, this can be exploited to align representations across different modalities and to combine multiple sources of indirect supervision. In this dissertation, we use DRaiL to model various language domains, including structured debates, argumentative essays, conversations and public opinions.

1.2 Algorithmic Frameworks and Learning Protocols

There are three main characteristics that contribute to successful language technologies: their generalizability, their efficiency, and their transparency. Different representations and learning paradigms have different strengths, and in most cases, improving one aspect comes at the expense of another. For instance, neural approaches offer high generalizability, but require a lot of resources, and are hard to interpret. Rule-based systems are transparent, but struggle to generalize. Probabilistic graphical models are easier to interpret, but require solving computationally intractable constrained optimization problems. Building on the thesis that neural-symbolic approaches offer the right balance to counteract this trade-off, this dissertation introduces algorithmic frameworks and learning protocols that emphasize these three aspects.

1.2.1 Efficient Neural-Symbolic Methods

To learn parameters in DRaiL, we propose a deep structured prediction framework that combines expressive textual encoders, relational embeddings and constrained inference. The objective functions used in DRaiL involve solving or approximating the MAP inference problem. While tractable solutions exist for traditional NLP tagging tasks, dealing with more complicated structures and arbitrary declarative constraints comes at a high computational cost. In this dissertation, we explore the use of randomized inference for deep structured models composed of expressive neural encoders, where theoretical guarantees are weak or nonexistent. We obtained competitive results at a fraction of the cost for a set of tasks involving complicated discourse structures.

1.2.2 Learning with Explanations

Identifying the reasoning steps taken to arrive to a decision is important to understand and judge the quality of predictive models. In a lot of cases the specific properties that explain our decisions are not readily available to us. This is especially the case in end-to-end deep learning models, where numerous, complex computations obfuscate the way in which the model reasons. Neural-symbolic representations provide us with a way to explicitly model emerging properties in a given domain. In cases where we do not have explicit supervision, we can use background knowledge to encode reasonable behaviors for these properties, and rely on sources of weak supervision to initialize them. By making properties explicit, we can also exploit human judgments to obtain and incorporate feedback.

In this dissertation, we explore the use of symbolic explanations to model two challenging discourse scenarios: identifying collaborative conversations and analyzing opinions about the COVID-19 vaccine. To identify collaborative conversations, we use discrete latent variables to explicitly model behaviors that indicative of successful collaborations, such as "idea development", "balanced contributions" and "engagement", and scored them using expressive language encoders. These behaviors are then connected to observed information using a symbolic reasoning framework implemented in DRaiL. This solution resulted in performance improvements, while providing a natural way to explain the final decision.

To analyze opinions about the COVID-19 vaccine, we build on Moral Foundations Theory [5], a theoretical framework for analyzing expressions of moral values by introducing five themes that are observed across cultures (e.g. care/harm, fairness/cheating). Moral Foundation Theory has been repeatedly used to explain user behaviors. In previous work, my collaborators and I made the observation that when divergent groups of people use the same moral foundation, their moral sentiment is directed at different targets. In Example 1, both arguments are using the *fairness* MF, but the targets vary (*refugees* vs. *legal immigrants*). To capture this, we introduced morality frames, a symbolic structure that makes the moral roles of entities explicit [6]. While this is a rich framework to analyze opinions about the COVID-19 vaccine, obtaining fine-grained, high-quality annotations is costly. To tackle this challenge, we leverage DRaiL to decompose the decision and explicitly encourage consistency between people with similar views (e.g. Two US citizens that are against the vaccine are likely to have negative views about Dr. Fauci). We model all decisions as latent, and leverage both declarative knowledge and distant sources of supervisions to learn in these settings.

1.2.3 Discovering and Grounding Explanations with Humans in the Loop

Experts across diverse academic and professional disciplines struggle with making sense of large amounts of linguistic data. However, uncovering latent themes from text automatically remains an open challenge in natural language processing. Traditionally, researchers and practitioners approach this challenge using noisy unsupervised techniques such as topic models [7] and clustering algorithms [8], or by manually identifying the relevant themes and annotating them in text [9]. In this dissertation, we combine computational and qualitative techniques to address this challenge. We focus our analysis on opinions about the COVID-19 vaccine, and propose interactive, humans-in-the-loop protocols to identify repeating themes in opinions about the vaccine, as well as to ground the identified themes in a large set of unlabeled opinions.

Human-in-the-loop approaches amplify the role of human experts in the process of learning and refining machine learning models. Most current human-in-the-loop protocols work directly over the space of inputs and their labels, for instance, by soliciting examples from people to augment the training data, or by disabling specific input features. While straightforward, this low-level representation does not take advantage of people's abilities to model concepts and higher-level abstractions. Neural-symbolic representations offer great opportunities to encode human expertise beyond labeling. This is advantageous for two reasons: 1) labeling is tedious, repetitive work, and previous studies have shown that uses prefer richer interaction protocols, and 2) higher-level abstractions have the potential to generalize to more scenarios, having a stronger impact in the model performance than adding or modifying a handful of training examples. In this dissertation, we leverage DRaiL to encode templatic knowledge coming from a small set of judgments provided by human experts, and use them to improve the grounding of the discovered themes in a large dataset of opinions about the COVID-19 vaccine.

1.3 Dissertation Organization, Collaborations and Publications

This dissertation is composed of work done in collaboration with other researchers, some of which has been previously published. The overall organization is as follows:

- Chapter 1 motivated neural-symbolic representations for natural language discourse scenarios, and has outlined the main contributions of this dissertation.

- Chapter 2 presents the relevant background and related work. A significant part of the literature review presented in Chapter 3 has appeared previously in all of our relevant publications [6], [10]–[12].

- Chapter 3 presents DRaiL, a general-purpose declarative neural-symbolic framework designed to deal with diverse natural language processing scenarios. DRaiL is the main contribution of this dissertation and it is used in all subsequent chapters. The work presented in this chapter was done in collaboration with Dr. Dan Goldwasser, and it was published in TACL 2021 [10].

- Chapter 4 presents a randomized deep structured prediction approach to learn DRaiL models efficiently. The work presented in this chapter was done in collaboration with

Manuel Widmoser, Dr. Jean Honorio and Dr. Dan Goldwasser, and it was published in EACL 2021 [11].

- Chapter 5 proposes two methodologies for learning DRaiL models with latent variables. The discussion is centered around two modeling scenarios. In the first scenario we have a high-level task and we wish to learn intermediate discrete explanations to support the decision. In the second scenario we have a complex problem composed of multiple interdependent decisions, for which we have no direct supervision. The work presented in the first scenario was done in collaboration with Ayush Jain, Steve Lancette, Dr. Mahak Goindani and Dr. Dan Goldwasser, and was published in SIGDIAL 2020 [12]. Part of this work was also part of Ayush Jain's Masters Thesis [13]. The work presented in the second scenario was done in collaboration with Tunazzina Islam, Monal Mahajan, Andrey Shor, Dr. Ming Yin, Dr. Lyle Ungar and Dr. Dan Goldwasser, and it will be published in NAACL 2022 [14].

- Chapter 6 proposes an interactive protocol to discover and ground latent variables in large language resources. The discussion is centered around two main challenges. In the first challenge we are concerned with learning to ground a set of named, hypothesized latent variables. In the second challenge we are concerned with discovering the space of relevant latent variables. In both cases, we leverage DRaiL to connect latent variables with other observed and predicted variables. The work presented in the first scenario was done in collaboration with Tunazzina Islam, Monal Mahajan, Andrey Shor, Dr. Ming Yin, Dr. Lyle Ungar and Dr. Dan Goldwasser, and it will be published in NAACL 2022 [14]. The second scenario presents exploratory work and has not yet been published.

2. RELEVANT BACKGROUND

This chapter summarizes the relevant background and related work, both from the methodological perspective and from the applications perspective. In Section 2.1 we will cover the relevant learning and representation frameworks. In Section 2.2 we will cover interactive approaches to NLP. Lastly, in Section 2.3 we will cover previous work on the discourse-level applications explored in dissertation, including opinion analysis, argumentation mining, and conversational analysis. In all cases, we will survey the related literature and highlight the connections to our work.

2.1 Relational Learning and Structured Prediction

This section surveys previous work on relational learning, structured prediction, and the integration of deep learning techniques with symbolic representations.

2.1.1 High-Level Languages for Graphical Models

Several high level languages for specifying graphical models have been suggested. BLOG [15] and CHURCH [16] were suggested for generative models. For discriminative models, we have Markov Logic Networks (MLNs) [17] and Probabilistic Soft Logic (PSL) [18]. Both PSL and MLNs combine logic and probabilistic graphical models in a single representation, where each formula is associated with a weight, and the probability distribution over possible assignments is derived from the weights of the formulas that are satisfied by such assignments. Like DRaiL, PSL uses formulas in clausal form (specifically collections of horn clauses). The main difference between DRaiL and these languages is that, in addition to graphical models, it uses distributed knowledge representations to represent dependencies. Other discriminative methods include FACTORIE [19], an imperative language to define factor graphs, Constrained Conditional Models (CCMs) [20], [21] an interface to enhance linear classifiers with declarative constraints, and ProPPR [22] a probabilistic logic for large databases that approximates local groundings using a variant of personalized PageRank.

2.1.2 Node Embedding and Graph Neural Networks

A recent alternative to graphical models is to use neural nets to represent and learn over relational data, represented as a graph. Similar to DRaiL's RelNets, the learned node representation can be trained by several different prediction tasks. However, unlike DRaiL, these methods do not use probabilistic inference to ensure consistency.

Node embeddings approaches [23]–[27] learn a feature representation for nodes capturing graph adjacency information, such that the similarity in the embedding space of any two nodes is proportional to their graph distance and overlap in neighboring nodes. Some frameworks [25], [27], [28] allow nodes to have textual properties, which provide an initial feature representation when learning to represent the graph relations. When dealing with multi-relational data, such as knowledge graphs, both the nodes and the edge types are embedded [29]–[32]. Finally, these methods learn to represent nodes and relations based on pair-wise node relations, without representing the broader graph context in which they appear. Graph neural nets [33]–[35] create contextualized node representations by recursively aggregating neighboring nodes.

2.1.3 Hybrid Neural-Symbolic Approaches

Several recent systems explore ways to combine neural and symbolic representations in a unified way. We group them into five categories.

Lifted Rules to Specify Compositional Networks

These systems use an end-to-end approach and learn relational dependencies in a latent space. Lifted Relational Neural Networks (LRNNs) [36] and RelNNs [2] are two examples. These systems map observed ground atoms, facts and rules to specific neurons in a network and define composition functions directly over them. While they provide for a modular abstraction of the relational inputs, they assume all inputs are symbolic and do not leverage expressive encoders.

Differentiable Inference

These systems identify classes of logical queries that can be compiled into differentiable functions in a neural network infrastructure. In this space we have Tensor Logic Networks (TLNs) [37] and TensorLog [4]. Symbols are represented as row vectors in a parameter matrix. The focus is on implementing reasoning using a series of numeric functions.

Rule Induction from Data

These systems are designed for inducing rules from symbolic knowledge bases, which is not in the scope of our framework. In this space we find Neural Theorem Provers (NTPs) [3], Neural Logic Programming [38], DRUM [39] and Neural Logic Machines (NLMs) [40]. NTPs use a declarative interface to specify rules that add inductive bias and perform soft proofs. The other approaches work directly over the database.

Deep Classifiers and Probabilistic Inference

These systems propose ways to integrate probabilistic inference and neural networks for diverse learning scenarios. DeepProbLog [1] extends the probabilistic logic programming language ProbLog to handle neural predicates. They are able to learn probabilities for atomic expressions using neural networks. The input data consists of a combination of feature vectors for the neural predicates, together with other probabilistic facts and clauses in the logic program. Targets are only given at the output side of the probabilistic reasoner, allowing them to learn each example with respect to a single query. On the other hand, Deep Probabilistic Logic (DPL) [41] combines neural networks with probabilistic logic for indirect supervision. They learn classifiers using neural networks and use probabilistic logic to introduce distant supervision and labeling functions. Each rule is regarded as a latent variable, and the logic defines a joint probability distribution over all labeling decisions. Then, the rule weights and the network parameters are learned jointly using variational EM. In contrast, DRaiL focuses on learning multiple interdependent decisions from data. Lastly, Deep Logic Models (DLMs) [42] learn a set of parameters to encode atoms in a probabilistic logic program. Similarly to Tensor Logic Networks [37] and TensorLog [4], they use differentiable inference, allowing the model to be trained end-to-end. DLMs can work with diverse neural architectures and backpropagate back to the base classifiers. The main difference between DLMs and DRaiL is that DRaiL ensures representation consistency of entities and relations across all learning tasks by employing RelNets.

Deep Structured Prediction for NLP

More generally, deep structured prediction approaches have been successfully applied to various NLP tasks such as named entity recognition and dependency parsing [43]–[48]. When the need arises to go beyond sentence-level, some approaches combine the output scores of independently trained classifiers using inference [49]–[53], while others implement joint learning for their specific domains [54], [55]. Our main differentiating factor is that we provide a general interface that leverages FOL clauses to specify factor graphs and express constraints.

To summarize these differences, we outline a feature matrix in Tab. 2.1. Given our focus in NLP tasks, we require a neural-symbolic system that (1) allows us to integrate state-of-the-art text encoders and NLP tools, (2) supports structured prediction across long texts, (3) lets us combine several modalities and their representations (e.g. social and textual information) and (4) results in an explainable model where domain constraints can be easily introduced.

2.1.4 Approximate Inference

Randomized approximation has been introduced as an alternative to exact inference. Zhang, Lei, Barzilay, *et al.*, 2014 [56] suggest a simple randomized greedy inference algorithm and empirically demonstrate its effectiveness for dependency parsing and other traditional NLP tasks [57]. The theoretical results in Honorio and Jaakkola, 2016 [58], based on the probably approximately correct Bayes framework, characterize these findings by providing generalization bounds. More recently, Ma, Chowdhury, Deshwal, *et al.*, 2019 [59] extended the work introduced by Zhang, Lei, Barzilay, *et al.*, 2014 [56] and Zhang, Li, Barzilay, *et*

Table 2.1. Comparing Systems

	Symbolic Features					Neural Features					
System	Symbolic Inputs	Raw Inputs	Decla-rative	Prob/Logic Inference	Rule Induction	Embed. Symbols	End-to-end Neural	Backprop. to Encoders	Architecture Agnostic	Multi-Task Learning	**Open Source**
MLN	✓		✓	✓							✓
FACTORIE	✓			✓							✓
CCM	✓	✓	✓	✓							✓
PSL	✓		✓	✓							✓
LRNNs	✓					✓	✓				
RelNNs	✓					✓	✓				✓
LTNs	✓		✓	✓		✓	✓				✓
TensorLog	✓		✓	✓		✓	✓				✓
NTPs	✓		✓	✓	✓	✓	✓				✓
Neural LP	✓				✓	✓	✓				✓
DRUM	✓				✓	✓	✓				✓
NLMs	✓				✓	✓	✓				✓
DeepProbLog	✓	✓	✓	✓	✓	✓		✓	✓		✓
DPL	✓	✓	✓	✓		✓		✓	✓		✓
DLMs	✓	✓	✓	✓		✓	✓	✓	✓		
DRaiL	✓	✓	✓	✓		✓		✓	✓	✓	✓

al., 2015 [57] to structured prediction tasks with large structured outputs by leveraging local classifiers to find good starting solutions and improve the accuracy of search. All of these methods were evaluated on linear structured models. In contrast, we explore randomized approaches for the non-linear case.

2.1.5 Latent Variable Models

There is an extensive body of work on learning discriminative models with discrete latent variables. Hidden-Unit Conditional Random Fields [60] extend standard linear conditional random fields by adding a binary latent variable that mediates the interaction between each observation and target variable on the chain. This allows the latent variables to be marginalized out during inference and learning, but cannot express more complex dependencies.

More expressive models include Multiple Relational Clusterings (MRC), a general framework for statistical predicate invention (SPI) [61]. The authors define SPI as the problem of discovering new concepts, properties and relations in structured data. Their model is based on Markov Logic [17], and new predicates are expressed in terms of the observable ones, using statistical techniques to guide the process and explicitly representing the uncertainty in the discovered predicates. MRC automatically invents predicates by clustering objects, attributes and relations.

Probabilistic Soft Logic (PSL) was also extended to learn with latent variables [62], [63]. Atoms in PSL can represent conditioning variables, target variables or hidden variables. Hinge-Loss Markov Random Fields (HL-MRF) are the formalism behind PSL. HL-MRFs are a class of undirected probabilistic graphical models for which inference is reduced to a convex optimization problem. HL-MRF with latent variables can be learned using hard EM, or more efficiently, with paired-dual learning.

Some recent efforts have looked into combining neural networks and latent variables for specific scenarios, such as incorporating discrete latent variables into recurrent neural networks for jointly modeling sequences of words and latent discourse relations between adjacent sentences [64], or learning sequential Markov models where discrete latent variables are sampled from the hidden states of a small LSTM [65], [66]. By relying on the DRaiL infrastructure, we propose a general neural-symbolic latent variable framework that can be applied to a wider range of problems and that is independent of the neural architecture used to represent the text.

2.2 Interactive NLP Frameworks

In this Section, we will survey the frameworks and approaches that have been suggested to interact with NLP systems. First, we will look at approaches that incorporate human feedback to build and debug NLP models. Then, we will focus on declarative and interactive approaches to incorporate sources of distant and weak supervision.

2.2.1 Building and Debugging NLP Models using Human Feedback

There has been increased interest in the XAI and NLP communities in leveraging human feedback to build and debug NLP models. Lertvittayakumjorn and Toni, 2021 [67] define explanation-based human debugging as the process of fixing or mitigating bugs in a trained model using human feedback given in response to explanations for the model. In this space, Ribeiro, Singh, and Guestrin, 2016 [68] proposed LIME, an explanation framework to explain complex classifiers. LIME learns interpretable classifiers on specific sections of the data such that the decision boundary matches the one learned by the complex model. In the case of textual data, they mapped complex representations to a unigram model, and showed that they can use human feedback to remove spurious explanations and improve model performance. More recently, Lertvittayakumjorn, Specia, and Toni, 2020 [69] proposed FIND, a debugging protocol that exploits exploits layer-wise relevance propagation [70] to obtain instance-level explanations for CNN classifiers. These explanations are aggregated using word clouds and presented to users, who can then deactivate the ones that do not make sense. The main difference between these two approaches is that FIND collects feedback on the model, and not the individual predictions.

While the NLP community is increasingly focused on large language models to represent text, there is not a lot of work on interactively debugging these models. The main difficulty comes from their complexity, as large LMs have numerous attention heads to compute contextual relations between words and subwords, feature attribution methods are insufficient to illustrate how the model reasons. To circumvent this issue, Zylberajch, Lertvittayakumjorn, and Toni, 2021 [71] proposed HILDF, a human-in-the-loop debugging framework that makes use of influence functions. Instead of focusing on specific parts of the input, influence functions help identify influential examples in the training set. These examples are then presented to uses for validation, and if selected, are used to augment the training set. Yao, Chen, Ye, *et al.*, 2021 [72] propose an alternative approach to deal with LM complexities, and solicit complex and compositional explanations from humans. In this framework, users are presented with instance-level explanations of the model predictions using heat-maps, then they are asked to explain any spurious patterns found and to suggest an adjustment of the importance scores. Users are prompted to use very specific templates to provide their explanations, so that they can be translated into executable first-order logic rules and used to find similar examples in the dataset. All templates are limited to providing explanations about specific input features.

In contrast to previous work, we do not focus solely on deactivating or updating model explanations, as we allow users to introduce new information that can help the model generalize better. Similarly to Yao, Chen, Ye, *et al.*, 2021 [72], we focus on *why* should examples be labeled a certain way, rather than on *what* is the correct label. However, we have more flexibility in the form of the feedback, allowing users to teach the model to reason beyond specific input features.

2.2.2 Declarative and Interactive Frameworks for Indirect Supervision

Our work is also related to declarative and interactive frameworks that support indirect supervision. Data programming [73] is paradigm to aggregate multiple weak label sources and generating probabilistic training labels for them. Snorkel [74] is an end-to-end system to apply the data programming paradigm. It combines different weak supervision sources to label training examples by providing a general language to interactively create labeling functions. Then, it learns a generative model over the labeling functions to estimate their accuracies and correlations without any supervision by relying on agreements and disagreements between the different labeling functions. This process results in a set of probabilistic labels that can be used to train discriminative machine learning models. Labeling functions, as defined in Snorkel, could be used as a way to instantiate predicates in DRaiL. In fact, we can think about our concept grounding procedure as an example of a labeling function. However, Snorkel is a framework for generating probabilistic labels for a single decision, that can then be used to learn machine learning models. In contrast, DRaiL -as well as other statistical relational learning frameworks- is designed to learn the parameters of a probabilistic graphical model. In these frameworks, we can model multiple inter-dependent decisions and constraints.

More closely related to our work, we have Deep Probabilistic Logic [41], a general framework that combines probabilistic logic with deep learning for indirect supervision. DPL leverages probabilistic logic as a language to integrate different sources of weak supervision and resolving noisy and contradictory information. Label decisions are modeled as latent variables and scored jointly using a probabilistic inference procedure, the resulting label is used to train a deep learning model in a supervised manner. DPL uses an EM framework to alternate between inference and learning. Like DRaiL, DPL has the advantage that it supports arbitrary high-order soft and hard constraints that capture the inter-dependencies among multiple instances. The main difference between DRaiL and DPL, is that DPL is designed to support one end-task learned with one neural network. In contrast, DRaiL is designed to combine multiple different decisions in a supervised manner, each tied to its own neural scoring function. Moreover, DRaiL uses a shared relational representation that learns entity-specific and relation-specific embeddings. Nevertheless, it easy to see how DPL could be expanded to support multiple decisions and neural nets. Similarly, DRaiL can be expanded to support latent variables (See Chapter 5).

2.3 Discourse Analysis

In this Section, we will cover previous work on the discourse-level applications explored in this dissertation. We will survey work on opinion analysis and argumentation mining, as well as work on general conversational analysis.

2.3.1 Opinion Analysis and Argumentation Mining

Identifying stances and arguments supporting them is a central challenge of argumentation mining [75], [76]. Stab and Gurevych, 2017 [77] introduce the task of extracting tree structures from argumentative essays that summarize the relations between claims and premises expressed in the text. They approach this task using an exhaustive set of hand-crafted features, linear local classifiers and Integer Linear Programming at test time. Niculae, Park, and Cardie, 2017 [54] tackle the same task by jointly learning to score multiple decisions while enforcing domain constraints. They explore structured SVMs and RNNs, using the approximate AD^3 inference algorithm [78].

A related line of work deals with predicting user stances in online debates. Some approaches model the problem as a text classification task [79], [80], while other approaches take a collective approach to model user behavior and interactions [50], [81]–[83]. In most cases, the task is approached by learning target-specific classifiers, trained and tested for each topic of interest.

More recently, other approaches have employed cross-target classification, where the classifier is adapted from different but closely related targets. Xu, Paris, Nepal, *et al.*, 2018 [84] manually identify pairs of domain-related sources and targets and use a neural model based on self-attention. Augenstein, Rocktäschel, Vlachos, *et al.*, 2016 [85] learn a target-dependent text representation using a bidirectional conditional encoding. Both of these approaches were developed for short spans of text, specifically tweets, and deal with a very reduced number of topics.

Li, Porco, and Goldwasser, 2018 [83] formulate the debate stance prediction task as a structured representation learning problem. They exploit the connections between text, users and stances to create a common representation for them. This way, they allow their model to share information between the representations of different debate topics.

The Fake News Challenge Stage 1 (FNC-1) shared task addressed a stance classification task [86]. The goal of the challenge is to determine the perspective (or stance) of a news article relative to a given headline. An article's stance can either agree or disagree with the headline, discuss the same topic, or be completely unrelated. This challenge can be regarded as a more general version of the cross-target stance prediction setup, while also dealing with longer texts. However, there is no debate structure or known dependencies between articles. The best reported results for the FNC-1 task were obtained using a stacked LSTM over word sequences of headline-document pairs, enriched with lexical features [87]. Commensurate results were obtained with a simpler feature-based approach and a feed-forward neural network [88]. Several recent works looked at modeling the persuasive strength of arguments and argument flow in debates [89]–[91], and more relevant to the scenarios that we explore, based on user attributes [92].

Morality and Framing Analysis

Usage of sociological theories like the Moral Foundation Theory (MFT) [5], [93] and Framing [94]–[96] in Natural Language Processing tasks has gained significant interest. The MFT has been widely used to study the effect of moral values on human behavioral patterns, such as charitable donations [97], violent protests [98] and social homophily [99]. Framing is a strategy used to bias the discussion on an issue towards a specific stance by emphasizing certain aspects that prime the reader to support the stance. Framing is used to study the political bias and polarization in social and news media [100]–[106]. MFT is frequently used to analyze political framing and agenda setting. For example, Fulgoni, Carpenter, Ungar, *et al.*, 2016 [107] analyzed framing in partisan news sources using MFT, and Dehghani, Sagae, Sachdeva, *et al.*, 2014 [108] studied the difference in moral sentiment usage between liberals and conservatives. In a related line of work, Brady, Wills, Jost, *et al.*, 2017 [109] found that moral/emotional political messages are diffused at higher rates on social media.

There is a body of previous work contributing to the detection of moral sentiments. Johnson and Goldwasser, 2018 [110] showed that policy frames [96] help in moral foundation prediction. Hoover, Portillo-Wightman, Yeh, *et al.*, 2020 [111] proposed a dataset of 35k tweets annotated for moral foundations to support supervised learning. Lin, Hoover, Portillo-Wightman, *et al.*, 2018 [112] used background knowledge for moral sentiment prediction, Xie, Junior, Hirst, *et al.*, 2019 [113] proposed a text based framework to account for moral sentiment change, and Garten, Boghrati, Hoover, *et al.*, 2016 [114] used pretrained distributed representations of words to extend the Moral Foundations Dictionary [115] for detecting moral rhetoric.

While existing work studies Moral Foundation Theory at the issue and sentence level, Roy and Goldwasser, 2021 [116] showed that there is a correlation between entity mention and the sentence-level moral foundation in the tweets by the U.S. politicians. In the scenarios that we explore in this dissertation, we extend this work by studying MFT directly at the entity level. Hence, our work is broadly related to work on entity-centric affect analysis [117]–[119].

Analyzing Opinions about the COVID-19 Vaccine

Recent studies have noted the prevalence of rumors and misinformation in the context of the COVID-19 pandemic [120]–[123]. Following this trend, several computational approaches have been proposed to detect misinformation related to COVID in news outlets and social media [124]–[127]. In the scenarios that we explore, we take a different approach and look at the problem of identifying opinions surrounding the COVID-19 vaccine, and explicitly modeling the rationale and moral sentiment that motivates them.

Some recent work also look at analyzing arguments about COVID and vaccine hesitancy more broadly. In most cases, they either take a traditional classification approach for predicting stances [128], [129], or use topic modeling techniques to uncover trends in word usage [129]–[132]. In contrast, we propose a holistic framework that combines different methodological techniques, including human-in-the-loop mechanisms, classification with distant supervision, and deep relational learning to connect stance prediction, reason analysis and fine-grained entity moral sentiment analysis.

2.3.2 Conversational Analysis

Analyzing conversational data and identifying social and linguistic indicators for collaborative and anti-social interactions has been explored previously in several studies, including dispute identification [133], counseling conversations [134] and most relevant to the scenarios that we explore, identifying constructive conversations [135], [136]. In the applications that we present in this dissertation, we adapt the conversational data provided by Napoles, Tetreault, Rosata, *et al.*, 2017 [137] to accommodate a more restrictive definition of *good* conversation, focusing on collaborative behavior. Conversations that are polite and socially pleasant without much content are not considered as collaborative in our case. Also, conversations that do not include balanced engagement from all the participants or contain few off-topic, insulting and rude posts are not considered collaborative as well.

From a technical perspective these studies attempt to characterize desired and undesired conversational behaviors using lexical and discourse features. For example, Danescu-Niculescu-Mizil, Sudhof, Jurafsky, *et al.*, 2013 [138] make use of domain-independent lexical and syntactic features on Wikipedia edits to study the relationship between politeness and social power. Other work [139]–[142] focuses on the persuasive power of arguments made during the conversational interactions.

Our technical approach to conversational analysis is different, instead of directly building on the raw inputs, we formulate the decision over a set of latent variables designed to capture fine-grained behaviors. Reasoning over conversational interactions using latent variables was previously suggested by Chaturvedi, Goldwasser, and Daumé III, 2014 [143], for predicting instructors' intervention in MOOCs, our task aims to characterize the entire conversation, rather than the actions of a single participant. Our latent variable formulation is used to characterize the conversational style, engagement and information flow. Other work focused on similar analysis in the supervised settings. For example, discourse relations between posts in conversational threads [144], and agreement and disagreement in social media dialogs [145].

3. DRaiL: A DECLARATIVE DEEP RELATIONAL LEARNING FRAMEWORK FOR NLP

Understanding natural language interactions in realistic settings requires models that can deal with noisy textual inputs, reason about the dependencies between different textual elements and leverage the dependencies between textual content and the context from which it emerges. Work in linguistics and anthropology has defined context as a frame that surrounds a focal communicative event and provides resources for its interpretation [146], [147].

As a motivating example, consider the interactions in the debate network described in Fig. 3.1. Given a debate claim (t_1), and two consecutive posts debating it (p_1, p_2), we define a textual inference task, determining whether a pair of text elements hold the same stance in the debate (denoted using the relation $\texttt{Agree}(\texttt{X}, \texttt{Y})$). This task is similar to other textual inference tasks [148] which have been successfully approached using complex neural representations [149], [150]. In addition, we can leverage the dependencies between these decisions. For example, assuming that one post agrees with the debate claim ($\texttt{Agree}(\texttt{t}_1, \texttt{p}_2)$), and the other one does not ($\neg\texttt{Agree}(\texttt{t}_1, \texttt{p}_1)$), the disagreement between the two posts can be inferred: $\neg\texttt{Agree}(\texttt{t}_1, \texttt{p}_1) \wedge \texttt{Agree}(\texttt{t}_1, \texttt{p}_2) \rightarrow \neg\texttt{Agree}(\texttt{p}_1, \texttt{p}_2)$. Finally, we consider the *social context* of the text. The disagreement between the posts can reflect a difference in the

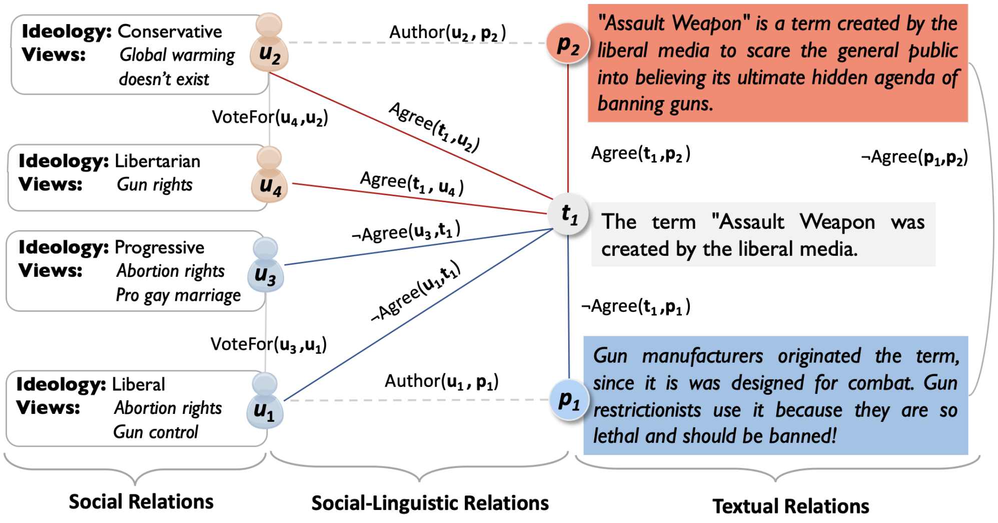

Ideology: Conservative
Views: Global warming doesn't exist
u_2
Author(u_2, p_2)
p_2
"Assault Weapon" is a term created by the liberal media to scare the general public into believing its ultimate hidden agenda of banning guns.
VoteFor(u_4,u_2)
Agree(t_1,u_2)
Agree(t_1,p_2)
¬Agree(p_1,p_2)
Ideology: Libertarian
Views: Gun rights
u_4
Agree(t_1, u_4)
t_1
The term "Assault Weapon was created by the liberal media.
Ideology: Progressive
Views: Abortion rights Pro gay marriage
u_3
¬Agree(u_3,t_1)
¬Agree(u_1,t_1)
¬Agree(t_1,p_1)
VoteFor(u_3,u_1)
Ideology: Liberal
Views: Abortion rights Gun control
u_1
Author(u_1, p_1)
p_1
Gun manufacturers originated the term, since it is was designed for combat. Gun restrictionists use it because they are so lethal and should be banned!
Social Relations
Social-Linguistic Relations
Textual Relations

Figure 3.1. Example Debate

perspectives their authors hold on the issue. While this information might not be directly observed, it can be inferred using the authors' social interactions and behavior. Given the principle of social homophily [151], stating that people with strong social ties are likely to hold similar views, authors' perspectives can be captured by representing their social interactions. Exploiting this information requires models that can align the social representation with the linguistic one.

Motivated by these challenges, we introduce DRaiL[1], a Deep Relational Learning framework, which uses a combined neural-symbolic representation for modeling the interaction between multiple decisions in relational domains. Similar to other neural-symbolic approaches [4], [152] our goal is to exploit the complementary strengths of the two modeling paradigms. Symbolic representations, used by logic-based systems and by probabilistic graphical models [17], [18], are interpretable, and allow domain experts to directly inject knowledge and constrain the learning problem. Neural models capture dependencies using the network architecture and are better equipped to deal with noisy data, such as text. However, they are often difficult to interpret and constrain according to domain knowledge.

Our main design goal in DRaiL is to provide a generalized tool, specifically designed for NLP tasks. Existing approaches designed for classic relational learning tasks [4], such as knowledge graph completion, are not equipped to deal with the complex linguistic input. While others are designed for very specific NLP settings such as word-based quantitative reasoning problems [1] or aligning images with text [152]. We discussed the differences between DRaiL and these approaches in Chapter 2. While the examples in this paper focus on modelings various argumentation mining tasks and their social and political context, the same principles can be applied to wide array of NLP tasks with different contextualizing information, such as images that appear next to the text, or prosody when analyzing transcribed speech, to name a few examples.

[1] ↑https://gitlab.com/purdueNlp/DRaiL/

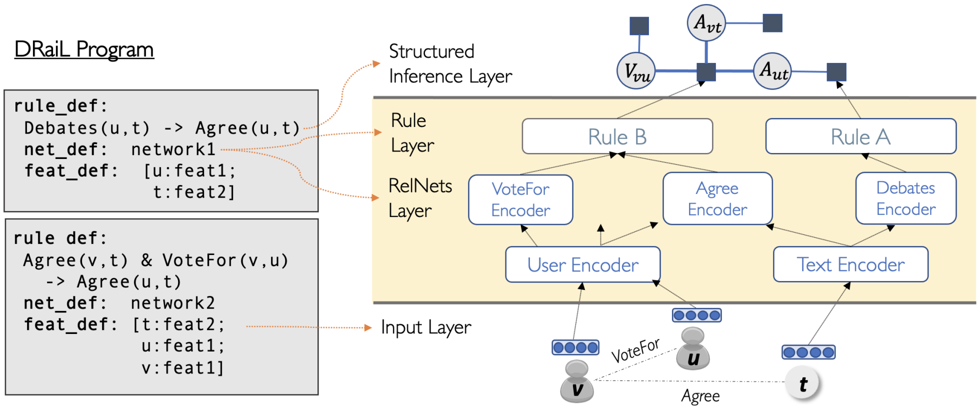
3.1 The DRaiL Framework

DRaiL uses a declarative language for defining deep relational models. Similar to other declarative languages [17], [18], it allows users to inject their knowledge by specifying dependencies between decisions using first-order logic rules, which are later compiled into a factor graph with neural potentials. In addition to probabilistic inference, DRaiL also models dependencies using a *distributed knowledge representation*, denoted RelNets, which provides a shared representation space for entities and their relations, trained using a relational multi-task learning approach. This provides a mechanism for explaining symbols, and aligning representations from different modalities. Following our running example, ideological standpoints, such as `Liberal` or `Conservative`, are discrete entities embedded in the same space as textual entities and social entities. These entities are initially associated with users, however using RelNets this information will propagate to texts reflecting these ideologies, by exploiting the relations that bridge social and linguistic information (see Fig. 3.1).

To demonstrate DRaiL's modeling approach, we introduce the task of *open-domain stance prediction with social context*, which combines social network analysis and textual inference over complex opinionated texts, as shown in Fig. 3.1. We complement our evaluation of DRaiL with two additional tasks, issue-specific stance prediction, where we identify the views expressed in debate forums with respect to a set of fixed issues [81], and argumentation mining [77], a document-level discourse analysis task.

DRaiL was designed for supporting complex NLP tasks. Problems can be broken down into domain-specific atomic components (which could be words, sentences, paragraphs or full documents, depending on the task), and dependencies between them, their properties and contextualizing information about them can be explicitly modeled.

In DRaiL dependencies can be modeled over the predicted output variables (similar to other probabilistic graphical models), as well as over the neural representation of the atoms and their relationships in a shared embedding space. This section explains the framework in detail. We begin with a high-level overview of DRaiL and the process of moving from a declarative definition to a predictive model.



```
rule_def:
 Debates(u,t) -> Agree(u,t)
 net_def:  network1
 feat_def:  [u:feat1;
             t:feat2]
```

```
rule def:
 Agree(v,t) & VoteFor(v,u)
   -> Agree(u,t)
 net_def:  network2
 feat_def: [t:feat2;
            u:feat1;
            v:feat1]
```

Figure 3.2. General Overview of DRaiL

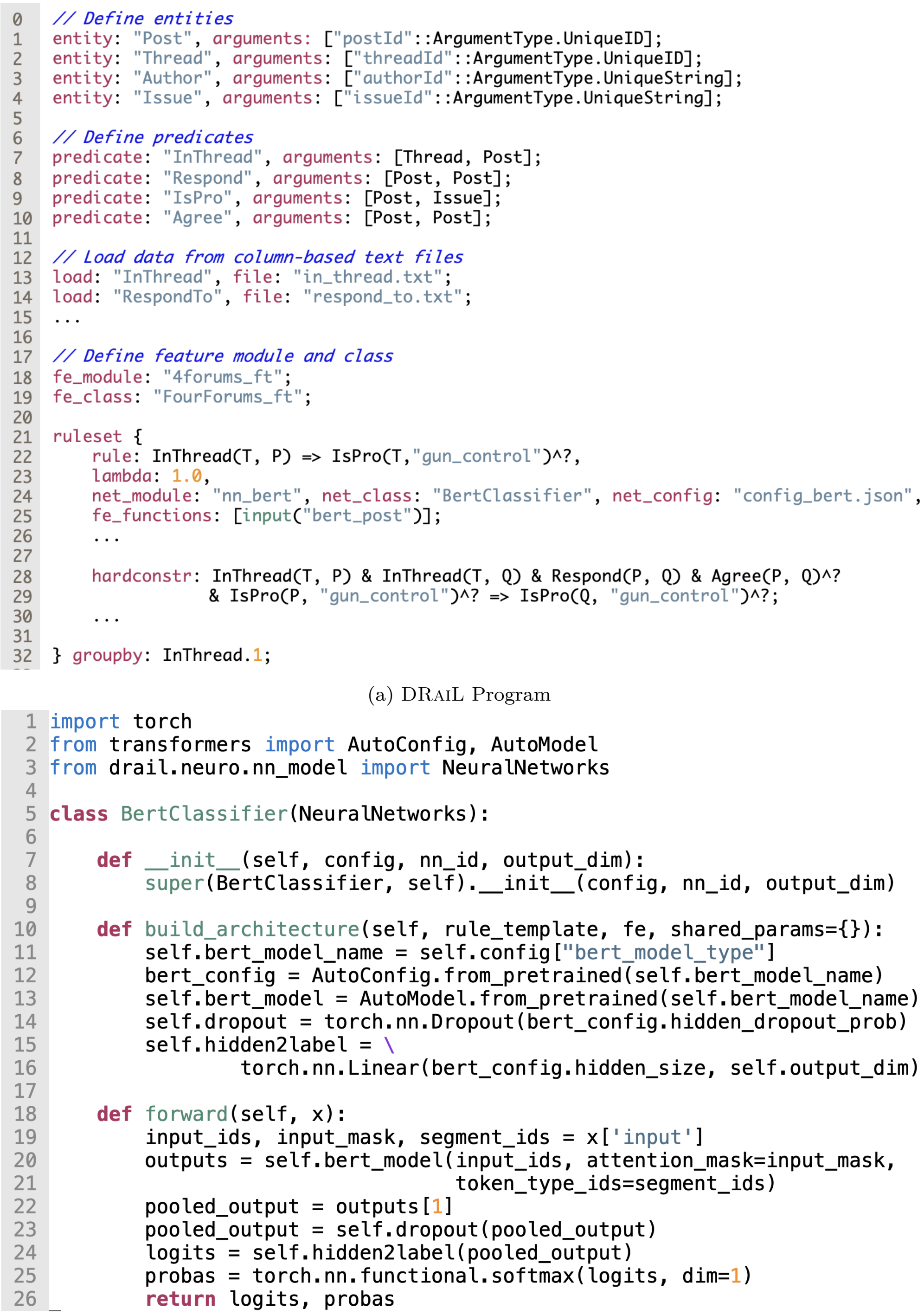
A DRaiL task is defined by specifying a finite set of *entities* and *relations*. Entities are either discrete symbols (e.g., POS tags, ideologies, specific issue stances), or attributed elements with complex internal information (e.g., documents, users). Decisions are defined using rule *templates*, formatted as horn clauses: $t_{LH} \Rightarrow t_{RH}$, where t_{LH} (*body*) is a conjunction of observed and predicted relations, and t_{RH} (*head*) is the output relation to be learned. Consider the debate prediction task in Fig. 3.1, it consists of several sub-tasks, involving textual inference ($\texttt{Agree}(\texttt{t}_1, \texttt{t}_2)$), social relations ($\texttt{VoteFor}(\texttt{u}, \texttt{v})$) and their combination ($\texttt{Agree}(\texttt{u}, \texttt{t})$). We illustrate how to specify the task as a DRaiL program in Fig. 3.2, by defining a subset of rule templates to predict these relations.

Each rule template is associated with a neural architecture and a feature function, mapping the initial observations to an input vector for each neural net. We use a shared *relational* embedding space, denoted RelNets, to represent entities and relations over them. As described in Fig. 3.2 ("RelNets Layer"), each entity and relation type is associated with an encoder, trained jointly across all prediction rules. This is a form of relational multi-task learning, as the same entities and relations are reused in multiple rules and their representation is updated accordingly. Each rule defines a neural net, learned over the relations defined on the body. They they take a composition of the vectors generated by the relations encoders as an input (Fig. 3.2, "Rule Layer"). DRaiL is architecture-agnostic, and neural modules for entities, relations and rules can be specified using PyTorch. Our experiments show that we can use different architectures for representing text, users, as well as for embedding discrete entities.

The relations in the horn clauses can correspond to hidden or observed information, and a specific input is defined by the instantiations -or *groundings*- of these elements. The collection of all rule groundings results in a factor graph representing our global decision, taking into account the consistency and dependencies between the rules. This way, the final assignments can be obtained by running an inference procedure. For example, the dependency between the views of users on the debate topic ($\texttt{Agree}(\texttt{u}, \texttt{t})$) and the agreement between them ($\texttt{VoteFor}(\texttt{u}, \texttt{v})$), is modeled as a factor graph in Fig. 3.2 ("Structured Inference Layer").

We include code snippets to show how to load data into DRAIL (Figure 3.3-a), as well as to how to define a neural architecture (Figure 3.3-b). Neural architectures and feature functions can be programmed by creating Python classes, and the module and classes can be directly specified in the DRAIL program (lines 13, 14, 24, and 29 in Figure 3.3-b).

We formalize the DRaiL language in Sec. 4.1. Then, in sections 3.1.2, 3.1.3 and 3.2, we describe the neural components and learning procedures.



```
// Define entities
entity: "Post", arguments: ["postId"::ArgumentType.UniqueID];
entity: "Thread", arguments: ["threadId"::ArgumentType.UniqueID];
entity: "Author", arguments: ["authorId"::ArgumentType.UniqueString];
entity: "Issue", arguments: ["issueId"::ArgumentType.UniqueString];

// Define predicates
predicate: "InThread", arguments: [Thread, Post];
predicate: "Respond", arguments: [Post, Post];
predicate: "IsPro", arguments: [Post, Issue];
predicate: "Agree", arguments: [Post, Post];

// Load data from column-based text files
load: "InThread", file: "in_thread.txt";
load: "RespondTo", file: "respond_to.txt";
...

// Define feature module and class
fe_module: "4forums_ft";
fe_class: "FourForums_ft";

ruleset {
    rule: InThread(T, P) => IsPro(T,"gun_control")^?,
    lambda: 1.0,
    net_module: "nn_bert", net_class: "BertClassifier", net_config: "config_bert.json",
    fe_functions: [input("bert_post")];
    ...

    hardconstr: InThread(T, P) & InThread(T, Q) & Respond(P, Q) & Agree(P, Q)^?
               & IsPro(P, "gun_control")^? => IsPro(Q, "gun_control")^?;
    ...

} groupby: InThread.1;
```

(a) DRaiL Program

```python
import torch
from transformers import AutoConfig, AutoModel
from drail.neuro.nn_model import NeuralNetworks

class BertClassifier(NeuralNetworks):

    def __init__(self, config, nn_id, output_dim):
        super(BertClassifier, self).__init__(config, nn_id, output_dim)

    def build_architecture(self, rule_template, fe, shared_params={}):
        self.bert_model_name = self.config["bert_model_type"]
        bert_config = AutoConfig.from_pretrained(self.bert_model_name)
        self.bert_model = AutoModel.from_pretrained(self.bert_model_name)
        self.dropout = torch.nn.Dropout(bert_config.hidden_dropout_prob)
        self.hidden2label = \
                torch.nn.Linear(bert_config.hidden_size, self.output_dim)

    def forward(self, x):
        input_ids, input_mask, segment_ids = x['input']
        outputs = self.bert_model(input_ids, attention_mask=input_mask,
                                  token_type_ids=segment_ids)
        pooled_output = outputs[1]
        pooled_output = self.dropout(pooled_output)
        logits = self.hidden2label(pooled_output)
        probas = torch.nn.functional.softmax(logits, dim=1)
        return logits, probas
```

(b) Neural Network Specification

Figure 3.3. Code Snippets

3.1.1 Modeling Language

We begin our description of DRaiL by defining the templating language, consisting of entities, relations and rules, and explaining how these elements are instantiated given relevant data.

Entities

Entities are named *symbolic* or *attributed* elements. An example of a symbolic entity is a political ideology (e.g. Liberal or Conservative). An example of an attributed entity is a user with age, gender and other profile information, or a document associated with textual content. In DRaiL entities can appear either as *constants*, written as strings in double or single quote (e.g. "`user1`") or as *variables*, which are identifiers, substituted with constants when grounded. Variables are written using unquoted upper case strings (e.g. `X`, `X1`). Both constants and variables are typed.

Relations

Relations are defined between entities and their properties, or other entities. Relations are defined using a unique identifier, a named *predicate*, and a list of typed arguments.

Atoms

Atoms consist of a predicate name and a sequence of entities, consistent with the type and arity of the relation's argument list. If the atom's arguments are all constants, it is referred to as a **ground atom**. For example, `Agree`("`user1`", "`user2`") is a ground atom representing whether "`user1`" and "`user2`" are in agreement. When atoms are not grounded (e.g. `Agree`(`X`, `Y`)) they serve as placeholders for all the possible groundings that can be obtained by replacing the variables with constants. Relations can either be *closed* (i.e., all of their atoms are observed) or *open*, when some of the atoms can be unobserved. In DRaiL, we use a question mark `?` to denote unobserved relations. These relations are the units that we reason over.

To help make these concepts concrete, consider the following example analyzing stances in a debate, as introduced in Fig 3.1. First, we define the entities:

$$\texttt{User} = \{\text{"}\texttt{u1}\text{"}, \text{"}\texttt{u2}\text{"}\}, \texttt{Claim} = \{\text{"}\texttt{t1}\text{"}\}, \texttt{Post} = \{\text{"}\texttt{p1}\text{"}, \text{"}\texttt{p2}\text{"}\}$$

Users are entities associated with demographic attributes and preferences. Claims are assertions over which users debate. Posts are textual arguments that users write to explain their position w.r.t the claim. We create these associations by defining a set of relations, capturing authorship $\texttt{Author}(\texttt{User}, \texttt{Post})$, votes between users $\texttt{VoteFor}(\texttt{User}, \texttt{User})?$, and the position users, and their posts, take w.r.t to the debate claim. $\texttt{Agree}(\texttt{Claim}, \texttt{User})?$, $\texttt{Agree}(\texttt{Claim}, \texttt{Post})?$. The authorship relation is the only closed one, e.g., the atom: $\mathcal{O} = \{\texttt{Author}(\text{"}\texttt{u1}\text{"}, \text{"}\texttt{p1}\text{"})\}$.

Rules

Rules are functions that map literals (atoms or their negation) to other literals. Rules in DRaiL are defined using templates formatted as horn clauses: $t_{LH} \Rightarrow t_{RH}$, where t_{LH} (*body*) is a conjunction of literals, and t_{RH} (*head*) is the output literal to be predicted, and can only be an instance of open relations. Horn clauses allow us to describe structural dependencies as a collection of "if-then" rules, which can be easily interpreted. For example, $\texttt{Agree}(\texttt{X}, \texttt{C}) \wedge \texttt{VoteFor}(\texttt{Y}, \texttt{X}) \Rightarrow \texttt{Agree}(\texttt{Y}, \texttt{C})$ expresses the dependency between votes and users holding similar stances on a specific claim. We note that rules can be rewritten in *disjunctive form* by converting the logical implication into a disjunction between the negation of the body and the head. For example, the rule above can be rewritten as $\neg\texttt{Agree}(\texttt{X}, \texttt{C}) \vee \neg\texttt{VoteFor}(\texttt{Y}, \texttt{X}) \vee \texttt{Agree}(\texttt{Y}, \texttt{C})$.

The DRaiL program

The DRaiL program consists of a set of rules, which can be weighted (i.e., soft constraints), or unweighted (i.e., hard constraints). Each *weighted rule* template defines a learning problem, used to score assignments to the head of the rule. Since the body may contain open atoms, each rule represents a factor function expressing dependencies between open atoms in the body and head. Unweighted rules, or *constraints*, shape the space of feasible assignments to open atoms, and represent background knowledge about the domain.

Given the set of grounded atoms $\mathcal{O}$, rules can be grounded by substituting their variables with constants, such that the grounded atoms correspond to elements in $\mathcal{O}$. This process results in a set of grounded rules, each corresponding to a potential function or to a constraint. Together they define a factor graph. Then, DRaiL finds the optimally scored assignments for open atoms by performing MAP inference. To formalize this process, we first make the observation that rule groundings can be written as linear inequalities, directly corresponding to their disjunctive form, as follows –

$$\sum_{\mathrm{i} \in I_r^+} y_{\mathrm{i}} + \sum_{\mathrm{i} \in I_r^-} (1 - y_{\mathrm{i}}) \geq 1 \tag{3.1}$$

Where I_r^+ (I_r^-) correspond to the set of open atoms appearing in the rule that are not negated (respectively, negated). Now, MAP inference can be defined as a linear program. Each rule grounding r, generated from template $t(r)$, with input features $\boldsymbol{x_r}$ and open atoms $\boldsymbol{y_r}$ defines the potential –

$$\psi_r(\boldsymbol{x_r}, \boldsymbol{y_r}) = \min \Big\{ \sum_{\mathrm{i} \in I_r^+} y_{\mathrm{i}} + \sum_{\mathrm{i} \in I_r^-} (1 - y_{\mathrm{i}}), 1 \Big\} \tag{3.2}$$

added to the linear program with a weight w_r. Unweighted rule groundings are defined as –

$$c(\boldsymbol{x_c}, \boldsymbol{y_c}) = 1 - \sum_{\mathrm{i} \in I_c^+} y_{\mathrm{i}} - \sum_{\mathrm{i} \in I_c^-} (1 - y_{\mathrm{i}}) \tag{3.3}$$

with $c(\boldsymbol{x_c}, \boldsymbol{y_c}) \leq 0$ added as a constraints to the linear program. This way, the MAP problem can be defined over the set of all potentials Ψ and the set of all constraints C as –

$$\begin{aligned} \arg\max_{\boldsymbol{y} \in \{0,1\}^n} P(\boldsymbol{y}|\boldsymbol{x}) \equiv & \arg\max_{\boldsymbol{y} \in \{0,1\}^n} \sum_{\psi_{r,t} \in \Psi} w_r \ \psi_r(\boldsymbol{x_r}, \boldsymbol{y_r}) \\ & \text{s.t. } c(\boldsymbol{x_c}, \boldsymbol{y_c}) \leq 0; \ \ \forall c \in C \end{aligned} \tag{3.4}$$

In addition to logical constraints, we also support arithmetic constraints than can be written in the form of linear combinations of atoms with an inequality or an equality. For example, we can enforce the mutual exclusivity of liberal and conservative ideologies for any user `X` by writing:

$$\texttt{Ideology}(\texttt{X}, \text{“}\texttt{con}\text{”}) + \texttt{Ideology}(\texttt{X}, \text{“}\texttt{lib}\text{”}) = \texttt{1}$$

We borrow some additional syntax from PSL to make arithmetic rules easier to use. [18] define a *summation atom* as an atom that takes terms and/or sum variables as arguments. A summation atom represents the summations of ground atoms that can be obtained by substituting individual variables and summing over all possible constants for sum variables. For example, we could rewrite the above ideology constraint as $\texttt{Ideology}(\texttt{X}, +\texttt{I}) = \texttt{1}$. Where $\texttt{Ideology}(\texttt{X}, +\texttt{I})$ represents the summation of all atoms with predicate `Ideology` that share variable `X`

DRaiL uses two solvers, Gurobi [153] and AD3 [78] for exact and approximate inference respectively.

To ground DRaiL programs in data, we create an in-memory database consisting of all relations expressed in the program. Observations associated with each relation are provided in column separated text files. DRaiL's compiler instantiates the program by automatically querying the database and grounding the formatted rules and constraints.

3.1.2 Neural Components

Let r be a rule *grounding* generated from *template* t, where t is tied to a neural scoring function Φ_t and a set of parameters θ_t (Rule Layer in Fig 3.2). In the previous section, we defined the MAP problem for all potentials $\psi_r(\boldsymbol{x}, \boldsymbol{y}) \in \Psi$ in a DRaiL program, where each potential has a weight w_r. Consider the following scoring function:

$$w_r = \Phi_t(\boldsymbol{x_r}, \boldsymbol{y_r}; \theta^t) = \Phi_t(x_{\text{rel}_0}, ..., x_{\text{rel}_{n-1}}; \theta^t) \tag{3.5}$$

Notice that all potentials generated by the same template share parameters. We define each scoring function Φ_t over the set of atoms on the left hand side of the rule template. Let $t = \text{rel}_0 \wedge \text{rel}_1 \wedge ... \wedge \text{rel}_{n-1} \Rightarrow \text{rel}_n$ be a rule template. Each atom rel_i is composed of a relation type, its arguments and feature vectors for them, as shown in Fig. 3.2, "Input Layer".

Given that a DRaiL program is composed of many competing rules over the same problem, we want to be able to share information between the different decision functions. For this purpose, we introduce RelNets.

3.1.3 RelNets

A DRaiL program often uses the same entities and relations in multiple different rules. The symbolic aspect of DRaiL allows us to constrain the values of open relations, and force consistency across all their occurrences. The neural aspect, as defined in Eq. 3.5, associates a neural architecture with each rule template, which can be viewed as a way to embed the output relation.

We want to exploit the fact that there are repeating occurrences of entities and relations across different rules. Given that each rule defines a learning problem, sharing parameters allows us to shape the representations using complementary learning objectives. This form of relational multi-task learning is illustrated it in Fig. 3.2, "RelNets Layer".

We formalize this idea by introducing relation-specific and entity-specific encoders and their parameters $(\phi_{\text{rel}}; \theta^{\text{rel}})$ and $(\phi_{\text{ent}}; \theta^{\text{ent}})$, which are reused in all rules. As an example, let's write the formulation for the rules outlined in Fig. 3.2, where each relation and entity encoder is defined over the set of relevant features:

$$
\begin{aligned}
w_{r_0} &= \Phi_{t_0}(\phi_{\text{debates}}(\phi_{\text{user}}, \phi_{\text{text}})) \\
w_{r_1} &= \Phi_{t_1}(\phi_{\text{agree}}(\phi_{\text{user}}, \phi_{\text{text}}), \phi_{\text{votefor}}(\phi_{\text{user}}, \phi_{\text{user}}))
\end{aligned}
$$

Note that entity and relation encoders can be arbitrarily complex, depending on the application. For example, when dealing with text, we could use BiLSTMs or a BERT encoder.

Our goal when using RELNETS is to learn entity representations that capture properties unique to their types (e.g. users, issues), as well as relational patterns that contextualize entities, allowing them to generalize better. We make the distinction between *raw* (or *attributed*) entities and *symbolic* entities. Raw entities are associated with rich, yet unstructured information and attributes, such as text or user profiles. On the other hand, symbolic entities are well defined concepts, and are not associated with additional information, such as political ideologies (e.g. *liberal*) and issues (e.g. *gun-control*). With this consideration, we identify two types of representation learning objectives:

Embed Symbol / Explain Data:

This objective aligns the embedding of symbolic entities and raw entities, grounding the symbol in the raw data, and using the symbol embedding to explain properties of previously unseen raw-entity instances. For example, aligning ideologies and text to (1) obtain an ideology embedding that is closest to the statements made by people with that ideology, or (2) interpret text by providing a symbolic label for it.

Translate / Correlate:

This objective aligns the representation of pairs of symbolic or raw entities. For example, aligning user representations with text, to move between social and textual information, as shown in Fig. 3.1, "Social-Linguistic Relations". Or capturing the correlation between symbolic judgements like agreement and matching ideologies.

3.2 Learning Approaches

The scoring function used for comparing output assignments can be learned *locally* for each rule separately, or *globally*, by considering the dependencies between rules.

3.2.1 Global Learning

The global approach uses inference to ensure that the parameters for all weighted rule templates are consistent across all decisions. Let Ψ be a factor graph with potentials $\{\psi_r\} \in \Psi$ over the all possible structures Y. Let $\boldsymbol{\theta} = \{\theta^t\}$ be a set of parameter vectors, and $\Phi_t(\boldsymbol{x_r}, \boldsymbol{y_r}; \theta^t)$ be the scoring function defined for potential $\psi_r(\boldsymbol{x_r}, \boldsymbol{y_r})$. Here $\hat{\mathbf{y}} \in Y$ corresponds to the current prediction resulting from the MAP inference procedure and $\mathbf{y} \in Y$ corresponds to the gold structure. We support two ways to learn $\boldsymbol{\theta}$:

(1) The structured hinge loss:

$$\max(0, \max_{\hat{\mathbf{y}} \in Y}(\Delta(\hat{\mathbf{y}}, \mathbf{y}) + \sum_{\psi_r \in \Psi} \Phi_t(\boldsymbol{x_r}, \boldsymbol{\hat{y}_r}; \theta^t)) - \sum_{\psi_r \in \Psi} \Phi_t(\boldsymbol{x_r}, \boldsymbol{y_r}; \theta^t) \tag{3.6}$$

(2) The general CRF loss:

$$\begin{aligned} -\log p(\mathbf{y}|\mathbf{x}) &= -\log \left(\frac{1}{Z(\mathbf{x})} \prod_{\psi_r \in \Psi} \exp \left\{ \Phi_t(\boldsymbol{x_r}, \boldsymbol{y_r}; \theta^t) \right\} \right) \\ &= -\sum_{\psi_r \in \Psi} \Phi_t(\boldsymbol{x_r}, \boldsymbol{y_r}; \theta^t) + \log Z(\mathbf{x}) \end{aligned} \tag{3.7}$$

Where $Z(\mathbf{x})$ is a global normalization term computed over the set of all valid structures Y:

$$Z(\mathbf{x}) = \sum_{\mathbf{y'} \in Y} \prod_{\psi_r \in \Psi} \exp \left\{ \Phi_t(\boldsymbol{x_r}, \boldsymbol{y'_r}; \theta^t) \right\}$$

When inference is intractable, approximate inference (e.g. AD^3) can be used to obtain $\boldsymbol{\hat{y}}$. To approximate the global normalization term $Z(\boldsymbol{x})$ in the general CRF case, we follow previous work [154], [155] and keep a pool β_k of k of high-quality feasible solutions during inference. This way, we can sum over the solutions in the pool to approximate the partition function $\sum_{\mathbf{y'} \in \beta_k} \prod_{\psi_r \in \Psi} \exp \{\Phi_t(\boldsymbol{x_r}, \boldsymbol{y'_r}; \theta^t)\}$.

In this paper, we use the structured hinge loss for most experiments, and include a discussion on the approximated CRF loss in Section 3.3.6.

3.2.2 Joint Inference

The parameters for each weighted rule template are optimized independently. Building on previous work [155], we show that joint inference serves as a way to greedily approximate the CRF loss, where we replace the normalization term in Eq. 3.7 with a greedy approximation over local normalization as:

$$\begin{aligned} &-\log\left(\frac{1}{\prod_{\psi_r \in \Psi} Z_L(\boldsymbol{x_r})} \prod_{\psi_r \in \Psi} \exp\left\{\Phi_t(\boldsymbol{x_r}, \boldsymbol{y_r}; \theta^t)\right\}\right) \\ &= -\sum_{\psi_r \in \Psi} \Phi_t(\boldsymbol{x_r}, \boldsymbol{y_r}; \theta^t) + \sum_{\psi_r \in \Psi} \log Z_L(\boldsymbol{x_r}) \end{aligned} \tag{3.8}$$

Where $Z_L(\boldsymbol{x_r})$ is computed over all the valid assignments $\boldsymbol{y'_r}$ for each factor ψ_r:

$$Z_L(\mathbf{x}_r) = \sum_{\boldsymbol{y'_r}} \exp\left\{\Phi_t(\boldsymbol{x_r}, \boldsymbol{y'_r}; \theta^t)\right\} \tag{3.9}$$

We refer to models that use this approach as JointInf.

3.3 Experimental Evaluation

We compare DRaiL to representative models for the following three categories: end-to-end neural networks, relational embedding methods, and probabilistic logic frameworks. Our goal is to examine how different types of approaches capture dependencies and what are their limitations when dealing with language interactions. These baselines are described in Sec. 3.3.1. We also evaluate different strategies using DRaiL in Sec. 3.3.2.

We focus on three tasks: open debate stance prediction (Sec. 3.3.3), issue-specific stance prediction (Sec. 3.3.4) and argumentation mining (Sec. 3.3.5). Details regarding the hyper-parameters used for all tasks can be found in Tab. 3.1.

Table 3.1. Hyper-parameter Tuning

Task	Param	Search Space	Selected Value
Open Domain (Local)	Learning Rate	2e-6,5e-6,2e-5,5e-5	2e-5
	Batch size	32 (Max. Mem)	32
	Patience	1,3,5	3
	Optimizer	SGD,Adam,AdamW	AdamW
	Hidden Units	128,512	512
	Non-linearity	-	ReLU
Open Domain (Global)	Learning Rate	2e-6,5e-6,2e-5,5e-5	2e-6
	Batch size	-	Full instance
Stance Pred. (Local)	Learning Rate	2e-6,5e-6,2e-5,5e-5	5e-5
	Patience	1,3,5	3
	Batch size	16 (Max. Mem)	16
	Optimizer	SGD,Adam,AdamW	AdamW
Stance Pred. (Global)	Learning Rate	2e-6,5e-6,2e-5,5e-5	2e-6
	Batch size	-	Full instance
Arg. Mining (Local)	Learning Rate	1e-4,5e-4,5e-3,1e-3,5e-2,1e-2	5e-2
	Patience	5,10,20	20
	Batch size	16,32,64,128	64
	Dropout	0.01,0.05,0.1	0.05
	Optimizer	SGD,Adam,AdamW	SGD
	Hidden Units	128,512	128
	Non-linearity	-	ReLU
Arg. Mining (Global)	Learning Rate	1e-4,5e-4,5e-3,1e-3,5e-2,1e-2	1e-4
	Patience	5,10,20	10
	Batch size	-	Full instance

3.3.1 Baselines

End-to-end Neural Networks

We test all approaches against neural networks trained locally on each task, without explicitly modeling dependencies. In this space, we consider two variants: IndNets, where each component of the problem is represented using an independent neural network, and E2E, where the features for the different components are concatenated at the input and fed to a single neural network.

Relational Embedding Methods

Introduced in Chapter 2, these methods embed nodes and edge types for relational data in a distributed space. They are typically designed to represent symbolic entities and relations. However, since our entities can be defined by raw textual content and other features, we define the relational objectives over our expressive entity and relation encoders. This adaptation has proven successful for domains dealing with rich textual information [156]. We test three relational knowledge objectives: TransE [29], ComplEx [31] and RotatE [32]. **Limitations:** (1) These approaches cannot constrain the space using domain knowledge, and (2) they cannot deal with relations involving more than two entities, limiting their applicability to higher order factors.

Probabilistic Logic Frameworks

We compare to Probabilistic Soft Logic (PSL) [18], a purely symbolic probabilistic logic, and TensorLog [4], a neural-symbolic one. In both cases, we instantiate the program using the weights learned with our base encoders. **Limitations:** These approaches do not provide a way to update the parameters of the base classifiers.

3.3.2 Modeling Strategies

Local vs. Global Learning

The trade-off between local and global learning has been explored for graphical models (MEMM vs. CRF), and for deep structured prediction [43], [55], [155]. While local learning is faster, the learned scoring functions might not be consistent with the correct global prediction. Following Han et al., 2019 [55], we initialize the parameters using local models.

RelNets

We will show the advantage of having relational representations that are shared across different decisions, in contrast to having independent parameters for each rule. Note that in all cases, we will use the global learning objective to train RelNets.

Modularity

Decomposing decisions into relevant modules has been shown to simplify the learning process and lead to better generalization [157]. We will contrast the performance of *modular* and *end-to-end* models to represent text and user information when predicting stances.

Representation Learning and Interpretability

We will do a qualitative analysis to show how we are able to *embed symbols* and *explain data* by moving between symbolic and sub-symbolic representations, as outlined in Section 3.1.3.

3.3.3 Open Domain Stance Prediction

Traditionally, stance prediction tasks have focused on predicting stances on a specific topic, such as abortion. Predicting stances for a different topic, such as gun control would require learning a new model from scratch. In this task, we would like to leverage the fact that stances in different domains are correlated. Instead of using a pre-defined set of debate topics (i.e., *symbolic* entities) we define the prediction task over claims, expressed in text, specific to each debate. Concretely, each debate will have a different claim (i.e., different value for `C` in the relation `Claim(T, C)`, where `T` corresponds to a debate thread). We refer to these settings as *Open-Domain* and write down the task in Fig. 3.4. In addition to the textual stance prediction problem (r0), where `P` corresponds to a post, we represent users (`U`) and define a user-level stance prediction problem (r1). We assume that additional users read the posts and vote for content that supports their views, resulting in another prediction problem (r2,r3). Then, we define representation learning tasks, which align symbolic (ideology, defined as `I`) and raw (users and text) entities (r4-r7). Finally, we write down all dependencies and constrain the final prediction (c0-c7).

// *Rules*

// *Learning objectives*

r0: $\mathtt{InThread(T,P) \land Claim(T,C) \Rightarrow Agree(P,C)?}$

r1: $\mathtt{Debates(T,U) \land Claim(T,C) \Rightarrow Agree(U,C)?}$

r2: $\mathtt{Debates(T,U) \land Votes(T,V) \Rightarrow VoteFor(V,U)?}$

r3: $\mathtt{Votes(T,V_1) \land Votes(T,V_2) \Rightarrow VoteSame(V_1,V_2)?}$

// *Auxiliary objectives*

r4: $\mathtt{InThread(T,P) \land Ideology(I) \Rightarrow HasIdeology(P,I)?}$

r5: $\mathtt{Claim(T,C) \land Ideology(I) \Rightarrow HasIdeology(C,I)?}$

r6: $\mathtt{Debates(T,U) \land Ideology(I) \Rightarrow HasIdeology(U,I)?}$

r7: $\mathtt{HasIdeology(A,I)? \land HasIdeology(B,I)? \Rightarrow Agree(A,B)?}$

// *Constraints*

// *Author constraints*

c0: $\mathtt{Agree(P,C)? \land Author(P,U) \Rightarrow Agree(U,C)?}$

c1: $\mathtt{\neg Agree(P,C)? \land Author(P,U) \Rightarrow \neg Agree(U,C)?}$

// *Debate constraints*

c2: $\mathtt{Agree(P_1,C)? \land Respond(P_1,P_2) \Rightarrow \neg Agree(P_2,C)?}$

c3: $\mathtt{\neg Agree(P_1,C)? \land Respond(P_1,P_2) \Rightarrow Agree(P_2,C)?}$

// *Social constraints*

c4: $\mathtt{Agree(U,C)? \land VoteFor(V,U)? \Rightarrow Agree(V,U)?}$

c5: $\mathtt{\neg Agree(U,C)? \land VoteFor(V,U)? \Rightarrow \neg Agree(V,U)?}$

c6: $\mathtt{Agree(V_1,C)? \land VoteSame(V_1,V_2)? \Rightarrow Agree(V_2,C)?}$

c7: $\mathtt{\neg Agree(V_1,C)? \land VoteSame(V_1,V_2)? \Rightarrow \neg Agree(V_2,C)?}$

Figure 3.4. DRaiL Program for O.D. Stance Prediction. T: Thread, C: Claim, P: Post, U: User, V: Voter, I: Ideology, A,B: Can be any in {Claim, Post, User}

Dataset

We collected a set of 7,555 debates from debate.org, containing a total of 42,245 posts across 10 broader political issues. For a given issue, the debate topics are nuanced and vary according to the debate question expressed in text (e.g. *Should semi-automatic guns be banned*, *Conceal handgun laws reduce violent crime*). Debates have at least two posts, containing up to 25 sentences each. In addition to debates and posts, we collected the user profiles of all users participating in the debates, as well as all users that cast votes for the debate participants. Profiles consist of attributes (e.g. gender, ideology). User data is considerably sparse.

We create two evaluation scenarios, *random* and *hard*. In the random split, debates are randomly divided into ten folds of equal size. In the hard split, debates are separated by political issue. This results in a harder prediction problem, as the test data will not share topically related debates with the training data. We perform 10-fold cross validation and report accuracy.

Entity and Relation Encoders

We represent posts and titles using a pre-trained BERT-small[2] encoder [158], a compact version of the language model proposed by Devlin et al., 2018 [150]. For users, we use feed-

[2] ↑We found negligible difference in performance between BERT and BERT-small for this task, while obtaining a considerable boost in speed

forward computations with ReLU activations over the profile features and a pre-trained node embedding [26] over the friendship graph. All relation and rule encoders are represented as feed-forward networks with one hidden layer, ReLU activations and a softmax on top. Note that all of these modules are updated during learning.

Results

Tab. 3.2 shows results for all the models described in Section 3.3.1. In E2E models, post and user information is collapsed into a single module (rule), whereas in IndNets, JointInf, Global and RelNets they are modeled separately. All other baselines use the same underlying modular encoders. We can appreciate the advantage of relational embeddings in contrast to IndNets for user and voter stances, particularly in the case of ComplEx and RotatE. We can attribute this to the fact that all objectives are trained jointly and entity encoders are shared. However, approaches that explicitly model inference, like PSL, TensorLog and DRaiL outperform relational embeddings and end-to-end neural networks. This is because they enforce domain constraints.

Table 3.2. General Results for Open Domain Stance Prediction. P:Post, U:User, V:Voter

	Model	**Random**			**Hard**		
		P	**U**	**V**	**P**	**U**	**V**
Local	IndNets	63.9	61.3	54.4	62.2	53.0	51.3
	E2E	66.3	71.2	54.4	63.4	68.1	51.3
Reln. Emb.	TransE	58.5	54.1	52.6	57.2	53.1	51.2
	ComplEx	61.0	63.3	**58.1**	57.3	55.0	55.4
	RotatE	59.6	58.3	54.2	57.9	54.6	51.0
Prob. Logic.	PSL	78.7	77.5	55.4	72.6	71.8	52.6
	TensorLog	72.7	71.9	56.2	70.0	67.4	**55.8**
DRaiL	E2E +Inf	80.2	79.2	54.4	76.9	75.5	51.3
	JointInf	80.7	79.5	55.6	75.2	74.0	52.5
	Global	81.0	79.5	55.8	75.3	74.0	53.0
	RelNets	**81.9**	**80.4**	57.0	**78.0**	**77.2**	53.7

We explain the difference between the performance of DRaiL and the other probabilistic logics by: (1) The fact that we use exact inference instead of approximate inference, (2) PSL

Table 3.3. Variations of the Model for O.D. Stance Prediction. P:Post, U:User, V:Voter

	Model	**Random**			**Hard**		
		P	**U**	**V**	**P**	**U**	**V**
Local	IndNets	63.9	61.3	54.4	62.2	53.0	51.3
	E2E	66.3	71.2	54.4	63.4	68.1	51.3
AC	JointInf	73.6	71.8	-	69.0	67.2	-
	Global	73.6	72.0	-	69.0	67.2	-
	RelNets	73.8	72.0	-	71.7	69.5	-
AC	JointInf	80.7	79.5	-	75.6	74.4	-
DC	Global	81.4	79.9	-	75.8	74.6	-
	RelNets	81.8	80.1	-	77.8	76.4	-
AC	JointInf	80.7	79.5	55.6	75.2	74.0	52.5
DC	Global	81.0	79.5	55.8	75.3	74.0	53.0
SC	RelNets	**81.9**	**80.4**	**57.0**	**78.0**	**77.2**	**53.7**

learns to weight the rules without giving priority to a particular task, whereas the JointInf model works directly over the local outputs, and *most importantly* (3) our Global and RelNets models back-propagate to the base classifiers and fine-tune parameters using a structured objective.

In Tab. 3.3 we show different versions of the DRaiL program, by adding or removing certain constraints. AC models only enforce author consistency, AC-DC models enforce both author consistency and disagreement between respondents, and finally, AC-DC-SC models introduce social information by considering voting behavior. We get better performance when we model more contextualizing information for the RelNets case. This is particularly helpful in the *Hard* case, where contextualizing information, combined with shared representations, help the model generalize to previously unobserved topics. With respect to the modeling strategies listed in Section 3.3.2, we can observe: (1) The advantage of using a global learning objective, (2) the advantage of using RelNets to share information and (2) the advantage of breaking down the decision into modules, instead of learning an end-to-end model. Then, we perform a qualitative evaluation to illustrate our ability to move between symbolic and raw information. Tab. 3.4 (Top) takes a set of statements and *explains* them by looking at the symbols associated with them and their score. For learning to map debate statements to ideological symbols, we rely on the partial supervision provided by the users that self-identify with a political ideology and disclose it on their public profiles. Note that we do not incorporate any explicit expertise in political science to learn to represent ideological information. We chose statements with the highest score for each of the ideologies. We can see that, in the context of guns, statements that have to do with some form of gun control have higher scores for the center-to-left spectrum of ideological symbols (moderate, liberal, progressive), whereas statements that mention gun rights and the ineffectiveness of gun control policies have higher scores for conservative and libertarian symbols.

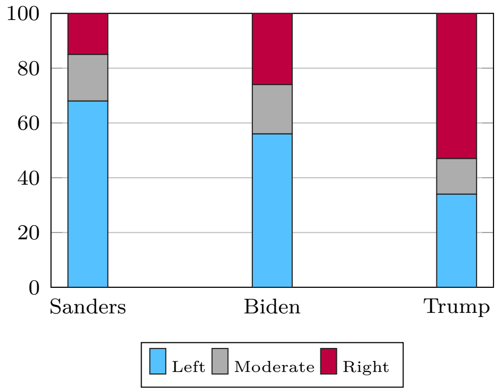
Table 3.4. Representation Learning Objectives: Explain Data (Top) and Embed Symbol (Bottom)

Issue	**Debate Statements**	**Con**	**Libt**	**Mod**	**Libl**	**Pro**
Guns	No laws should be passed restricting the right to bear arms	**.98**	.01	.00	.01	.00
	Gun control is an ineffective comfort tactic...	.08	**.65**	.22	.02	.03
	Gun control is good for society	.14	.06	**.60**	.15	.06
	In the US handguns ought to be banned	.03	.01	.01	**.93**	.02
	The USA should ban most guns and confiscate them	.00	.00	.01	.00	**.99**

Issue	**Ideology**	**Statements close in the embedding space**
LGBTQ+	**Libl**	gay marriage ought be legalized, gay marriage should be legalized, same-sex marriage should be federally legal
	Con	Leviticus 18:22 and 20:13 prove the anti gay marriage position, gay marriage is not bad, homosexuality is not a sin nor taboo

To complement this evaluation, in Tab. 3.4 (Bottom), we *embed* ideologies and find three example statements that are close in the embedding space. In the context of LGBTQ+ issues, we find that statements closest to the liberal symbol are those that support the legalization of same-sex marriage, and frame it as a constitutional issue. On the other hand, the statements closest to the conservative symbol, frame homosexuality and same-sex marriage as a moral or religious issue, and we find statements both supporting and opposing same-sex marriage. This experiment shows that our model is easy to interpret, and provides an explanation for the decision made.

Finally, we evaluate our learned model over entities that have not been observed during training. To do this, we extract statements made by three prominent politicians from *on-*



Politician	**Issue**	**Statement**	**Label**
Sanders	Guns	For background checks, and closing loopholes	Left
	Guns	Intervene with mental illness, to prevent mass shootings	Mod
	Abortion	Advocate for family planning and funding for contraceptives	Left
Biden	Guns	Guns need to have trigger locks	Left
	Abortion	Accepts catholic church view that life begins at conception	Right
	Abortion	Ensure access to and funding for contraception	Left
Trump	Guns	No limits on guns; they save lives	Right
	Abortion	I am pro-life; fight ObamaCare abortion funding	Right
	Abortion	Planned Parenthood does great work on women's health	Left

Figure 3.5. Statements Made by Politicians Classified Using our Model

theissues.org. Then, we try to explain the politicians by looking at their predicted ideology. Results for this evaluation can be seen in Fig. 3.5. The top Fig. shows the proportion of statements that were identified for each ideology: left (liberal or progressive), moderate and right (conservative). We find that we are able to recover the relative positions in the political spectrum for the evaluated politicians: Bernie Sanders, Joe Biden and Donald Trump. We find that Sanders is the most left leaning, followed by Biden. In contrast, Donald Trump stands mostly on the right. We also include some examples of the classified statements. We show that we are able to identify cases in which the statement does not necessarily align with the known ideology for each politician.

// Rules
// Learning Objectives
r0: $\texttt{InThread}(\texttt{T},\texttt{P}) \Rightarrow \texttt{IsPro}(\texttt{P},\texttt{I})?$
r1: $\texttt{Respond}(\texttt{P}_2,\texttt{P}_1) \Rightarrow \texttt{Agree}(\texttt{P}_1,\texttt{P}_2)?$
// Constraints
// Dependency Constraints
c0: $\texttt{Agree}(\texttt{P}_1,\texttt{P}_2)? \wedge \texttt{IsPro}(\texttt{P}_1,\texttt{I})? \Rightarrow \texttt{IsPro}(\texttt{P}_2,\texttt{I})?$
c1: $\texttt{Agree}(\texttt{P}_1,\texttt{P}_2)? \wedge \neg\texttt{IsPro}(\texttt{P}_1,\texttt{I})? \Rightarrow \neg\texttt{IsPro}(\texttt{P}_2,\texttt{I})?$
c2: $\neg\texttt{Agree}(\texttt{P}_1,\texttt{P}_2)? \wedge \texttt{IsPro}(\texttt{P}_1,\texttt{I})? \Rightarrow \neg\texttt{IsPro}(\texttt{P}_2,\texttt{I})?$
c3: $\neg\texttt{Agree}(\texttt{P}_1,\texttt{P}_2)? \wedge \neg\texttt{IsPro}(\texttt{P}_1,\texttt{I})? \Rightarrow \texttt{IsPro}(\texttt{P}_2,\texttt{I})?$
// Author Constraints
c4: $\texttt{Author}(\texttt{P}_1,\texttt{A}) \wedge \texttt{Author}(\texttt{P}_2,\texttt{A}) \wedge \texttt{IsPro}(\texttt{P}_1,\texttt{I})? \Rightarrow \texttt{IsPro}(\texttt{P}_2,\texttt{I})?$
c4: $\texttt{Author}(\texttt{P}_1,\texttt{A}) \wedge \texttt{Author}(\texttt{P}_2,\texttt{A}) \wedge \neg\texttt{IsPro}(\texttt{P}_1,\texttt{I})? \Rightarrow \neg\texttt{IsPro}(\texttt{P}_2,\texttt{I})?$

Figure 3.6. DRaiL program for Issue-Specific Stance Prediction. T: Thread, P: Post, I: Issue, A: Author

3.3.4 Issue-Specific Stance Prediction

Given a debate thread on a specific issue (e.g. abortion), the task is to predict the stance w.r.t. the issue for each one of the debate posts [81]. Each thread forms a tree structure, where users participate and respond to each other's posts. We treat the task as a collective classification problem, and model the agreement between posts and their replies, as well as the consistency between posts written by the same author. The DRaiL program for this task can be observed in Fig. 3.6.

Dataset

We use the 4Forums dataset from the Internet Argument Corpus [81], consisting of a total of 1,230 debates and 24,658 posts on abortion, evolution, gay marriage and gun control. We use the same splits as Li et al., 2018 [83] and perform 5-fold cross validation.

Entity and Relation Encoders

We represented posts using pre-trained BERT encoders [150] and do not generate features for authors. As in the previous task, we model all relations and rules using feed-forward

Table 3.5. General Results for Issue-Specific Stance and Agreement Prediction (Macro F1). AB: Abortion, E: Evolution, GM: Gay Marriage, GC: Gun Control

	Model	AB		E		GM		GC	
		S	A	S	A	S	A	S	A
Local	IndNets	66.0	61.7	58.2	59.7	62.6	60.6	59.5	61.0
Reln.	TransE	62.5	62.9	53.5	65.1	58.7	69.3	55.3	65.0
Embed.	ComplEx	66.6	73.4	60.7	72.2	66.6	72.8	60.0	70.7
	RotatE	66.6	72.3	59.2	71.3	67.0	74.2	59.4	69.9
Prob.	PSL	81.6	74.4	69.0	64.9	83.3	74.2	71.9	71.7
Logic.	TensorLog	77.3	61.3	68.2	51.3	80.4	65.2	68.3	55.6
	JointInf	82.8	74.6	64.8	63.2	84.5	73.4	70.4	66.3
DRaiL	Global	88.6	**84.7**	72.8	72.2	**90.3**	81.8	76.8	72.2
	RelNets	**89.0**	83.5	**80.5**	**76.4**	89.3	**82.1**	**80.3**	**73.4**

networks with one hidden layer and ReLU activations. Note that we fine-tune all parameters during training.

Results

In Tab. 4.3 we can observe the general results for this task. We report macro F1 for post stance and agreement between posts for all issues. As in the previous task, we find that ComplEx and RotatE relational embeddings outperform IndNets, and probabilistic logics outperform methods that do not perform constrained inference. PSL outperforms JointInf for evolution and gun control debates, which are the two issues with less training data. Whereas JointInf outperforms PSL for debates on abortion and gay marriage. This could indicate that re-weighting rules may be advantageous for the cases with less supervision. Finally, we see the advantage of using a global learning objective and augmenting it with shared representations. Tab. 4.5 compares our model with previously published results.

3.3.5 Argument Mining

The goal of this task is to identify argumentative structures in essays. Each argumentative structure corresponds to a tree in a document. Nodes are predefined spans of text

Table 3.6. Previous work on Issue-Specific Stance Prediction (Stance Acc.)

Model	A	E	GM	GC	Avg
BERT [150]	67.0	62.4	67.4	64.6	65.4
PSL [50]	77.0	80.3	80.5	69.1	76.7
Struct. Rep. [83]	86.5	82.2	87.6	83.1	84.9
DRaiL RelNets	**89.2**	**82.4**	**90.1**	**83.1**	**86.2**

and can be labeled either as *claims, major claims* or *premises*, and edges correspond to support/attack relations between nodes. Domain knowledge is injected by constraining sources to be premises and targets to be either premises or major claims, as well as enforcing tree structures. We model nodes, links and second order relations, *grandparent* ($a \rightarrow b \rightarrow c$), and *co-parent* ($a \rightarrow b \leftarrow c$) [54]. Additionally, we consider link labels, denoted stances. The DRaiL program for this task can be observed in Fig. 3.7.

// Rules
// Learning Objectives
r0: $\texttt{InPar(C,P)} \Rightarrow \texttt{NodeType(C,T)}?$
r1: $\texttt{InPar}(\texttt{C}_1,\texttt{P}) \wedge \texttt{InPar}(\texttt{C}_2,\texttt{P}) \Rightarrow \texttt{Link}(\texttt{C}_1,\texttt{C}_2)?$
r2: $\texttt{Link}(\texttt{C}_1,\texttt{C}_2)? \Rightarrow \texttt{AttackStance}(\texttt{C}_1,\texttt{C}_2)?$
// Auxiliary Objectives
r3: $\texttt{InPar}(\texttt{C}_1,\texttt{P}) \wedge \texttt{InPar}(\texttt{C}_2,\texttt{P}) \wedge \texttt{InPar}(\texttt{C}_3,\texttt{P}) \Rightarrow \texttt{Grandp}(\texttt{C}_1,\texttt{C}_2,\texttt{C}_3)?$
r4: $\texttt{InPar}(\texttt{C}_1,\texttt{P}) \wedge \texttt{InPar}(\texttt{C}_2,\texttt{P}) \wedge \texttt{InPar}(\texttt{C}_3,\texttt{P}) \Rightarrow \texttt{Coparent}(\texttt{C}_1,\texttt{C}_2,\texttt{C}_3)?$
// Constraints
// Higher order dependencies
c0: $\texttt{Grandp}(\texttt{C}_1,\texttt{C}_2,\texttt{C}_3)? \wedge \texttt{Link}(\texttt{C}_1,\texttt{C}_2)? \Rightarrow \texttt{Link}(\texttt{C}_2,\texttt{C}_3)?$
c1: $\texttt{Grandp}(\texttt{C}_1,\texttt{C}_2,\texttt{C}_3)? \wedge \texttt{Link}(\texttt{C}_2,\texttt{C}_3)? \Rightarrow \texttt{Link}(\texttt{C}_1,\texttt{C}_2)?$
c2: $\texttt{Coparent}(\texttt{C}_1,\texttt{C}_2,\texttt{C}_3)? \wedge \texttt{Link}(\texttt{C}_1,\texttt{C}_3)? \Rightarrow \texttt{Link}(\texttt{C}_2,\texttt{C}_3)?$
c3: $\texttt{Coparent}(\texttt{C}_1,\texttt{C}_2,\texttt{C}_3)? \wedge \texttt{Link}(\texttt{C}_2,\texttt{C}_3)? \Rightarrow \texttt{Link}(\texttt{C}_1,\texttt{C}_3)?$
// Source is always a premise
c4: $\texttt{Link}(\texttt{C}_1,\texttt{C}_2)? \Rightarrow \texttt{NodeType}(\texttt{C}_1,\texttt{"Premise"})?$
// Multiclass constraint
c5: $\texttt{HasType}(\texttt{C},+\texttt{T})? = 1$
// Enforce tree structure
c6: $\texttt{Link}(\texttt{C}_1,+\texttt{C}_2)? \leq 1$
c7: $\texttt{Link}(\texttt{C}_1,\texttt{C}_2)? \Rightarrow \texttt{Path}(\texttt{C}_1,\texttt{C}_2)?$
c8: $\texttt{Path}(\texttt{C}_1,\texttt{C}_2)? \wedge \texttt{Path}(\texttt{C}_2,\texttt{C}_3)? \Rightarrow \texttt{Path}(\texttt{C}_1,\texttt{C}_3)?$
c9: $\texttt{InPar}(\texttt{C}_1,\texttt{P}) \wedge \texttt{InPar}(\texttt{C}_2,\texttt{P}) \wedge (\texttt{C}_1 = \texttt{C}_2) \Rightarrow \neg\texttt{Path}(\texttt{C}_1,\texttt{C}_2)?$

Figure 3.7. DRaiL program for Argument Mining. P: Paragraph, C: Component, T: Type

Dataset

We used the UKP dataset [77], consisting of 402 documents, with a total of 6,100 propositions and 3,800 links (17% of pairs). We use the splits used by Niculae et al., 2017 [54], and report macro F1 for components and positive F1 for relations.

Entity and Relation Encoders

To represent the component and the essay, we used a BiLSTM over the words, initialized with Glove embeddings [159], concatenated with a feature vector following Niculae et al., 2017 [54]. For representing the relation, we use a feed-forward computation over the components, as well as the relation features used in Niculae et al., 2017 [54].

Results

We can observe the general results for this task in Tab. 4.1. Given that this task relies on constructing the tree from scratch, we find that all methods that do not include declarative constraints (IndNets and relational embeddings) suffer when trying to predict links correctly. For this task, we did not apply TensorLog given that we could not find a way to express tree constraints using their syntax. Once again, we see the advantage of using global learning, as well as sharing information between rules using RelNets.

Table 3.7. General Results for Argument Mining

	Model	Node	Link	Avg	Stance
Local	IndNets	70.7	52.8	61.7	63.4
Reln. Embed.	TransE	65.7	23.7	44.7	44.6
	ComplEx	69.1	15.7	42.4	53.5
	RotatE	67.2	20.7	44.0	46.7
Prob. Logic	PSL	76.5	56.4	66.5	64.7
DRaiL	JointInf	78.6	59.5	69.1	62.9
	Global	**83.1**	61.2	72.2	**69.2**
	RelNets	82.9	**63.7**	**73.3**	68.4

Tab. 4.2 shows the performance of our model against previously published results. While we are able to outperform models that use the same underlying encoders and features, recent work by Kuribayashi et. al [160] further improved performance by exploiting contextualized word embeddings that look at the whole document, and making a distinction between argumentative markers and argumentative components. We did not find a significant improvement by incorporating their ELMo-LSTM encoders into our framework[3], nor by replacing our BiLSTM encoders with BERT. We leave the exploration of an effective way to leverage contextualized embeddings for this task for future work.

Table 3.8. Previous Work on Argument Mining

Model	Node	Link	Avg
Human upper bound	86.8	75.5	81.2
ILP Joint [77]	82.6	58.5	70.6
Struct RNN strict [54]	79.3	50.1	64.7
Struct SVM full [54]	77.6	60.1	68.9
Joint PointerNet [161]	84.9	60.8	72.9
Kuribayashi et. al, 2019 [160]	**85.7**	**67.8**	**76.8**
DRaiL RelNets	82.9	63.7	73.3

3.3.6 Loss Function Analysis

In this section we perform an evaluation of the CRF loss in DRaiL for the issue-specific stance prediction task. Note that one drawback of the CRF loss (Eq. 3.7) is that we need to accumulate the gradient for the approximated partition function. When using entity encoders with a lot of parameters (e.g. BERT), the amount of memory needed for a single instance increases. We were unable to fit the full models in our GPU. For the purpose of these tests, we froze the BERT parameters after local training and updated only the relation and rule parameters. To obtain the solution pool, we use gurobi's pool search mode to find β high-quality solutions. This also increases the cost of search at inference time.

[3] ↑We did not experiment with their normalization approach, extended BoW features, nor AC/AM distinction.

Development set results for the debates on abortion can be observed in Tab. 3.9. While increasing the size of the solution pool leads to better performance, it comes at a higher computational cost.

Table 3.9. Stance Prediction (Abortion) Dev Results for Different Training Objectives

Model	Stance	Agree	Avg	Secs p/epoch
Hinge loss	82.74	78.54	80.64	132
CRF($\beta = 5$)	83.09	81.03	82.06	345
CRF($\beta = 20$)	84.10	82.16	83.13	482
CRF($\beta = 50$)	84.19	81.80	83.00	720

3.3.7 Runtime Analysis

In this section, we perform a run-time analysis of all probabilistic logic systems tested in this chapter. All experiments were run on a 12 core 3.2Ghz Intel i7 CPU machine with 63GB RAM and an NVIDIA GeForce GTX 1080 Ti 11GB GDDR5X GPU. Fig. 3.8 shows the overall training time (per fold) in seconds for each of the evaluated tasks. Note that the figure is presented in logarithmic scale. We find that DRaiL is generally more computationally expensive than both TensorLog and PSL. This is expected given that DRaiL back-propagates to the base classifiers at each epoch, while the other frameworks just take the local predictions as priors. However, when using a large number of arithmetic constraints (e.g. Argument Mining), we find that PSL takes a really long time to train.

We found no significant difference when using ILP or AD^3. We presume that this is due to the fact that our graphs are small and that Gurobi is a highly optimized commercial software.

Finally, we find that when using encoders with a large number of parameters (e.g. BERT) in tasks with small graphs, the difference in training time between training local and global models is minimal. In these cases, back-propagation is considerably more expensive than inference, and global models converge in fewer epochs. For Argument Mining, local models are at least twice as fast. BiLSTMs are considerably faster than BERT, and inference is more expensive for this task.

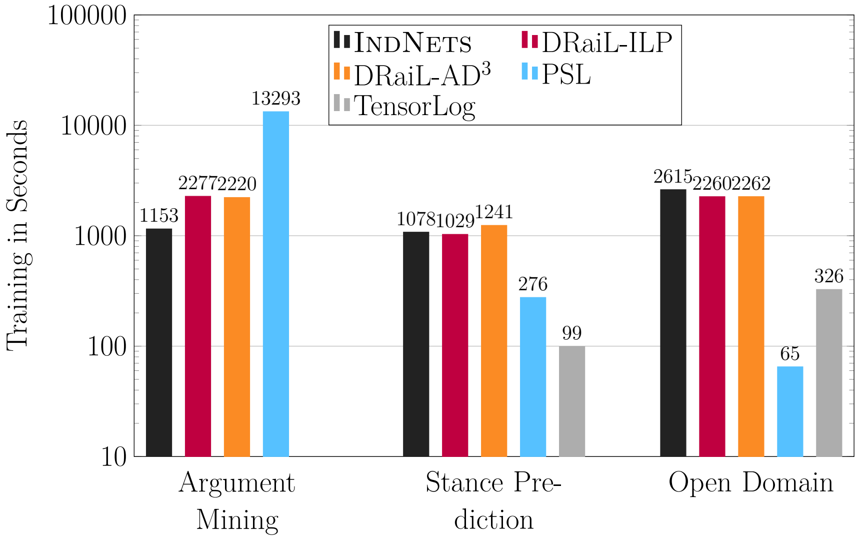


Figure 3.8. Avg. Overall Training Time (per Fold)

3.4 Summary

In this chapter, we motivate the need for a declarative neural-symbolic approach that can be applied to NLP tasks involving long texts and contextualizing information. We introduce a general framework to support this, and demonstrate its flexibility by modeling problems with diverse relations and rich representations, and obtain models that are easy to interpret and expand. The code, data and documentation for DRaiL and the application examples in this chapter have been released to the community, to help promote this modeling approach for other applications.

4. LEARNING DRaiL MODELS EFFICIENTLY WITH RANDOMIZED INFERENCE

Many discourse-level NLP tasks require modeling complex interactions between multiple sentences, paragraphs or even documents. For example, analyzing opinions in online conversations [50], [82] requires modeling the dependencies between the opinions in individual posts, the disagreement between posts in long conversational threads and the overall view of users, given all their posts.

Learning in these settings is extremely challenging. It requires highly expressive models that can capture the claims made in each document, either by using a rich, manually crafted feature set, or by using neural architectures to learn an expressive representation [54], [162]. In addition, reasoning about the interaction between these decisions is often computationally challenging, as it requires incorporating domain-specific constraints into the search procedure, making exact inference intractable. As a result, most current work relies on highly engineered solutions, which are difficult to adapt. Instead of training structured predictors that model the interaction between decisions during training, they combine locally trained classifiers at test time [77].

In the previous chapter, we introduced DRaiL, a declarative modeling framework that leverages *deep structured prediction* to combine rich neural representations with an inference-layer, forcing consistency between different interdependent decisions [163]. To solve the structured inference problem, DRaiL incorporates both exact and approximate inference algorithms. While successful, the computational complexity of these algorithms has the potential to escalate dramatically when the space of dependencies and constraints increases.

In previous work, randomized inference algorithms were proposed for structured NLP tasks, such as tagging and dependency parsing, in the context of linear models [56], [57], [59]. This approach offers an efficient alternative to exact inference. Instead of finding the optimal output state, the algorithms make greedy updates to a randomly initialized (or locally initialized) output assignment state. Our main contribution is to explore these ideas in the context of deep structured models composed of expressive text encoders, where theoretical guarantees are weak or nonexistent. Moreover, we do this for discourse-level tasks involving a rich set of domain constraints. To do this, we consider two variations of this approach. In the first, the algorithm samples and traverses only legal states (i.e., consistent with the constraints imposed by domain knowledge). In the second, these restrictions are ignored and only applied at test time. Adapting the sampling procedure to the specific constraints imposed by each domain is difficult, motivating the second approach as a generic alternative.

We focus on two discourse-level tasks, stance prediction in discussion forums, described above, and parsing argumentation structures in essays [77]. The latter consists of constructing an argumentation tree that represents the type-of, and relation-between, the arguments made in the essay. Models for both tasks typically employ declarative inference for incorporating domain knowledge. Our experiments are designed to quantify the trade-off between different modeling choices, both in terms of task performance and computational cost. We compare exact ILP models, approximate inference based on the popular AD^3 algorithm [78] and the two randomized inference algorithms. Our experiments show that in all cases, deep structured prediction outperforms traditional shallow approaches, structured learning outperforms inference over locally trained models, and generic randomized inference performs competitively to exact inference.

4.1 Modeling Approach

We look at two challenging structured prediction problems that deal with long texts where dependencies span across different paragraphs, documents and authors. To deal with these setups, we define neural factor graphs $G = \{\Psi\}$ where each decision $\psi_\mathrm{i} \in \Psi$ is associated with a neural architecture ρ_i and a set of parameters $\boldsymbol{\theta}_\mathrm{i}$. In this section, we introduce the tasks in detail.

4.1.1 Argument Mining

This task aims to identify argumentative structures in essays. Each argumentative structure forms a tree, and there is a forest per document. Nodes correspond to spans of text in the document and they can be labeled either as *claims, major claims* or *premises.* Edges correspond to stances (i.e., support/attack relations between nodes). The spans of texts are given, and we need to label nodes, predict which pairs of nodes are connected by an edge and label the edges. Domain knowledge can be exploited as there are only valid edges between pairs of premises, a premise and a claim, or a claim and a major claim. At the same time, all trees are rooted at major claims. Similarly to previous work, we model second order relationships: **grandparent** ($a \rightarrow b \rightarrow c$), and **co-parent** ($a \rightarrow b \leftarrow c$) [54], [164].

Figure 4.1 has a visual representation of the structure. In this problem, each forest defines a factor graph Ψ and G is the collection of all documents. We define a set of five neural architectures corresponding to the five types of decisions that we need to make: $NN = \{\rho_{\text{node}}, \rho_{\text{link}}, \rho_{\text{stance}}, \rho_{\text{grandparent}}, \rho_{\text{coparent}}\}$, each with its own set of parameters $\boldsymbol{\theta} = \{\theta_{\text{node}}, \theta_{\text{link}}, \theta_{\text{stance}}, \theta_{\text{grandparent}}, \theta_{\text{coparent}}\}$. Note that this corresponds to the task introduced in Chapter 3, Section 3.3.5. In principle, we can substitute each $(\rho_{\text{i}}, \theta_{\text{i}})$ with any neural architecture. We include details about the architectures in the experimental section.

4.1.2 Stance Prediction

Given a debate thread on a specific political issue, the task is to predict the stance of each post w.r.t. the issue (e.g., pro-life or pro-choice on abortion) [81]. Following previous work, we model the problem as a collective classification task and consider all posts in a given thread. To do this, we add the task of predicting stance agreement between consecutive posts. As observed in Figure 4.1, each thread forms a tree, where users participate and respond to each other's posts. For a thread labeling to be valid, we need to enforce consistency between the node and edge labels.

In this case, each discussion thread defines a factor graph Ψ and G is the collection of threads. We define two neural architectures $NN = \{\rho_{\text{stance}}, \rho_{\text{agreement}}\}$, each with its own set of parameters $\boldsymbol{\theta} = \{\theta_{\text{stance}}, \theta_{\text{agreement}}\}$. Note that this corresponds to the task introduced in Chapter 3, Section 3.3.4. As in the previous setup, each $(\rho_{\text{i}}, \theta_{\text{i}})$ can be substituted by any neural architecture, more details are outlined in the experimental section.

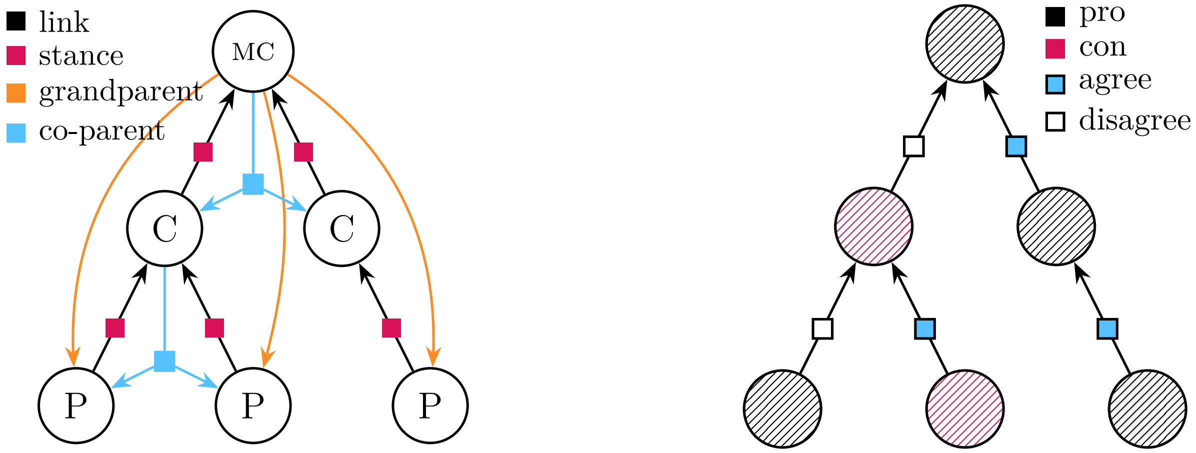


Figure 4.1. Argument Mining (left), Stance Prediction (right)

4.2 Learning Approach

We learn a joint neural model that uses inference during training to ensure consistency across all decisions. Let Ψ be a factor graph with potentials $\psi_\mathrm{i} \in \Psi$ over all possible structures Y. Let $\boldsymbol{x}_\mathrm{i}$ be the input vector to potential ψ_i. Let $\boldsymbol{\theta} = \{\theta^\mathrm{i}\}$ be a set of parameter vectors associated with a set of neural networks $\boldsymbol{\rho} = \{\rho_\mathrm{i}\}$, and $\rho_\mathrm{i}(\boldsymbol{x}_\mathrm{i}, \boldsymbol{y}_\mathrm{i}; \theta^\mathrm{i})$ is the score for potential ψ_i resulting from a forward pass. Here $\boldsymbol{y} \in Y$ corresponds to the gold structure and $\hat{\boldsymbol{y}} \in Y$ to the prediction resulting from the MAP inference procedure:

$$\begin{gathered}\arg\max_{\boldsymbol{y}\in Y} \sum_{\psi_\mathrm{i}\in\Psi} \rho_\mathrm{i}(\boldsymbol{x}_\mathrm{i}, \boldsymbol{y}_\mathrm{i}; \theta^\mathrm{i}) \\ s.t.\ c(\boldsymbol{x}, \boldsymbol{y})\ \ \forall c \in C\end{gathered} \tag{4.1}$$

Where C is a set of domain-specific constraints defined over the factor graph Ψ. To learn $\boldsymbol{\theta}$, we use the structured hinge loss $L(\boldsymbol{x}, \boldsymbol{y}, \hat{\boldsymbol{y}}; \boldsymbol{\theta})$ defined as:

$$\max\Big(0, \max_{\hat{\mathbf{y}}\in Y}\Big(\Delta(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \sum_{\psi_\mathrm{i}\in\Psi} \rho_\mathrm{i}(\boldsymbol{x}_\mathrm{i}, \hat{\boldsymbol{y}}_\mathrm{i}; \theta^\mathrm{i})\Big) - \sum_{\psi_\mathrm{i}\in\Psi} \rho_\mathrm{i}(\boldsymbol{x}_\mathrm{i}, \boldsymbol{y}_\mathrm{i}; \theta^\mathrm{i})\Big) \tag{4.2}$$

Where $\Delta(\boldsymbol{y}, \hat{\boldsymbol{y}})$ is the hamming loss, and we perform loss augmented inference. Given that this learning procedure corresponds to one of the learning protocols supported by DRaiL, we implement our models using DRaiL. Please, refer to Chapter 3, Section 3.2 for further learning details, and Sections 3.3.5 and 3.3.4 for the full DRaiL program specifications for these two tasks.

4.3 Randomized Inference

In this section, we describe the randomized inference procedure used for each task. We define the relevant domain constraints for each case, and explain how we sample solutions that respect them. To implement these inference algorithms in DRaiL, we extend the inference class for each scenario. Note that these implementations are domain-specific, but DRaiL allows us the possibility to override the general inference class for each case. Given that we have access to all the instances and constraints through the base inference class, the implementation is straight-forward.

Finally, we include a discussion about the theoretical bounds for the linear case.

4.3.1 Argument Mining

For randomized inference on argument mining, we adapt the randomized greedy algorithm proposed by Zhang, Lei, Barzilay, *et al.*, 2014 [56]. Algorithm 1 outlines the overall procedure. We will consider that each paragraph $p \in P$ of an essay contains a single tree. We obtain a local optimum tree $\hat{y}$ by using the hill climbing algorithm, which is further described below. After that, $\hat{y}$ is labeled and added to the forest Y. We iterate over each paragraph (line 4) and subsequently score the forest as:

$$\bar{\mathcal{S}}(Y) = \sum_{\hat{y} \in Y} \mathcal{S}(\hat{y}) = \sum_{\hat{y} \in Y} w + h \left\| y - \hat{y} \right\|_1 \tag{4.3}$$

Where $w = \sum_{\psi_{\mathrm{i}} \in \Psi} \rho_{\mathrm{i}}(\boldsymbol{x}_{\mathrm{i}}, \hat{\boldsymbol{y}}_{\mathrm{i}}; \theta^{\mathrm{i}})$ is the sum of the scores of the potentials for the predicted tree $\hat{y}$. We additionally add a weighted Hamming distance term to the scoring function. $h \left\| y - \hat{y} \right\|_1$ gets close to w if the distance is low, and close to zero if it is high. Whenever

Algorithm 1 Randomized Inference

```
1: $\hat{Y} \leftarrow \{\}$
2: for number of restarts do
3:     $Y \leftarrow \{\}$
4:     for each $p \in P$ do
5:         $\hat{y} \leftarrow$ hill_climbing($p$)
6:         label($\hat{y}$)
7:         $Y \leftarrow Y \cup \hat{y}$
8:     end for
9:     if $\bar{\mathcal{S}}(Y) > \bar{\mathcal{S}}(\hat{Y})$ then
10:        $\hat{Y} \leftarrow Y$
11:    end if
12: end for
13: return $\hat{Y}$ =0
```

the score of the locally improved forest is better than the forest found so far, Y becomes the new currently best scoring forest $\hat{Y}$. Since hill climbing might get stuck in a local optimum, we repeat line 3-9 for a constant number of restarts.

Algorithm 2 Hill Climbing

```
$\hat{y}_0 \leftarrow$ initialize tree randomly for paragraph $p$
label($\hat{y}_0$)
$\hat{y} \leftarrow \hat{y}_0$
$t \leftarrow 0$
repeat
    $\mathcal{L} \leftarrow$ top-down level node list of $\hat{y}$
    for $\mathrm{i} = 1, ..., |\mathcal{L}|$ do
        for $\mathrm{j} = \mathrm{i} - 1, ..., 0$ do
            $\hat{y}_{t+1} \leftarrow$ connect subtree of $\mathcal{L}_\mathrm{i}$ to $\mathcal{L}_\mathrm{j}$
            label($\hat{y}_{t+1}$)
            if $\mathcal{S}(\hat{y}_{t+1}) > \mathcal{S}(\hat{y})$ then
                $\hat{y} \leftarrow \hat{y}_{t+1}$
            end if
            $t \leftarrow t + 1$
        end for
    end for
until no improvement in this iteration
return $\hat{y}$ =0
```

Algorithm 2 describes the hill climbing procedure. It initially draws uniformly a tree $\hat{y}_0$ at random. Then the greedy algorithm applies local updates on $\hat{y}_t$ and attempts to achieve a better scoring tree $\hat{y}_{t+1}$. This is done by iterating through a top-down level node list $\mathcal{L}$ of $\hat{y}_t$. Denote i as the current position in the list, then the entire subtree of $\mathcal{L}_\mathrm{i}$ is connected to the node $\mathcal{L}_\mathrm{j}$, whereas $\mathrm{j} = \mathrm{i} - 1, \mathrm{i} - 2, \ldots, 0$. If the score of $\hat{y}_{t+1}$ is higher than the score of

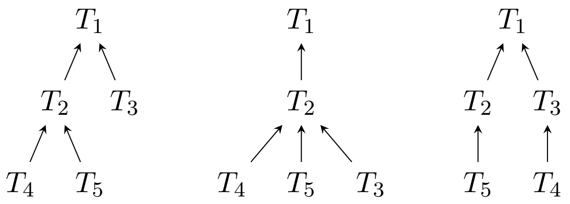


Figure 4.2. Greedy local update of a tree $\hat{y}_t$ (left) to $\hat{y}_{t+1}$ and $\hat{y}_{t+2}$ without score improvement

$\hat{y}_t$, the newly generated tree is kept. The algorithm continues until the score can no longer be improved and therefore yields a local optimum tree. Figure 4.2 depicts how such local updates are performed, $\mathcal{L} = (T_1, T_2, T_3, T_4, T_5)$.

It might be the case that a paragraph contains more than a single tree, therefore, when a tree is initially drawn at random, we introduce an additional *phantom node* which serves as the new root. This modification no longer restricts hill climbing on trees only. Moreover, it allows us having multiple roots and we treat the second layer of the tree like the top layer.

Domain-Specific Constraints

For node labeling, we exploit domain knowledge. *Major claims* can only occur in the first or last paragraph, and there has to be at least one major claim in each essay. A root gets labeled as major claim with some fixed probability depending on the paragraph (first or last), holding the condition that there has to exist at least one. Any other root is labeled as a *claim* in each paragraph. Nodes having an edge to a major claim are labeled as claims as well. All remaining nodes are *premises*. An edge can have either the label *support* or *attack* and we draw all edge labels randomly with a probability of 0.9 being a support label. The node and edge labels are determined after each iteration since scoring depends on both, links and labels.

In section 4.4, we evaluate our models using randomized inference with and without domain-specific constraints. In the latter case, all labels are chosen at random.

4.3.2 Stance Prediction

A debate thread provides a fixed structure, thus nodes and links are predefined and no improvement of the tree structure needs to be done. However, nodes and edges still need to be labeled and can be improved. Initially, we pick the node labels, which can either be *pro* or *con*. Following the observations made by [59], we leverage local classifiers and greedily choose the label with the highest score for each node.

Domain-Specific Constraints

To respect the dependencies between node and edges labels, we use the following heuristic: If two consecutive nodes u and v have different stances, the edge (u, v) receives a *disagreement* label, if they share the same stance, (u, v) gets an *agreement* label. When author constraints are considered as well, we additionally force stances of posts to be equal when written by the same author.

We attempt to improve node labels by flipping them randomly and subsequently adjust the edge labels. This is done until an iteration no longer improves the overall score. We restart the algorithm for a constant number of times in order to increase the chance of achieving a global optimum. In the experiments, we evaluate our models using randomized inference with and without domain specific constraints. When constraints are not used, a random node is flipped and its adjacent edge adjusted, without enforcing consistency in the whole tree.

4.3.3 Theoretical Bounds for the Linear Case

The error of the *constrained* randomized algorithms can be bound for the linear case. Let's define the norm of the set of parameter vectors $\boldsymbol{\theta}$ as follows: $\|\boldsymbol{\theta}\| = \sqrt{\sum_{\theta^{\mathrm{i}} \in \boldsymbol{\theta}} \|\theta^{\mathrm{i}}\|^2}$, where $\|\theta^{\mathrm{i}}\|$ is the Euclidean norm of the parameter vector θ^{i}. Let n be the number of training samples. From Theorem 2 and Claim ii in [58], for $\rho_{\mathrm{i}}(\boldsymbol{x}_{\mathrm{i}}, \boldsymbol{y}_{\mathrm{i}}; \theta^{\mathrm{i}})$ linear in θ^{i}, the generalization bound (i.e., the difference between the test error and training error) is on the order of $\|\boldsymbol{\theta}\|^2/n + \|\boldsymbol{\theta}\|/\sqrt{n} + \max(1/\log 2, \|\boldsymbol{\theta}\|^2) \log^{3/2} n/\sqrt{n}$. The above generalization bound is decreasing in n, and increasing in $\|\boldsymbol{\theta}\|$, which suggests the use of a large training set, and the penalization of the norm $\|\boldsymbol{\theta}\|$ during learning. In our experiments, we show that in practice we can obtain competitive results by implementing the randomized algorithms for the non-linear case.

4.4 Experimental Evaluation

We learn our models using four different inference procedures: (1) **ILP** defines the inference problem as an integer linear program and uses the Gurobi solver[1] to perform exact inference, (2) **AD^3/ILP** translates the ILP program into an AD^3 instance to perform approximate inference, (3) **Rand-C** uses the randomized method with domain constraints, and (4) **Rand** uses the randomized method without domain constraints. Note that we always use exact inference to evaluate on both the development and test sets. For completeness, we add an entry **AD^3** where we use AD^3 for both training and testing.

All experiments were run on a 32 core 3.2Ghz Intel Xeon CPU machine with 128GB RAM and an NVIDIA GeForce GTX 1080 Ti 11GB GDDR5X GPU. We performed an exhaustive search for hyper-parameters on the development set. We tuned the learning rate (lr ∈ {1e-6, 2e-6, 5e-6, 1e-5, 2e-5, 5e-5, 1e-4, 2e-4, 5e-4, 1e-3, 2e-3, 5e-3, 1e-2, 2e-2, 5e-2, 1e-1}), patience (p $\in \{5, 10, 15, 20\}$), and number of restarts (r $\in \{1, 5, 10, 15, 20, 30, 50, 100\}$). The weight decay was fixed at 1e-5. We found that results were stable for local and global models, for different sets of constraints and across inference algorithms.

4.4.1 Argument Annotated Persuasive Essays

Dataset

We used the UKP dataset [77], consisting of 402 documents, with a total of 6,100 propositions and 3,800 links (17% of pairs). We use the train/dev/test splits used by Niculae, Park, and Cardie, 2017 [54], and report macro F1 for components and positive F1 for relations.

[1] ↑https://www.gurobi.com/products/gurobi-optimizer

Learning and Representation

We did five repetitions and reported the average performance. Each repetition used a different seed to initialize the model parameters. For training, we used stochastic gradient descent, a patience of 10, weight decay of 1e-5, and five restarts for randomized inference. For local models, we used a learning rate of 0.05 and for structured learning we used a learning rate of 1e-4. Similarly to previous work on deep structured prediction [55], we obtained better results by performing structured learning over locally trained models, instead of training them from scratch. To represent the component and the essay, we used a BiLSTM over the words, initialized with GloVe embeddings [159], concatenated with a feature vector following Niculae, Park, and Cardie, 2017 [54], without the word features. For representing the relation, we use the components, as well as the relation features used in Niculae, Park, and Cardie, 2017 [54]. For shallow models, we use a bag-of-words representation for the text and concatenate it with the rest of the features into a single feature vector.

We test two versions of the model: (1) **Base** includes node labeling, link prediction and link labeling, and (2) **Full** adds grandparent and co-parent factors. Domain constraints are introduced in all models.

Table 4.1. F1 for argument mining using **deep structured prediction**, Avg results using **shallow** models included in parenthesis

Model	Inference	Node	Link	Avg	Stance
Local	–	70.7	52.8	61.7 (60.7)	63.4
	L+I	76.5	56.9	66.7 (66.5)	62.5
Base	ILP	83.0	57.6	70.3 (67.2)	68.0
	AD^3/ILP	**83.2**	58.2	**70.7** (67.2)	**68.4**
	AD^3	83.0	57.6	70.3 (67.2)	68.0
	Rand-C	82.8	58.4	70.6 (67.7)	**68.4**
	Rand	82.9	**58.5**	**70.7** (67.7)	68.0
Full	ILP	83.1	61.2	72.2 (65.3)	**69.2**
	AD^3/ILP	83.7	62.0	72.9 (65.3)	68.5
	AD^3	83.5	61.1	72.3 (65.3)	**69.2**
	Rand-C	**83.8**	62.6	73.2 (66.3)	68.4
	Rand	83.4	**63.2**	**73.3** (65.9)	68.4

Results

We can analyze the results across three dimensions:

Structured Learning

The advantage of leveraging more structural dependencies can be seen in Table 4.1. The model gets increasingly better as more dependencies are considered, and using global learning outperforms learning local models and using inference just at prediction time (L+I).

Deep vs. Shallow

There is a consistent trend showing that deep structured models are more expressive than their shallow counterparts, as we can see by comparing average results in Table 4.1. To obtain good results using linear classifiers, Stab and Gurevych, 2017 [77] relied on an exhaustive set of features (Table 4.2). These numbers cannot be replicated by using just word-features and the feature set suggested by Niculae, Park, and Cardie, 2017 [54], as our shallow models and their structured SVM results show. In contrast, deep models and word embeddings are able to leverage this information without additional features. In addition, we find that deep models have a shorter overall training time (3.3x faster for the full model). This can be attributed to the compact embedding representation used in deep models, in contrast to the large sparse one-hot vectors used in linear models. Similarly to previous work [54], we find that higher-order factors and strict constraints are more helpful when using deep structured models than in their shallow counterparts.

Randomized vs. ILP/AD3

When using deep structured prediction, we did not find a statistically significant difference in the performance of the models that were trained with ILP/AD3 vs. the ones that were trained with constrained and non-constrained randomized inference.

Table 4.2. Previous work on UKP Dataset

Model	Node	Link	Avg
Human upper bound	86.8	75.5	81.2
ILP Joint [77]	82.6	58.5	70.6
Struct RNN strict [54]	79.3	50.1	64.7
Struct RNN full [54]	76.9	50.4	63.6
Struct SVM strict [54]	77.3	56.9	67.1
Struct SVM full [54]	77.6	60.1	68.9
Joint PointerNet [161]	84.9	60.8	72.9
kuribayashi-etal-2019-empirical, 2019 [160]	**85.7**	**67.8**	**76.8**
BERT [150]	71.1	50.8	61.0
BERT-doc	79.5	55.8	67.7
BERT-doc + Inf (Base)	79.9	58.1	69.0
BERT-doc + Structured Prediction (Base)	82.1	60.0	71.1
Deep Full ILP	83.1	61.2	72.2
Deep Full Rand-C	**83.8**	62.6	73.2
Deep Full Rand	83.4	**63.2**	**73.3**

Comparison to Previous Work

We obtain competitive results with respect to previous work that relies on the same underlying embeddings or features, as observed in Table 4.2. Recently, Kuribayashi, Ouchi, Inoue, *et al.*, 2019 [160] were able to further improve performance by exploiting contextualized embeddings that look at the whole document, instead of embedding the arguments in isolation, and by making a distinction between argumentative markers and argumentative components. We attempted document-level contextualized embeddings using BERT and were not able to replicate their success[2]. Moreover, we found no significant improvement on the structured prediction models when replacing our BiLSTM encoders with either BERT or document-level BERT. We leave the exploration of an effective way to leverage contextualized embeddings for future work. As for stance prediction, Stab and Gurevych, 2017 [77] identify stances over the resulting structure and obtain a macro F1 of 68.0. Our full models obtain commensurate results, 69.2, 68.4 for ILP and randomized inference, respectively.

[2] ↑We did not experiment with their extended BoW features, nor AC/AM distinction.

4.4.2 Debate Stance Prediction

Dataset

We use a subset of the 4Forums dataset from the Internet Argument Corpus [81], which consist of a total of 418 discussion threads on four political issues, containing 24,658 posts. We use the same splits as Li, Porco, and Goldwasser, 2018 [83]. Most previous work reports accuracy. However, given that labels are highly imbalanced, we also report macro F1.

Learning and Representation

We model the problem as a collective classification task by predicting disagreement between consecutive posts in a given thread. We represented posts using a BERT encoder. For disagreement, we just represented pairs of posts without additional information. We do 5-fold cross validation and report the average performance. For training, we used AdamW, weight decay of 1e-5, a patience of three, and 50 restarts for randomized inference. For local models, we used a learning rate of 5e-5 and for structured models we used a learning rate of 2e-6. For structured learning, we initialize the parameters using the local models. Note that we keep fine-tuning BERT during training.

We test two versions of the model: (1) **Base** includes consistency between node and edge labels, and (2) **AC** adds author constraints enforcing the same stance for all posts by the same author.

Results

Following the argumentation mining use case, we analyze the results across the following dimensions:

Structured Learning

We can see that the performance of all structured models outperforms learning local models and using inference just at prediction time (L+I), both for post stance (Table 4.3) and for disagreement (Table 4.4).

Table 4.3. **Post stance** on 4Forums. A: Abortion, E: Evolution, GM: Gay Marriage, GC: Gun Control

Model	Infer.	A		E		GM		GC	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
Majority		56.8	28.4	65.9	33.0	66.0	33.0	67.9	34.0
Local		66.0	64.3	65.2	54.3	70.0	61.5	68.2	54.1
Base	L+I	71.0	70.4	63.3	59.2	73.6	69.4	66.8	60.2
	ILP	**72.4**	**71.8**	**64.7**	60.6	**75.1**	72.6	70.5	65.8
	AD^3/ILP	**72.4**	**71.8**	63.2	59.4	**75.1**	72.6	69.7	64.3
	AD^3	**72.4**	**71.8**	64.5	**60.7**	**75.1**	72.6	**71.0**	**66.0**
	Rand-C	71.9	71.5	63.1	60.0	75.0	**73.0**	65.4	60.8
	Rand	71.5	71.1	61.3	58.0	74.3	72.0	64.1	60.2
AC	L+I	83.6	84.6	73.3	69.7	84.8	81.9	68.2	60.9
	ILP	87.5	**88.0**	76.1	73.8	**91.2**	**90.3**	74.2	69.9
	AD^3/ILP	86.2	85.8	**76.7**	**73.9**	90.0	89.0	**74.4**	70.7
	AD^3	85.0	84.8	62.7	60.3	87.4	86.3	72.8	67.9
	Rand-C	**87.8**	87.6	**76.7**	73.7	88.9	87.7	73.4	**71.3**
	Rand	86.6	86.4	73.4	70.9	89.9	88.9	72.7	68.8

Randomized vs. ILP/AD3

In the case of stance prediction, there is a significant trend in the performance of the different inference methods. Learning with exact inference generally outperforms the randomized constrained procedure, and the latter outperforms its non-constrained version. The difference is more pronounced in the case of **AC** models. However, we find that relative to its simplicity, the randomized procedures obtain highly competitive performance.

Comparison to Previous Work

Table 4.5 compares our models to previous work on this dataset. Sridhar, Foulds, Huang, *et al.*, 2015 [50] use Probabilistic Soft Logic (PSL) to learn a global assignment for the post labels. They use local classifiers to obtain the input scores to PSL. The main difference between their approach and ours is that we are able to back-propagate the global error back into the classifiers, and we find that it improves performance considerably. Even though we use BERT encoders in our structured procedure, we can see that BERT alone is not able

Table 4.4. Disagreement on 4Forums. A: Abortion, E: Evolution, GM: Gay Marriage, GC: Gun Control

Model	Infer.	A		E		GM		GC	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
Majority		77.8	38.9	66.4	33.2	73.7	36.9	64.3	32.2
Local		76.0	58.1	63.4	56.0	71.3	56.9	67.0	61.4
Base	L+I	70.8	60.3	62.6	58.4	63.3	58.7	61.4	58.9
	ILP	74.2	**62.9**	63.6	**59.6**	**71.5**	61.4	63.8	59.9
	AD^3/ILP	74.2	**62.9**	62.8	58.8	**71.5**	61.4	64.2	61.5
	AD^3	74.2	**62.9**	**64.3**	59.5	**71.5**	61.4	64.2	59.9
	Rand-C	**76.0**	61.6	64.1	57.3	71.2	60.1	64.8	**61.8**
	Rand	**76.0**	60.7	62.7	58.5	70.4	**61.5**	**65.3**	59.4
AC	L+I	83.2	78.7	72.1	70.0	71.0	68.0	68.8	67.0
	ILP	88.0	82.2	**76.5**	**73.6**	**86.1**	**81.5**	75.4	73.6
	AD^3/ILP	86.3	80.5	74.8	72.4	83.9	79.2	72.8	71.5
	AD^3	84.9	76.4	66.8	61.4	84.4	78.4	68.1	65.8
	Rand-C	**88.2**	**82.4**	74.3	71.7	82.5	78.1	**78.5**	**76.3**
	Rand	87.7	81.4	75.0	71.9	84.1	78.7	75.8	74.4

to solve the task. Lastly, we compare to the structured representation learning method of Li, Porco, and Goldwasser, 2018 [83] and find that we are able to improve on abortion and gay marriage only. Note that these two are the issues with more data available (8,000 and 7,000 posts respectively). The main difference with their approach and ours is that they push author profile information into the learned representation. We hypothesize that this is key to obtain good performance for gun control, which contains only 4,000 posts.

Table 4.5. Previous work on 4Forums (Post Acc)
*Note that [83] use author profile information in their models, whereas we only look at text

Model	A	E	GM	GC	Avg
BERT [150]	66.0	65.2	70.0	68.2	67.4
PSL [50]	77.0	80.3	80.5	69.1	76.7
Struct. Rep. [83]*	86.5	**82.2**	87.6	**83.1**	**84.9**
Deep AC ILP	87.5	76.1	**91.2**	**74.2**	**82.3**
Deep AC Rand-C	**87.8**	**76.7**	88.9	73.4	81.7
Deep AC Rand	86.6	73.4	89.9	72.7	80.7

4.4.3 Inference Analysis

In our experiments, randomized inference always outperforms ILP and AD^3 in terms of run-time. Figure 4.3 shows the speedup factor per epoch against ILP and AD^3. In argument mining, AD^3 is faster than ILP, except on our full model, where both perform similarly. We noticed that ILP consumes a lot of time in initialization and encoding. The randomized inference approach is able to predict argumentative structures 9.1x faster than ILP for our base model, and 7.5x faster than AD^3 for our full model. For stance prediction on 4Forums, ILP is considerably faster than AD^3, we presume that this is due to the fact that Gurobi is a highly optimized commercial software, and our graphs are small. Randomized inference is 11x faster than ILP on the base model and beats AD^3 by a factor of 27 when author constraints are used.

We also measured pure inference time over five training runs and took the average. Figure 4.4 shows (in logarithmic scale) plain inference run-time in seconds on the training set

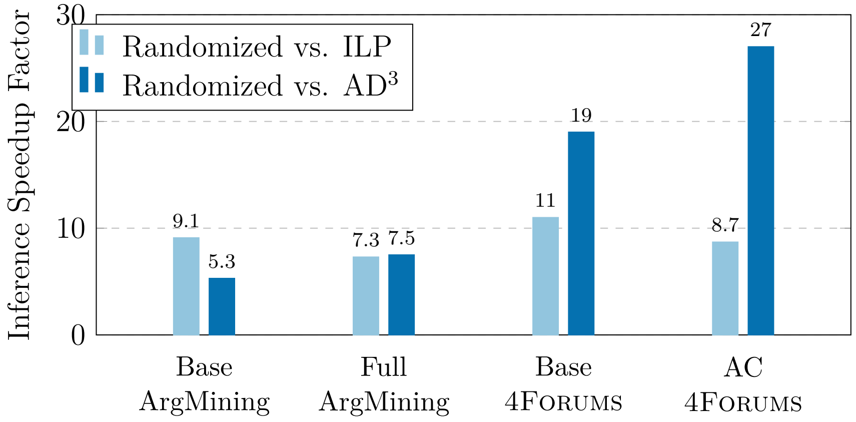


Figure 4.3. Comparing inference speedup per epoch

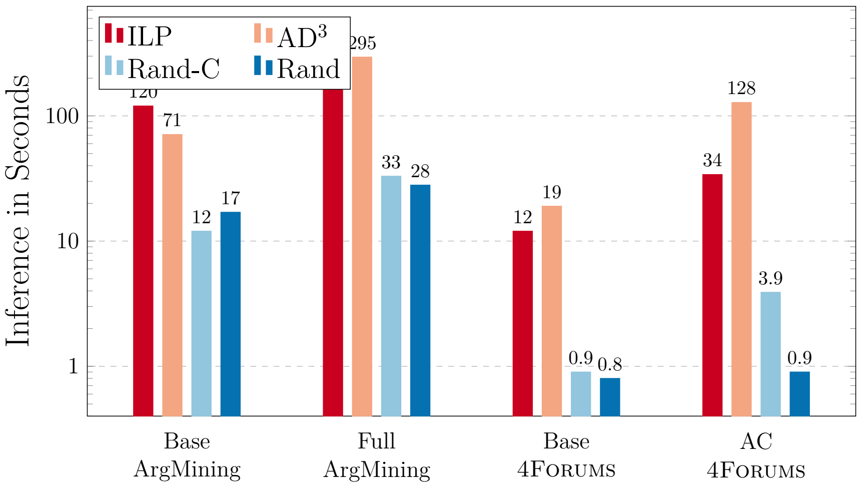
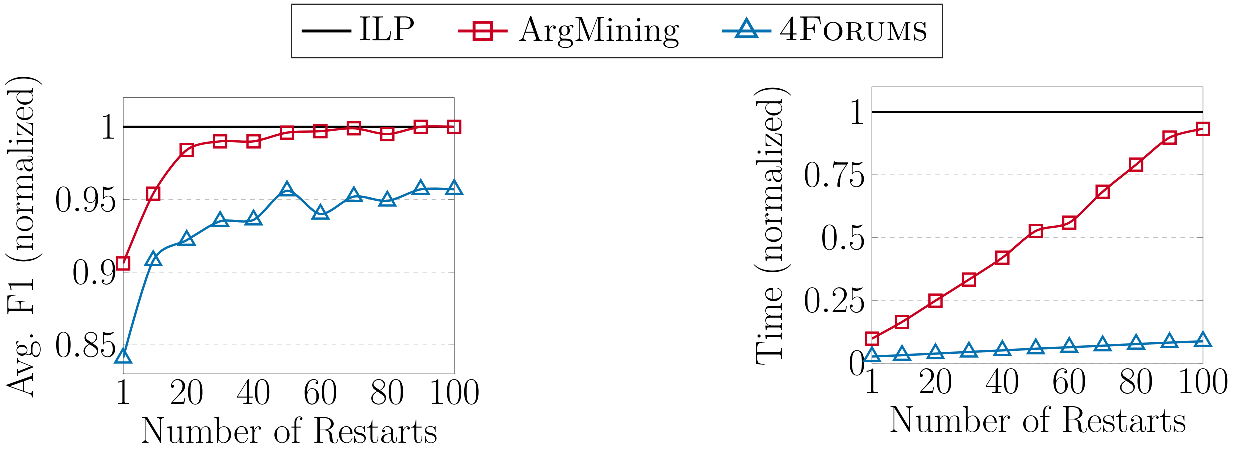


Figure 4.4. Pure inference run-time in seconds

for all of our models. We can observe that randomized inference without domain constraints has almost the same performance as the constrained version. Again, we find that randomized inference considerably outperforms ILP and AD^3.

Additionally, we evaluated our model at test time by replacing exact inference with randomized inference, incrementally increasing the number of restarts. Figure 4.5 shows the performance and runtime of the Rand-C algorithm with respect to exact inference $\left(\text{i.e., } \frac{\text{Rand-C}}{\text{ILP}}\right)$. Figure fig:test-inferencea shows that the global optimum is closely approached after just 20 restarts for the argument mining task, as opposed to stance prediction on 4Forums, where a higher number of restarts is required. This is in line with our reported results in Sec-



(a) (b)

Figure 4.5. Impact of performance and run-time depending on the number of restarts for randomized inference

tions 4.4.1 and 4.4.2. Figure fig:test-inferenceb shows that randomized inference is about twice as fast than ILP when using 50 restarts for the Argument Mining task, and it starts to approach the time needed for ILP after 100 restarts. On the other hand, the randomized algorithm on 4Forums continues to be an order of magnitude faster even when doing 100 repetitions. Note that as the number of restarts keeps increasing, the randomized procedure will eventually surpass the time needed to perform exact inference.

4.5 Summary

We studied the effectiveness of randomized inference for deep structured prediction and obtained positive results for two challenging discourse-level tasks. We showed that, in practice, we can train complex structured models, using expressive neural architectures, and get competitive results at a lower computational cost. Moreover, we saw that combining expressive representations and inference is a promising direction for modeling discourse-level structures. Future directions include expanding the discussion to other tasks involving more complex structures.

5. LEARNING DRaiL MODELS WITH LATENT EXPLANATIONS

In Chapter 3 we motivated the use of neural-symbolic methods to model natural language scenarios involving multiple inter-dependent decisions, and showed the advantages of exploiting their complementary strengths. In the cases that we studied, we assumed that annotations were available for all the different decisions. However, meaningful linguistic and contextual structures are sometimes hidden or unavailable to practitioners.

One way in that traditional graphical models deal with hidden structure is to represent it using latent variables. In statistics and machine learning, latent variables are variables that are not directly observed but inferred, through a model, from other variables that are observed or predicted. For example, Goldwasser and Zhang, 2016 [165] analyze satirical articles by assigning latent categories to the entities and actions expressed in the text, then these categories are mapped to a decision on whether the article is satirical or not. The motivation was to group similar entities and actions, and capture patterns regardless of specific word use. While there is no supervision for these categories, the latent variable model commits to an explanation for a decision, and forces consistency between decisions and explanations. In the context of natural language processing, latent variables have proven useful for a wide range of tasks, including semantic parsing [166], document clustering and topic modeling [167], machine translation [168], common-sense inferences [165] and social media analysis [63]. However, most of this work has been done using traditional probabilistic graphical models.

In Chapter 3, we proposed a neural-symbolic framework that combines probabilistic graphical models and neural networks, retaining the ability of directly representing dependencies between variables in the output space, while taking advantage of the expressive distributed representations that neural methods afford. In this chapter, we propose to incorporate latent predicates into our framework. By extending DRaiL to handle latent predicates, models will be able to learn distributed representations of the different variables and their context, while being able to connect interdependent decisions, and marginalize over properties and relations that are unobserved at training time. The advantages of latent neural-symbolic models are two-fold: (1) they add inductive bias to the model, helping to constrain the representation learned by the neural modules, and potentially allowing it to generalize better, and (2) they provide an additional interpretation for the model through explicit discrete internal states. This is particularly useful when used in conjunction with neural networks, which are inherently difficult to interpret. We limit our discussion to latent variables corresponding structural dependencies in three scenarios. In first two scenarios, we have a clear idea of what is the missing structure, and we have varying levels of direct supervision for it. In the third scenario, we assume that a structure exists, but the categories and dependencies that comprise it might not be known, and need to be discovered. For example, we might not know what are the most frequent arguments to refuse the COVID-19 vaccine, but we assume that they exist.

Scenario 1: Learning to Explain High-Level Decisions In this scenario, we have knowledge about the properties and constraints of the problem, as well as direct supervision for the high-level decision. However, we do not have any source of supervision for the variables and explanations that influence the high-level decision. For example, we might have a dataset of arguments annotated as pro or anti vaccine. In addition to this, we might know what are the most frequent reasons for people to refuse vaccines, but no direct supervision for them. Using our neural-symbolic framework, we could explicitly represent frequent arguments as latent variables and write down the dependencies that connect them to the supervision that we have (e.g. An argument that exhibits distrust of authoritative figures is more likely to be anti-vaccine.) Following this intuition, we ask the following questions: (1) *Can we represent the missing information as latent variable assignments and connect them to meaningful explanations?* and (2) *Does explicitly modeling the latent structure help us generalize better?*

Scenario 2: Learning Models from Explanations In this scenario, we have knowledge about the properties and constraints of the problem. However, we do not have any form direct supervision for our target domain. For example, we might have a dataset of arguments about the COVID-19 vaccine, and we might know what are the most frequent reasons for people to refuse vaccines, but we do not have any supervision for either the stance or the reasons. Using our neural-symbolic framework, we could exploit the inter-dependencies between the different decisions to guide the learning process and facilitate domain adaptation (e.g. A negative opinion about Fauci is more likely to indicate an anti-vaccine stance). By doing this, we are providing explanations to the model ahead of time to help constrain the output space. Given that we have no supervision for the target domain, we could leverage external and weak sources of supervision. For example, we could use entity-level sentiment datasets like MPQA 3.0. In this scenario, we ask the following questions: (1) *Can we leverage external or weakly annotated resources using a shared distributed representation?* and (2) *Does constraining the different decisions using symbolic explanations help offset the disadvantages of not having direct supervision?*

Scenario 3: Discovering the Space of Possible Explanations In this scenario, we do not have first-hard knowledge about the properties and constraints of the problem, but we know that they exist. For example, we might assume that there are repeated themes around the COVID-19 vaccine debate, but we do not know what they are. We tackle this scenario by leveraging human interaction, and we present it in Chapter 6.

We outline two alternative methodologies for dealing with latent variables in our explanation driven neural-symbolic framework. The first methodology uses the modified structured hinge loss to account for latent variables [169], this methodology is appropriate when there is a mix of predicted and latent variables, some of which we have supervision for, as well as a theory of the way they influence each other (Scenario 1). The second methodology uses an Expectation-Maximization style approach, in which constrained symbolic inference is used in the expectation step to obtain training labels, and local normalization is used in the maximization step to learn neural models. This methodology is appropriate in cases where we do not have direct supervision for any of the relevant variables, but we have a theory of the way they influence each other (Scenario 2). To perform this analysis and demonstrate our framework, we will model two real discourse tasks: identifying collaborative conversations online, and analyzing the COVID-19 vaccine debate.

5.1 Scenario 1: Learning to Explain High-level Decisions

5.1.1 Case Study: Identifying Collaborative Conversations

Online conversations are rampant on social media channels, news forums, course websites and various other discussion websites consisting of diverse groups of participants. While most efforts have been directed towards identifying and filtering negative and abusive content [133], [170], [171], this scenario focuses on characterizing and automatically identifying the positive aspects of online conversations [135], [136], [172]. The main focus is on *collaborative conversations*, which help achieve a shared goal such as gaining new insights about the discussion topic by identifying behaviors like response informativeness and engagement, among others. This case study corresponds to work done in collaboration with Ayush Jain for his Masters thesis [13].

Instead of looking at the outcomes of conversations (e.g., task completion [135]), the analysis presented in this case study is centered on conversational behaviors, specifically looking at indications of *collaborative* behavior that is conducive to group learning and problem-solving. These include purposeful interactions centered around a specific topic, as well as open and respectful exchanges that encourage participants to elaborate on previous ideas. To help clarify these concepts, consider the following conversation snippet.

User A : We should invest in more resources to encourage young people to be responsible citizens.

Response Option 1 : I wonder if more initiatives at grassroots level can help them to identify and understand issues of their local community more deeply.

Response Option 2 : Good point, I agree.

This snippet compares two possible responses to User A's post. Option 1 offers a balanced contribution, developing the idea presented in the original post and allowing the conversation to proceed. Option 2, while polite and positive, is *not* collaborative as the initial idea is not expanded on. In fact, agreement is often used as a polite way to end conversations without contributing additional content. Despite the positive sentiment, capturing the absence of balanced content contribution and the absence of idea development as different discourse behaviors, one can infer that it is not a collaborative conversation.

While humans could tell the two apart by detecting constructive discourse behaviors, automatically capturing these behaviors is highly challenging. Anecdotal evidence, collected by extracting features from conversation transcripts, can lead to conflicting information, as identifying collaborative behavior relies on complex interactions between posts. Our key intuition is that reasoning and enforcing consistency over these behaviors can help capture the conversational dynamics and lead to more accurate predictions.

Our technical approach follows this intuition. We design a hybrid relational model that combines neural networks and declarative inference to capture high-level discourse behaviors. Since we only have access to the raw conversational text, we model these behaviors as discrete latent variables, used to support and justify the final decision – whether the conversation is collaborative or not.

Explicitly modeling discourse behaviors as latent variables allows us to add inductive bias, constraining the representation learned by the neural model. It also provides a natural way to "debug" the learning process, by evaluating the latent variables activation. Our experiments show that the joint model involving global learning of different latent discourse behaviors improves performance. We use the Yahoo News Annotated Comments Corpus napoles2017ynacc, which was expanded by Jain, 2020 [13] for the collaborative task.

5.1.2 Task Definition

Collaborative conversations are purposeful interactions, often revolving around a desired outcome, in which interlocutors build on each others' ideas to help move the discussion forward. Collaborative conversations are an important tool in collaborative problem solving [173] and require collaboration skills [174], [175]. We focus on identifying indicators of successful collaboration. The task builds on the work of Napoles, Pappu, and Tetreault, 2017 [136], who released a dataset annotated for engaging, respectful and informative conversations. Jain, 2020 [13] further annotated it for collaborative conversations, in which participants build on each other's words, provide constructive critique, and elaborate on suggested ideas, generalizing them and synthesizing new ideas and knowledge in the process. The annotations correspond to binary labels (i.e. collaborative or non-collaborative). During the annotation process, several repeating behaviors that helped characterize and separate between collaborative and non-collaborative conversations were identified, and annotated for a small subset of the examples. The resulting set of behaviors are outlined below.

Non-Collaborative Discourse Behaviors

(A) **Low Idea Development:** users who: (1) deviate from the thread topic and change the topic, (2) ignore previously raised ideas and give preference to their own, (3) repeat or reinforce previous viewpoints.

(B) **Low User Engagement:** users who: (1) show little interest, (2) add shallow contributions, such as jokes or links.

(C) **Negative Sentiment:** relevant when disagreements are not resolved politely and respectfully.

(D) **Rudeness:** use of abusive, rude or impolite words.

Collaborative Discourse Behaviors

(A) **High Idea Development:** when users stay on topic (with respect to the original post) and new ideas are formed and developed based on preceding turns.

(B) **Reference to Previous Posts:** users refer to the previous post to advance the conversation.

(C) **Back and Forth:** users support and appreciate the ideas shared by others, and are polite when expressing disagreements.

(D) **Positive Sentiment:** resulting in positive interactions among users, expressed through polite conversation or informal emoticons.

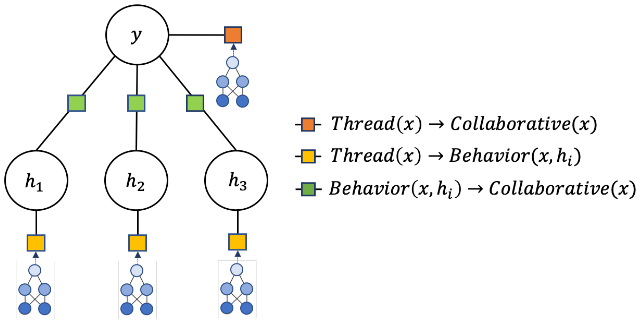

y
h_1
h_2
h_3
🟧 $Thread(x) \rightarrow Collaborative(x)$
🟨 $Thread(x) \rightarrow Behavior(x, h_i)$
🟩 $Behavior(x, h_i) \rightarrow Collaborative(x)$

Figure 5.1. Factor Graph for Collaborative Conversations

(E) **High User Engagement:** User engagement leading to insightful discussions, meaningful to its participants.

(F) **Balanced Content Distribution:** between all members in the group.

(G) **Questions:** raised by participants to advance the conversation.

5.1.3 Modeling Approach

Identifying collaborative conversations requires characterizing the nuanced behaviors outlined in the previous section. In previous work, this analysis was defined by extracting social and discourse features directly from the raw data. In contrast, we view this decision as a probabilistic reasoning process over the relevant conversational behaviors. Since these behaviors are not directly observed, and have to be inferred from the raw conversational features, we treat them as discrete latent variables which are assigned together-with, and consistent-with, the final classification task.

Formally, each behavior is captured by a binary latent variable, denoted as $\mathbf{h} = \langle h_1, ..., h_k \rangle$, indicating if it's active or not in the given thread. These decisions are then connected with the final prediction, denoted y, a binary output value. This results in a factor graph (Figure 5.1).

We model the problem in DRAIL. Initially, we define the rule:

$$\texttt{Thread(T)} \Rightarrow \texttt{IsCollaborative(T)} \tag{5.1}$$

Mapping the thread to the predicted value directly. Then, we augment the program with rules capturing individual discourse behaviors, and associate the predictions of these rules with the final prediction task.

To do this, we define the set of latent conversational behaviors B ∈ *{Idea Development, Reference to Previous Post, Sentiment, Balanced Content, Back and Forth, Questioning Activity, User Engagement, Rudeness and Controversial}*, and define two rules for each behavior in B, as follows:

$$\begin{aligned} &\texttt{Thread}(\texttt{T}) \Rightarrow \texttt{LatentBehavior}(\texttt{T},\texttt{B}) \\ &\texttt{LatentBehavior}(\texttt{T},\texttt{B}) \Rightarrow \texttt{IsCollaborative}(\texttt{T}) \end{aligned} \tag{5.2}$$

Corresponding to prediction of the specific latent behavior B in conversational thread T, and capturing the relationship between the latent behavior and the collaborative prediction.

5.1.4 Learning Approach

To deal with the challenges of this scenario, we extend DRaiL to handle latent predicates. In Chapter 3 we defined a DRaiL program as a set of weighted rules written as horn-clauses. Each rule r, with input features $\boldsymbol{x_r}$ and open atoms $\boldsymbol{y_r}$, defines the potential $\psi_r(\boldsymbol{x_r}, \boldsymbol{y_r})$ representing dependencies between the open atoms in the body and head (See Chap. 3 for more details). Then, DRaiL finds the optimally scored assignments for open atoms by performing MAP inference over the set of all potentials Ψ and the set of all constraints C

$$\begin{aligned} \arg\max_{\boldsymbol{y}\in\{0,1\}^n} P(\boldsymbol{y}|\boldsymbol{x}) \equiv &\arg\max_{\boldsymbol{y}\in\{0,1\}^n} \sum_{\psi_{r,t}\in\Psi} w_r \ \psi_r(\boldsymbol{x_r}, \boldsymbol{y_r}) \\ &s.t. \ c(\boldsymbol{x_c}, \boldsymbol{y_c}) \leq 0; \ \ \forall c \in C \end{aligned}$$

Learning was done using the structured hinge loss:

$$\max(0, \max_{\hat{\mathbf{y}} \in Y}(\Delta(\hat{\mathbf{y}}, \mathbf{y}) + \sum_{\psi_r \in \Psi} \Phi_t(\boldsymbol{x_r}, \boldsymbol{\hat{y}_r}; \theta^t)) - \sum_{\psi_r \in \Psi} \Phi_t(\boldsymbol{x_r}, \boldsymbol{y_r}; \theta^t) \tag{5.3}$$

Where $\boldsymbol{\theta} = \{\theta^t\}$ is a set of parameter vectors, and $\Phi_t(\boldsymbol{x_r}, \boldsymbol{y_r}; \theta^t)$ is the scoring function defined for potential $\psi_r(\boldsymbol{x_r}, \boldsymbol{y_r})$, $\hat{\mathbf{y}} \in Y$ corresponds to a given prediction resulting from the MAP inference procedure and $\mathbf{y} \in Y$ corresponds to the gold structure.

To incorporate learning with latent predicates, we now define the set $\boldsymbol{h_r}$ of latent atoms for each rule r, and re-define the structured hinge loss as:

$$\max(0, \max_{\hat{\mathbf{y}}, \hat{\mathbf{h}} \in Y}(\Delta(\hat{\mathbf{y}}, \mathbf{y}) + \sum_{\psi_r \in \Psi} \Phi_t(\boldsymbol{x_r}, \boldsymbol{\hat{h}_r}, \boldsymbol{\hat{y}_r}; \theta^t)) - \max_{\mathbf{h}}(\sum_{\psi_r \in \Psi} \Phi_t(\boldsymbol{x_r}, \boldsymbol{h_r}, \boldsymbol{y_r}; \theta^t)) \tag{5.4}$$

In this case, we perform MAP inference twice: once predicting both open and latent atoms (left-hand side), and a second one by fixing the open atoms with their observed values and predicting just latent atoms (right-hand side).

To specify which predicates are latent, we include a reserved word in DRaiL. Then, writing rules that use latent predicates is as simple as:

$$\begin{aligned} &\texttt{latent} : \texttt{LatentPred} \\ &\texttt{ObservedPred(X)} \Rightarrow \texttt{LatentPred(X)} \\ &\texttt{LatentPred(X)} \Rightarrow \texttt{TargetPred(X)} \end{aligned} \tag{5.5}$$

And learn the model by connecting the latent predicates to other decisions. Note that we can write any dependencies or constraints over these predicates (See Chapter 3 for supported operations and syntax).

Table 5.1. Features per Behavior

Behavior	**Features**
Sentiment	Degree of sentiment and intensity
Balanced Content	Sentences per post, words per posts, post depth
Controversial	Upvote/downvote ratio, $u - d$, $u + d$, $u/(u + d)$
Reference to Previous Posts	2nd per. pronouns, quotes of prev. posts, @username tags
Back and Forth	(Dis)agreement markers, content indicators, post references
Idea Flow	Lexical chains [176]
Rudeness	Profanity, bad words, short posts indicators
User Activity	Number of posts, number of threads
Questioning activity	Question marks, question forms, question types

5.1.5 Experimental Evaluation

Dataset

The dataset consists of 2,327 conversations from the Yahoo News Annotated Comments Corpus [137] annotated as "collaborative/non-collaborative" by three annotators with 81% inter-annotator accuracy [13]. We use 2,130 conversations for training, 97 for validation and 100 for testing. The data is imbalanced, with more conversations being non-collaborative (64%, 69% and 67% for training, validation and testing, respectively). Additionally, fine-grained discourse behaviors are annotated for a sample set of 103 conversations.

Experimental Settings

Each rule template is associated with an initial feature representation and a neural architecture to learn its scoring function. We used feedforward networks for all rules, with one hidden layer and a softmax on top. All hidden layers use sigmoid activations. The number of hidden units are: 400 for the local rule, 50 for idea flow and 100 for all remaining behaviors. We used a learning rate of 0.01. All of these parameters, as well as the weights for the different rules, were tuned using the validation set. The features used are outlined in Table 5.1.

Results

We compare the model that explicitly reasons about conversational behaviors and their relationships (*DRaiL Latent*), with a local neural model that predicts whether a conversation is collaborative or not by using all discourse features as inputs to a single rule (*NN Local*). To motivate the use of neural networks, we include two *Linear SVM* baselines, one that uses bag-of-words and one that uses the set of all discourse features (Table 5.1). These results, outlined in Table 5.2, demonstrate the advantage of modeling competing discourse behaviors as latent variables and making a joint decision using inference, as opposed to just representing them using input features.

Table 5.2. Predicting Collaborative Conversations (Fixed splits)

Model	Prec.	Rec.	F1
Linear SVM(BoW)	0.60	0.58	0.59
Linear SVM(BoW + disc.)	0.63	0.61	0.62
NN Local	0.65	0.64	0.64
DRaiL Latent	**0.69**	**0.68**	**0.69**

We conduct an additional experiment to evaluate the quality of the predicted latent behaviors. To do this, we use the subset of annotated behaviors, and evaluate the activations produced by our latent variable model. We compare their correctness **before** learning (based on initialization parameters) and **after** learning. Inference is used in both cases. Table 5.3 describes the results. We can see that performance consistently improved after global training compared to the initialization point, a clear indication of the connection between the latent information and the predicted conversational outcome. Identifying rude behaviors yields the highest F1 score (0.62), which can be expected as the decision relies on lexical information (negative and abusive words). Similarly, it is relatively easy to identify balanced content behavior, given that structural features (outlined in table 5.1) are very informative. Lexical chains, representing the repeated occurrence of a single word or of several closely related words over the course of a post [176], are also successful at capturing idea flow behaviors. However, controversial and back and forth behaviors are more challenging.

Table 5.3. Predicting Individual Latent Behaviors on Annotated Sample Set Before and After Global Learning

Individual Behavior	F1 (**before**)	F1 (**after**)
Idea Flow	0.371	0.574
Controversial	0.390	0.420
Balanced Content	0.541	0.610
Sentiment	0.462	0.548
User Activity	0.521	0.570
Reference to Previous Posts	0.299	0.427
Questioning Activity	0.427	0.511
Rudeness	0.514	0.620
Back and Forth	0.470	0.520

Laslty, we performed an ablation study to see if the global model is driven by any particular discourse behavior (Table 5.4). We observe that performance drops significantly if the sentiment behavior is removed. Just using rules related to idea flow, sentiment and balanced content behaviors leads to an F1 score of 0.62.

Table 5.4. Ablation Study

Model	Precision	Recall	F1
All	0.690	**0.680**	**0.687**
All except Sentiment	0.483	0.495	0.489
All except Idea Flow	0.635	0.554	0.591
All except Balanced Content	0.581	0.593	0.586
All except Questioning Activity	0.578	0.588	0.582
Idea Flow + Sentiment + Balanced Content	0.645	0.607	0.625
Idea Flow + Sentiment + User Activity	0.665	0.404	0.502
Sentiment + Balanced Content + Controversial + Questioning Activity	**0.693**	0.546	0.610

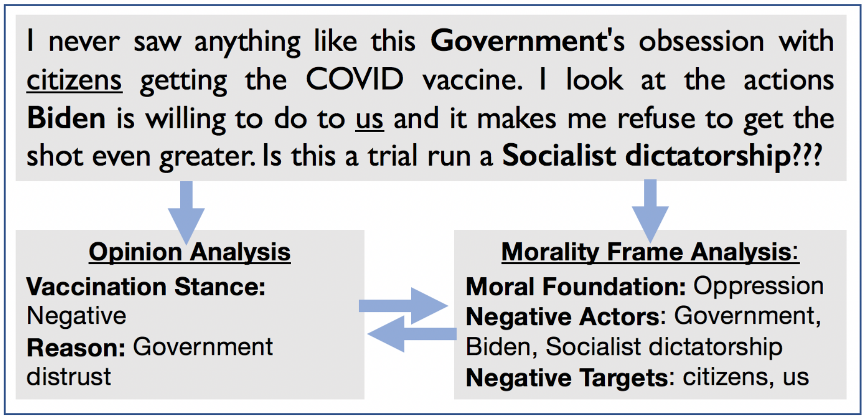
5.2 Scenario 2: Learning from Symbolic Explanations

5.2.1 Case Study: Analyzing The COVID-19 Vaccine Debate

One of the unfortunate side-effects of the COVID-19 pandemic is a global infodemic flooding social media with low quality and polarizing information about the pandemic, influencing its perception and risks associated with it [177]. As studies have shown [178], these influences have clear real-world implications, in terms of public acceptance of treatment options, vaccination and prevention measures.

Most computational approaches tackling the COVID-19 infodemic view it a misinformation detection problem, i.e., identifying false claims and analyzing reactions to them on social media [179]–[181]. This approach, while definitely a necessary component in fighting the infodemic, does not provide policy makers and health-professionals with much needed information, characterizing the reasons and attitudes that underlie the health and well-being choices individuals make.

In this scenario, we tackle a holistic analysis of the COVID-19 vaccination debate, providing multiple interconnected views of the opinions expressed in text. Figure 5.2 describes an example of the intended output. Our analysis identifies the *stance* expressed in the post



I never saw anything like this **Government's** obsession with citizens getting the COVID vaccine. I look at the actions **Biden** is willing to do to us and it makes me refuse to get the shot even greater. Is this a trial run a **Socialist dictatorship**???

Opinion Analysis
Vaccination Stance: Negative
Reason: Government distrust
Morality Frame Analysis:
Moral Foundation: Oppression
Negative Actors: Government, Biden, Socialist dictatorship
Negative Targets: citizens, us

Figure 5.2. Holistic Analysis Framework of Social Media Posts, Connecting entity-level Moral Perspectives, Stance and Arguments Justifying it.

(anti-vaccination) and the *reason* for it (distrust of government). Given the ideologically polarized climate of social media discussion on this topic, we also aim to characterize the moral attitudes expressed in the text (oppression), and how different entities mentioned in it are perceived (``Biden, Government'' are oppressing, ``citizens, us'' are oppressed).

We operationalize this analysis by breaking it down into several classification tasks: stance prediction, reason identification, moral foundation prediction, entity role assignment and entity-level sentiment analysis. Then, we explicitly model the inter-dependencies between them, and build expectations about likely attitudes in the context of each stance. As a motivating example, consider the reason ``distrust in government'', which can be associated with the ``oppression'' moral foundation, only in cases where its actor is an entity related to government functions. We model these expectation as a probabilistic inference process using DRaiL, by incorporating consistency constraints over the judgements made by our model, and predicting jointly the most likely analysis, consisting of all analysis dimensions. The full model, as well as the individual sub-tasks, are explained in detail in Section 5.2.3.

While our analysis in this chapter focuses on a specific issue, vaccination hesitancy, we believe that our analysis framework should be easily adaptable to new issues. Relying on human insight to characterize and operationalize stance and reason identification is one aspect, that characterizes *issue-specific* considerations. Moral Foundation Theory, by its definition abstracts over specific debate topics, and offers a general account for human morality. However, from a practical perspective, models for predicting these highly abstract concepts are trained on data specific to a debate topic and might not generalize well. Instead of retraining the model from scratch, we hypothesize that given an initial model constructed using out-of-domain data, and a small amount of in-domain labeled data, we can obtain acceptable performance by modeling the interaction between reasons, stances and moral foundations. We do this by modeling all variables as latent in a constrained EM-style framework. This learning procedure is explained in detail in Section 5.2.4.

5.2.2 Task Definition

Opinion Analysis

To analyze opinions about the COVID-19 vaccine, we model the vaccination stance expressed in each tweet (i.e. pro-vaccine, anti-vaccine, neutral) and the underlying reason behind such stance. For example, in Fig. 5.2 the tweet expresses an anti-vaccine stance, and mentions their distrust of the Biden administration as the reason to take this stance.

There are three main challenges involved in this analysis: 1) predicting the stance, 2) constructing the space of possible reasons, and 3) mapping tweets to the relevant reasons. Stance prediction is an established NLP classification task [182], and we approach it as such. To uncover reasons, we build on a health informatics study that characterized the arguments made against the COVID-19 vaccine in social media [183]. In this work, researchers come up with a code-book of 12 main themes frequently used as reasons to refuse or cast doubt on the vaccine. Given that these themes are expressed in natural language, we formulate reason identification as a sentence similarity task. To account for pro-vaccine stances, we expand each of these themes with its positive counterpart (e.g. distrust in government → trust in government).

Table 5.5. Moral Foundations

Care/Harm: Underlies virtues of kindness, gentleness, and nurturance.
Fairness/Cheating: Generates ideas of justice, rights, and autonomy.
Loyalty/Betrayal: Underlies virtues of patriotism and self-sacrifice for the group. It is active anytime people feel that it's "one for all, and all for one."
Authority/Subversion: Underlies virtues of leadership and followership, including deference to legitimate authority and respect for traditions.
Purity/Degradation: Underlies religious notions of striving to live in an elevated, less carnal, more noble way. It underlies the widespread idea that the body is a temple which can be desecrated by immoral activities and contaminants.
Liberty/Oppression: The feelings of reactance and resentment people feel toward those who dominate them and restrict their liberty.

Morality Frame Analysis

Moral Foundations Theory [5] suggests that there are at least six basic foundations that account for the similarities and recurrent themes in morality across cultures, each with a positive and negative polarity (See Tab. 5.5).

To analyze moral perspectives in tweets, we build on the definition of morality frames proposed by Roy, Pacheco, and Goldwasser, 2021 [6], where moral foundations are regarded as frame predicates, and associated with positive and negative entity roles.

While Roy, Pacheco, and Goldwasser, 2021 [6] defined different roles types for each moral foundation (e.g. *entity causing harm*, *entity ensuring fairness*), we aggregate them into two general role types: **actor** and **target**, each with an associated polarity (positive, negative). An **actor** is a "do-er" whose actions or influence results in a positive or negative outcome for the **target** (the "do-ee"). For each moral foundation in a given tweet, we identify the "entity doing good/bad" (positive/negative actor) and "entity benefiting/suffering" (positive/negative target). For example, the statement "We are suffering from the pandemic" expresses **harm** as the moral foundation, where "pandemic" is a **negative actor**, and "we" is a **negative target** (i.e. the entity suffering from the actor's actions). There can be zero, one or multiple actors and targets in a given tweet. Entities can correspond to specific individuals or groups (e.g., I, democrats, people of a given demographic), organizations (e.g., political parties, CDC, FDA, companies), legislation or other political actions (e.g., demonstrations, petitions), disease or natural disasters (e.g., Covid, global warming), scientific or technological innovations (e.g., the vaccine, social media, the Internet), among others.

We break down the task of predicting morality frames into four classification tasks. For each tweet, our goal is to predict whether it is making moral judgement or not, and identify its prominent moral foundation. For each entity mentioned in the tweet, we predict whether it is a target or a role, and whether it has positive or negative polarity.

5.2.3 Modeling Approach

We propose a joint probabilistic model that reasons about the arguments made, their morality frames, stances, reasons, and the dependencies between them. We implement our model using DRaiL, and use horn-clauses of the form $p_0 \wedge p_1 \wedge ... \wedge p_n \Rightarrow h$ to describe relational properties. Each logical rule defines a probabilistic scoring function over the relations expressed in its body and head.

Base rules/classifiers: We define three base rules to score whether a tweet $\mathtt{t_i}$ has a moral judgment, what is its prominent moral foundation $\mathtt{m}$, and what is its vaccination stance.

$$\begin{aligned} r_0 &: \mathtt{Tweet(t_i)} \Rightarrow \mathtt{IsMoral(t_i)} \\ r_1 &: \mathtt{Tweet(t_i)} \Rightarrow \mathtt{HasMF(t_i, m)} \\ r_2 &: \mathtt{Tweet(t_i)} \Rightarrow \mathtt{VaxStance(t_i, s)} \end{aligned} \tag{5.6}$$

To score the moral role of an entity $\mathtt{e_i}$ mentioned in tweet $\mathtt{t_i}$, we write two rules. The first one scores whether the entity $\mathtt{e_i}$ is an actor or a target, and the second one scores its polarity (positive or negative).

$$\begin{aligned} r_3 &: \mathtt{Mentions(t_i, e_i)} \Rightarrow \mathtt{HasRole(e_i, r)} \\ r_4 &: \mathtt{Mentions(t_i, e_i)} \Rightarrow \mathtt{EntPolarity(e_i, p)} \end{aligned} \tag{5.7}$$

Note that these rules do not express any dependencies. They function as base classifiers that map tweets and entities to their most probable labels.

Dependency between roles and moral foundations: The way an entity is portrayed in a tweet can be highly indicative of its moral foundation. For example, people are likely to mention *children* as a *negative actor* in the context of *care/harm.* To capture this, we explicitly model the dependency between an entity, its moral role, and the MF.

$$\begin{aligned} r_5 : \mathtt{Mentions(t_i, e_j)} \wedge \mathtt{HasRole(e_i, r)} \\ \wedge\, \mathtt{EntPolarity(e_i, p)} \Rightarrow \mathtt{HasMf(t_i, m)} \end{aligned} \tag{5.8}$$

Dependency between stances and moral foundations: There is a significant correlation between the stance of a tweet with respect to the vaccine debate, and its moral foundation. For example, people who oppose the vaccine are more likely to express the liberty/oppression MF. To capture this, we model the dependency between the stance of a tweet and its MF.

$$r_6 : \mathtt{VaxStance(t_i, s)} \Rightarrow \mathtt{HasMf(t_i, m)} \tag{5.9}$$

Dependency between reasons and moral foundations/stances: Explicitly modeling the dependency between repeating reasons and other decisions can help us add inductive bias into our model, potentially simplifying the task. For example, we can enforce the difference between two opposing views that use similar wording, and that could otherwise be treated similarly by a text-based model (e.g. *"natural methods of protection against the disease are better than vaccines"* vs. *'vaccines are better than natural methods of protection* *against the disease"*). We add two rules to capture this dependency, one between reasons and moral foundations, and one between reasons and stances.

$$\begin{aligned} &r_7 : \texttt{Mentions}(\texttt{t}_\texttt{i}, \texttt{r}) \Rightarrow \texttt{HasMf}(\texttt{t}_\texttt{i}, \texttt{m}) \\ &r_8 : \texttt{Mentions}(\texttt{t}_\texttt{i}, \texttt{r}) \Rightarrow \texttt{VaxStance}(\texttt{t}_\texttt{i}, \texttt{s}) \end{aligned} \tag{5.10}$$

Hard constraints: To enforce consistency between different decisions, we add two unweighted rules (or hard constraints). These rules are not associated with a scoring function and must always hold true. We enforce that, if a tweet is predicted to be moral, then it needs to also be associated to a specific moral foundation. Likewise, if a tweet is not moral, then no MF should be assigned to it.

$$\begin{aligned} &c_0 : \texttt{IsMoral}(\texttt{t}_\texttt{i}) \Rightarrow \neg\texttt{HasMf}(\texttt{t}_\texttt{i}, \texttt{none}) \\ &c_1 : \neg\texttt{IsMoral}(\texttt{t}_\texttt{i}) \Rightarrow \texttt{HasMf}(\texttt{t}_\texttt{i}, \texttt{none}) \end{aligned} \tag{5.11}$$

Whenever the tweets have the same stance, we include a constraint to enforce consistency between the polarity of different mentions of the same entity. roy-etal-2021-identifying showed that enforcing consistency for mentions of the same entity within a political party was beneficial. Given the polarization of the COVID-19 vaccine, we use the same rationale.

$$\begin{aligned} c_3 : &\texttt{Mentions}(\texttt{t}_\texttt{i}, \texttt{e}_\texttt{i}) \wedge \texttt{Mentions}(\texttt{t}_\texttt{j}, \texttt{e}_\texttt{j}) \wedge \texttt{SameVaxStance}(\texttt{t}_\texttt{i}, \texttt{t}_\texttt{j}) \wedge \texttt{EntPolarity}(\texttt{e}_\texttt{i}, \texttt{p}) \\ &\Rightarrow \texttt{EntPolarity}(\texttt{e}_\texttt{j}, \texttt{p}) \end{aligned} \tag{5.12}$$

5.2.4 Learning Approach

To learn DRaiL models in the weak supervision setting, we use an Expectation-Maximization style protocol, outlined in Algorithm 3.

Algorithm 3 *Weak Supervision Learning Protocol*

```
1: Random initialization for all $\boldsymbol{\theta_r}$
2: for $r \in$ base rules do
3:     $\boldsymbol{\theta_r} \leftarrow$ distant supervision classifier
4: end for
5: while not converged do
6:     $\boldsymbol{Y}_{\texttt{gold}} \leftarrow$ DRaiL_MAP_inference(k)
7:     Train all rules locally using $\boldsymbol{Y}_{\texttt{gold}}$
8: end while=0
```

This process consists of two consecutive steps: (1) initializing the parameters of the base rules using weakly or distantly supervised classifiers (lines 2-4), and (2) learning DRaiL models over the target domain (lines 5-8).

Initializing base rules using weak supervision

The first step in Algorithm 3 is to initialize the parameters of the base rules using weakly or distantly supervised classifiers. To do this, we source external datasets for vaccination stance, moral foundations, moral roles and entity-level sentiment.

For moral foundation prediction, we use the dataset proposed by Johnson and Goldwasser, 2018 2018, consisting of 2,000 tweets by US congress members annotated for the five core moral foundations. We also use the Moral Foundation Twitter Corpus [184], consisting of 35,000 tweets annotated for moral foundations. The topics across these two datasets span political issues (e.g. gun control, immigration) and events (e.g. Hurricane Sandy, Baltimore protests). Given that neither of these two datasets contain examples for the *liberty/oppression* moral foundation, we curate a small lexicon by looking for synonyms and antonyms of the words *liberty* and *oppression*. Then, we use this lexicon to annotate the congresstweets dataset [1]. We annotate a tweet as *liberty/oppression* if it contains at least four keywords, which results in around 2,000 tweets. The derived lexicon for liberty/oppression can be seen in Tab. 5.6

To learn to predict roles, we use the subset of the Johnson and Goldwasser, 2018 [110] dataset annotated for roles by Roy, Pacheco, and Goldwasser, 2021 [6], which contains

[1] ↑https://github.com/alexlitel/congresstweets

Table 5.6. Liberty/Oppression Lexicon.

liberty, independence, freedom, autonomy, sovereignty
self-government, self-rule, self-determination, home-rule
civil liberties, civil rights, human rights, autarky,
free-rein, latitude, option, choice, volition, democracy,
oppression, persecution, abuse, maltreatment, ill treatment,
dictator, dictatorship, autocracy, tyranny, despotism,
repression, suppression, subjugation, enslavement,
exploitation, dependence, constraint, control, totalitarianism

Table 5.7. ProVax Hashtags

FullyVaccinated, GetTheVax, GetVaccinatedASAP,
VaccineReady, VaxUpIL, TeamVaccine, GetTheJab,
VaccinesSaveLives, RollUpYourSleeve, DontMissYourVaccine,
letsgetvaccinated, TakeTheVaccine, takethevaccine,
COVIDIDIOTS, SafeVaccines, ThisIsOurShotCA,
LetsGetVaccinated, getthevaccine, GetVaccinated
PandemicOfTheUnvaccinated, VaccineStrategy, igottheshot,
vaccinationdone, ThisIsOurShot, VaccinateNiagara,
TwoDoseSummer, OurVaccineOurPride, IGotMyShot,
FreeVaccineForAll, VaccineEquity, COVIDIOTS, GetTheVaccine,
GetVaxxed, VaccineJustice, getthejab, VaccineForAll,
covidiot, gettheshot, RollUpYourSleevesMN, GoVAXMaryland,
WorldImmunizationWeek, VaccinesWork, getvaccinated,
GetVaccinatedNow, VaxUp, PlanYourVaccine,
VaccinateEveryIndian, TakeYourShot, Vaccines4All,
VaccinnateWithConfidence, firstdose, YesToCOVID19Vaccine,
NYCVaccineForAll, Vaccine4All, getvaxxed, VaccinEquity,

roughly 3,000 tweet-entity-role triplets. For entity sentiment, we combine the Roy, Pacheco, and Goldwasser, 2021 [6] dataset with the MPQA 3.0 entity sentiment dataset [185], which contains about 1,600 entity-sentiment pairs.

For stance, we annotate a dataset of 85,000 unlabeled covid tweets using a set of prominent antivax and provax hashatgs. For the antivax case, we rely on the hashtags proposed by Muric, Wu, and Ferrara, 2021 [186]. For the provax case, we manually annotate hashtags that have a clear provax message, and that are used in at least 50 tweets in the unlabeled dataset. The full set of hashtags used can be found in Tabs. 5.7 and 5.8.

Table 5.8. AntiVax Hashtags

abolishbigpharma, noforcedflushots, NoForcedVaccines,
ArrestBillGates, notomandatoryvaccines,
betweenmeandmydoctor, NoVaccine, bigpharmafia,
NoVaccineForMe, bigpharmakills, novaccinemandates,
BillGatesBioTerrorist, parentalrights, billgatesevil,
parentsoverpharma, BillGatesIsEvil, saynotovaccines,
billgatesisnotadoctor, stopmandatoryvaccination,
billgatesvaccine, cdcfraud, cdctruth, v4vglobaldemo, cdcwhistleblower
vaccinationchoice, covidvaccineispoison, VaccineAgenda
depopulation, vaccinedamage, DoctorsSpeakUp, vaccinefailure,
educateb4uvax, vaccinefraud, exposebillgates, vaccineharm,
forcedvaccines, vaccineinjuries, Fuckvaccines, vaccineinjury
idonotconsent, VaccinesAreNotTheAnswer, informedconsent,
vaccinesarepoison, learntherisk, vaccinescause,
medicalfreedom, vaccineskill, medicalfreedomofchoice,
momsofunvaccinatedchildren, mybodymychoice

Learning DRaiL models

Once the base rules have been initialized using weak supervision, we turn our attention to learning DRaiL models over the target dataset. We alternate between MAP inference using all rules to obtain training labels (expectation step, line 6 in Algorithm 3), and training the neural networks locally using these labels (maximization step, , line 7 in Algorithm 3). We receive an optional parameter k indicating the amount of direct supervision to be used, if there is any direct supervision available. When k is provided, k% of the in-domain labels are seeded during MAP inference.

The main intuition behind this approach is that constraining the output variables with expectations about the way they relate to each-other (e.g. anti-vaccine stances are more likely to evoke the liberty/oppression moral foundation), will help us refine the candidate labels for the target domain. By alternating between inference and learning, we aim to continuously refine the target domain labels and improve the resulting rule weights. In this step, we use all the rules outlined in Section 5.2.3.

5.2.5 Experimental Evaluation

Dataset

There is no existing corpus of COVID-19 vaccine arguments annotated for morality frames and vaccination stance, so we collected and annotated our own.

Dataset Construction

To collect our dataset, we searched for tweets between Apr. and Oct. 2021 mentioning specific keywords. To create the list of keywords, we read multiple articles about COVID mentioning vaccination status, vaccine hesitancy, misinformation, vaccine constraints, health issues, religious sentiment and other vaccine-related debates, and made a list of repeating statements. Then, we consulted three researchers, two in Computational Social Science and one in Psychology, and constructed a list of relevant keywords that are indicative of morally charged discussions. The full list of keywords can observed in Table 5.9.

Table 5.9. List of the keywords for data collection.

covid vaccine, covid vaccination, covid vaccine tyranny, covid vaccine oppression, covid vaccine mandate, covid vaccine conspiracy, covid vaccine anti-vax, covid vaccine religion, covid vaccine satan, covid vaccine god, covid vaccine jesus, covid vaccine islam, covid vaccine muslim, covid vaccine christianity, covid vaccine christian, covid vaccine hindu, covid vaccine jews, covid vaccine catholic, covid vaccine buddhism, covid vaccine religious, covid vaccine biden failure, covid vaccine passport, covid vaccine loyalty, covid vaccine cheating, covid vaccine freedom, covid vaccine betrayal, covid vaccine liberty, covid vaccine black people, covid vaccine propaganda, covid vaccine hesitancy, covid vaccine hesitant, covid vaccine microchip, covid vaccine bill, covid vaccine pregnancy, covid vaccine pregnant, covid vaccine approval, covid vaccine biden, covid vaccine fda, covid vaccine cdc, covid vaccine fauci, covid-19 china, vaccine passport, vaccination mandate, covid vaccine death, covid vaccine military, experimental covid vaccine, covid vaccine authorization, vaccine oppression, vaccine satan, covid vaccine bill gates, covid vaccine side effect, covid vaccine adverse events

Manual Annotation

Moral foundation and vaccination stance labels can be annotated directly for each tweet. To identify entities, annotators are able to highlight the relevant text spans, and choose its role label (i.e. positive/negative actor or target). We annotate our dataset using three in-house annotators pursuing a Ph.D. in Computer Science. We award the annotators \$ 0.75 per tweet and bonus ($2 * \$0.75 = \$1.5$) for completing two practice examples. Our work is IRB approved, and we follow their protocols.

Inter-annotator agreement

We calculate the agreement among annotators using Krippendorff's α [187], where $\alpha = 1$ suggests perfect agreement, and $\alpha = 0$ suggests chance-level agreement. We found $\alpha = 60.82$ for moral foundations, and $\alpha = 78.71$ for stance. For roles, we calculate the character by character agreement between annotations. For example, if one annotator has marked "Dr Fauci" as a target in a tweet, and another has marked "Fauci", it will be considered as an agreement on the characters "Fauci" but disagreement on "Dr". Doing this, we found $\alpha = 83.46$. When removing characters marked by all three annotators as "non-role", the agreement drops to $\alpha = 67.15$.

Resulting annotated dataset

We use a majority vote to get moral foundation and vaccination stance labels, and obtain 750 annotated tweets. Similarly, we define a text span to be an entity mention E, having a moral role R and polarity P, in a tweet T, if it is annotated as such by at least two annotators. Our resulting dataset contains 891 (T,E,R,P) tuples.

Unlabeled COVID-19 vaccine corpus

In addition to our annotated dataset, we collect a corpus of 85,000 tweets in English mentioning the covid vaccine, uniformly distributed between Jan. and Oct. 2021. These tweets are unlabeled, and are used to augment data for indirect supervision (See Section 5.2.4).

Experimental Settings

In DRaiL, each rule r is associated with a neural architecture, which serves as a scoring function to obtain the rule weight w_r. We use BERT-base-uncased [150] for all classifiers. For the rules that model dependencies (Eqs. 5.8, 6.3, 6.4), we concatenate the CLS token with a 1-hot vector of the symbols on the left hand side of the rule (i.e. role, sentiment, stance and reason), before passing it through a classifier. For rules that have the entity on the left-hand side (Eqs. 5.7, 5.8), we use both the tweet and the entity as an input to BERT, using the SEP token. We trained supervised models using local normalization in DRaiL, and leveraged distant supervision using protocol outlined in Alg. 3. In all cases, we used a learning rate of $2\text{e}-5$, a maximum sequence length of 100, and AdamW. In all experiments, we perform 5-fold cross-validation over the annotated dataset and report the micro-averaged results.

Results

Tab. 5.10 shows our general results for morality frames and vaccination stance. We evaluate our base classifiers and show the impact of modeling dependencies using DRaiL. The joint model results in a significant improvement for morality, moral foundation and vaccination stance. For entities, role and polarity remain stable.

Table 5.10. General Results (F1 Scores). NM: Non Moral

Model	**Moral/NM**		**Moral Found.**		**Actor/Target**		**Ent. Polarity**		**Vax Stance**	
	Macro	Weighted	Macro	Weighted	Macro	Weighted	Macro	Weighted	Macro	Weighted
Random	54.96	55.36	11.07	15.15	45.57	45.72	34.63	36.69	49.16	49.23
Majority Class	37.05	43.62	8.33	23.98	34.63	36.69	46.54	58.15	35.77	39.84
Lexicon Matching	58.97	60.01	25.28	35.85	-	-	-	-	-	-
Base (distant sup.)	69.77	68.88	28.79	41.27	71.94	72.05	63.88	74.30	69.46	70.35
Base (direct sup.)	68.94	69.71	35.28	42.92	**84.71**	**84.75**	**72.92**	**84.31**	66.91	67.36
DRaiL	**80.53**	**81.17**	**53.29**	**62.27**	84.60	84.64	71.53	83.35	**72.06**	**72.53**

We show an ablation study in Tab. 5.11. First, we can see how all dependencies contribute to the performance improvement, role-MF being the most impactful. We can also see that explicitly modeling morality constraints improves both the morality prediction and the MF prediction, suggesting an advantage to breaking down this decision. We observe that the stance-polarity constraint does not have a significant impact, but does not hurt performance either, suggesting that our classifiers already capture this information. Lastly, we can see that the performance for roles and polarity remains stable, potentially because these classifiers have a strong starting point.

Table 5.11. Ablation Study (Weighted F1). MC: Morality Constraint, SPC: Stance-Polarity Constraint

Model	**M/NM**	**MF**	**Act/Tar**	**Polar.**
BERT	69.71	42.92	**84.75**	84.31
+RoleMF	69.71	55.54	84.64	84.13
+RoleMF+MC	79.00	57.68	84.64	84.13
+StanceMF	69.71	47.85	84.75	84.31
+StanceMF+MC	72.37	48.63	84.75	84.31
+StanceMF+MC+SPC	72.32	48.63	84.75	**84.35**
+ReasonMF	69.71	53.15	84.75	84.31
+ReasonMF+MC	72.60	53.41	84.75	84.31
+ReasonStance+SPC	69.71	42.92	84.64	83.26
+ ALL	**81.17**	**62.27**	84.64	83.26

Lastly, we evaluate the impact of our indirect supervision protocol by slowly augmenting the amount of direct supervision available and summarize findings in Fig. 5.3. We can see that by leveraging out of domain data and dependencies between variables, we obtain competitive results with just 25% of direct supervision, and we can outperform the fully supervised classifiers using 50% of the annotated labels.

5.3 Summary

We proposed extensions to our neural-symbolic framework, introduced in Chapter 3, to handle discrete latent information. We explored two alternative learning protocols to: (1) leverage the inter-dependencies between different contextualizing variables to adapt to new domains, and (2) discover underlying linguistic and contextual structure to explain higher-

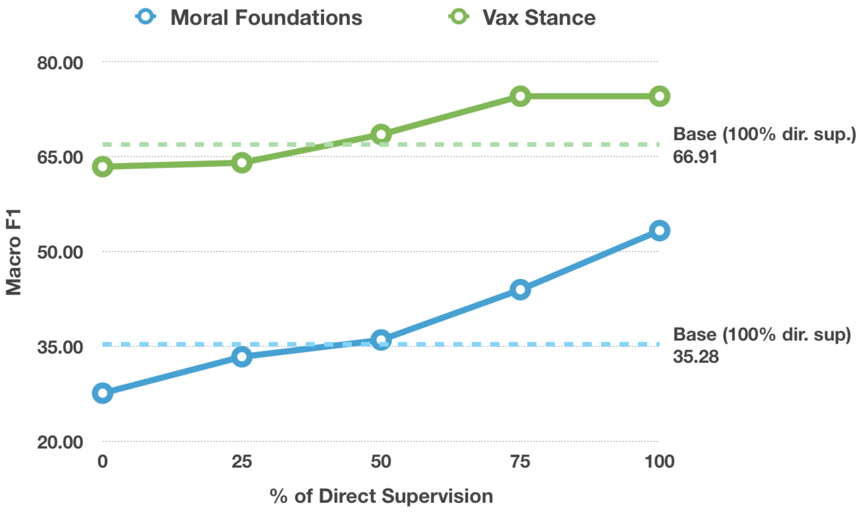


Figure 5.3. Performance in low-supervision settings

level decisions. Moreover, we applied our methods to two challenging discourse scenarios dealing with real-world data.

Given the amount of textual data generated daily, there are broader opportunities for exploiting sources of weak and distant supervision than what we explored in this dissertation. However, we showed that combining distributed representations, distant sources of supervision, and symbolic structured knowledge is a promising direction to both learn to explain high-level decisions and to learn in low-supervision settings.

6. DISCOVERING AND GROUNDING LATENT EXPLANATIONS WITH HUMANS-IN-THE-LOOP

In Chapter 5 we showcased the strengths of neural-symbolic approaches to deal with latent information in language domains with rich language and contextual structure. There are two main challenges that arise when modeling latent information. First, we need to identify the space of possible variables and second, we need to ground these variables in the language data. In the scenarios that we explored in Chapter 5 we made two strong assumptions: (1) That the space of relevant contextual variables was known beforehand, and (2) That we had good mechanisms to ground these variables in text. However, in realistic settings this is not always the case. In this Chapter, we dive deeper into these challenges and explore protocols that leverage human expertise to interactively define the space of relevant variables, and allow experts to improve the grounding of these variables in large language resources.

Human-in-the-loop approaches have been proposed to enable humans to actively participate in the learning process and help debug and refine models [188], [189]. Most existing techniques solicit people to provide feedback on individual predictions [190], or allow people to augment the dataset by providing additional examples for a given label [191]. While straightforward, working in the space of the raw inputs does not take advantage of the ability of humans to make abstractions and reason over them, like forming concepts to generalize from observations to new examples [192], turning raw sensory inputs into high-level semantic knowledge [193], and deductively drawing inferences via conceptual rules and statements [194]. To address these challenges, we suggest an *interactive analysis framework* that leverages multiple interconnected views of the language domain.

We specifically focus on a timely topic, analyzing attitudes explaining *vaccination hesitancy*, introduced in Chapter 5, Section 5.2.1 (See Figure 5.2). Our analysis identifies the *stance* expressed in the post and the *reason* for it. Given the ideologically polarized climate of social media discussion on this topic, we also aim to characterize the *moral attitudes* expressed in the text, and how different *entities* mentioned in it are perceived. The main challenge in this type of analysis is the operationalization of these different abstract analysis dimensions. While stance prediction and morality frame labeling are well-defined, established NLP task, constructing the space of possible reasons justifying stances on a given topic remains an open challenge, traditionally approached using noisy unsupervised techniques such as topic models [7], or by manually identifying and annotating them in text [9]. In Chapter 5 we relied on previous data-driven studies were the space of possible reasons was manually coded by domain experts, who observed repeating themes in the COVID-19 vaccination debate [181], [195]. In this Chapter, we devise a humans-in-the-loop framework that allows users to dynamically explore the space of repeating themes in a large corpus, and provides them with a set of intervention operations to add, remove and name themes, identify good and bad examples for them, and augment them with additional sentences that help characterize the themes better. We tackle the two main challenges outlined above: (1) Interactively grounding statements about the vaccine to a set of known themes, and (2) Interactively discovering the space of latent themes. To evaluate the resulting themes and their groundings, as well as to help users partition and explore the large language resource, we rely on DRaiL, the declarative neural-symbolic framework introduced in Chapter 3. The advantage of using DRaiL is that we can incorporate the supporting analysis dimensions both at inference time to ground emerging themes, as well as a way to provide explanations for different segments of the data during the exploration phase. For example, we can characterize a set of textual arguments by their aggregated statistics on vaccination stance, moral sentiments and salient entity sentiments.

The contents of this Chapter are organized as follows: first, we review the COVID-19 opinion analysis task and introduce the set of exploratory and intervention operations supported by our framework. Then, we formalize our protocol and perform an two-stage analysis. The first stage deals with leveraging human interaction to improve the grounding of themes when there is an initial grouping or partition. The second stage deals with exploring the space of themes when there is no prior knowledge or grouping of the latent themes.

6.1 Case Study: Analyzing the COVID-19 Vaccine Debate

As a case study, we build on the COVID-19 Vaccine analysis framework introduced in Chapter 5. We have a set of tweets mentioning the COVID-19 vaccine, and we want to identify the broader stance (pro-vaccine, anti-vaccine), repeating themes used to justify stances, and the moral sentiments and frames emphasized in each argument. While in the previous Chapter we gave equal importance to the prediction of all the analysis dimensions, in this Chapter we focus on strategies to identify the *repeating themes* that are used to justify stances, as well as identifying the individual tweets that mention them. Stance prediction and morality frame analysis can be defined as traditional classification tasks. However, constructing the space of possible reasons justifying stances on a given topic remains an open challenge. In this Chapter, we present an interactive tool that allows users to *explore* a large repository of tweets to find and identify themes, and allows them to *intervene* by guiding and correcting the system. For this purpose, we use the dataset of 85,000 unlabeled tweets about the COVID-19 vaccine introduced in Chapter 5. All of these tweets were posted in English by users located in the United States, and are uniformly distributed between January and October of 2021. To avoid repetitions, we filter out all retweets ahead of time.

6.2 Interactive Protocol

We propose a simple protocol that combines NLP and machine learning techniques, interactive interfaces and qualitative methods to assist experts in characterizing large language repositories about the COVID-19 vaccine. Our protocol takes a large repository of statements in natural language, and leverages computational techniques to propose an initial partition of the data, such that statements that mention the same theme are clustered together. Then, it provides experts with a set of *exploratory operations* that allows them to further explore and partition the space, and to evaluate the quality of the discovered clusters and the grounded statements. Then, experts have a discussion phase, in which they can follow qualitative techniques to discover patterns. Finally, we provide them with a set of *intervention operations* to name patterns, as well as to provide examples and judgements to improve the quality of the partitions. A diagram outlining this protocol can be observed in Fig. 6.1.

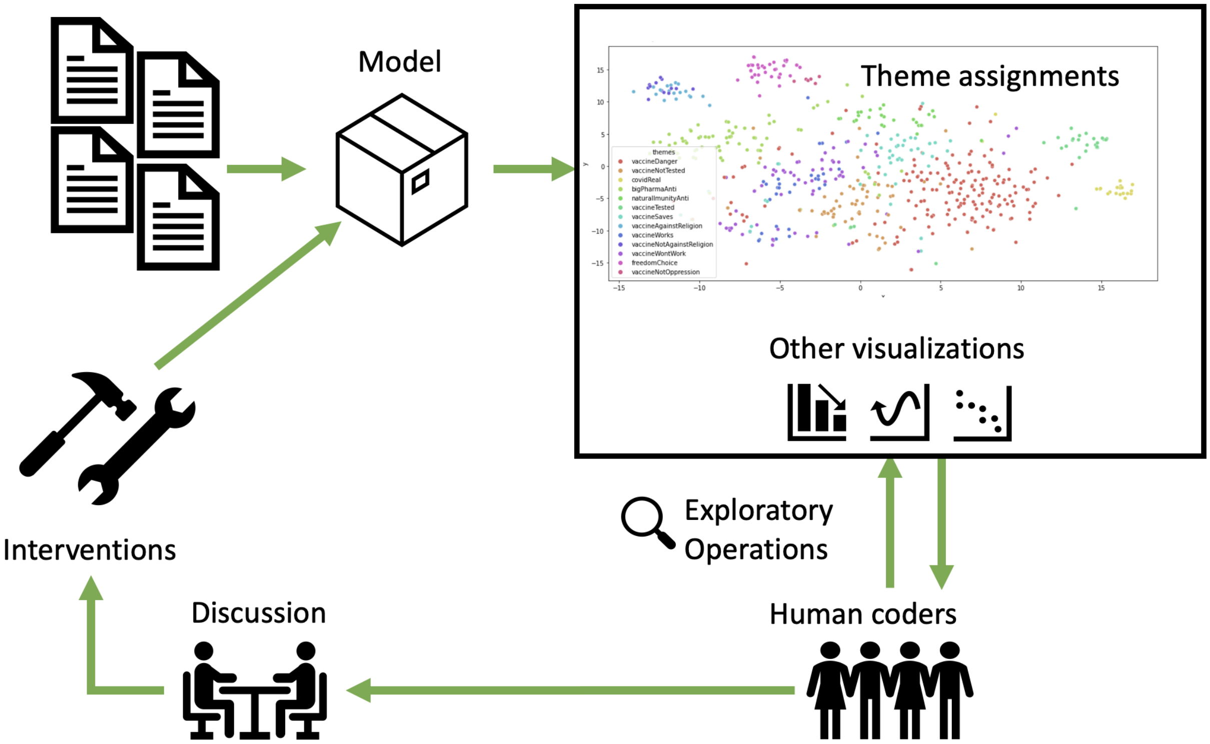


Figure 6.1. Interactive Protocol

In the next section, we present a tool that implements a set of exploratory and intervention operations to support this protocol. Note that our protocol is not necessarily tied to any specific tool, and different tools could follow the same methodology.

6.3 Interactive Tool

To support our interaction protocol, we developed a tool for experts to interact with the language resource. This tool is a simple GUI equipped with a finite set of exploratory and intervention operations for users to explore the domain, discover clusters and introduce knowledge and judgements into the system. In the following sections, we explain the representations used for the textual instances and themes, as well as the exploratory and intervention operations.

6.3.1 Representing Themes and Instances

In our experiments, we represent example instances using their Sentence BERT [196] embeddings. We represent themes using a handful of explanatory phrases and a small set of examples, and also calculate their SBERT embeddings. To measure the closeness between an instance and a theme, we compute the cosine similarity between the instance and all of the explanatory phrases and examples for the theme, and take the maximum similarity score among them. Note that our tool and the operations presented are agnostic of the representation used. The underlying embedding objective, as well as the "closeness" scoring function can be easily replaced.

6.3.2 Finding, Defining and Grounding Themes

Users can either manually create themes by providing a name and a small set of explanatory phrases and examples, or they can automatically partition the space of textual instances to find clusters of similar phrases. In the latter case, users are able to name the clusters and manually select example instances that represent it. We experiment with two grounding mechanisms: a Nearest Neighbors approach that assigns instances to its closest theme, and an structured inference approach that leverages DRaiL. More details about these grounding mechanisms are provided in Sections 6.4.1 and 6.4.2.

6.3.3 Exploratory Operations

Exploratory operations allow experts to inspect the current state of the system, both to evaluate the quality of the theme grounding, as well as to explore the data space and discover new emerging themes. We divide exploratory operations in two types: discovery operations and quality assurance operations.

Discovery Operations

These operations allow users to explore the space of textual instances and get a sense of what themes emerge in the data. We describe each of them below.

Finding Clusters

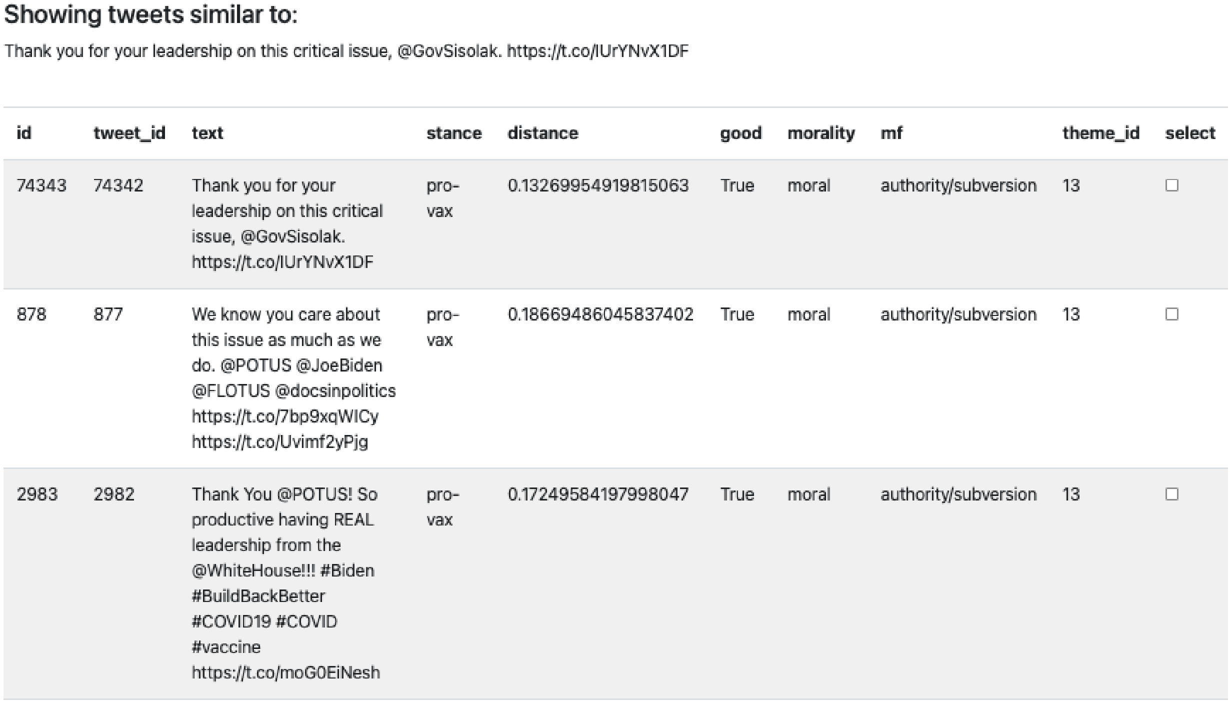
We allow users to find clusters in the space of unassigned instances. To do this, we run a clustering algorithm using the instance representations described in Section 6.3.1. In the experiments presented here, we use the K-means clustering algorithm [197]. However, our protocol is agnostic of the clustering algorithm used.

Performing Text-based Queries

To allow users to expand their search, we provide a text-based query operation. Users can then type any query in natural language and find examples that are close to the query in the embedding space. A screenshot of this functionality is shown in Figure 6.3.

Finding Similar Instances

Once grounded examples are shown by executing either of the two operations above, users have the ability to select each example and find other examples that are close in the embedding space. A screenshot of this functionality is shown in Figure 6.2



Showing tweets similar to:

Thank you for your leadership on this critical issue, @GovSisolak. https://t.co/lUrYNvX1DF

id	tweet_id	text	stance	distance	good	morality	mf	theme_id	select
74343	74342	Thank you for your leadership on this critical issue, @GovSisolak. https://t.co/lUrYNvX1DF	pro-vax	0.13269954919815063	True	moral	authority/subversion	13	□
878	877	We know you care about this issue as much as we do. @POTUS @JoeBiden @FLOTUS @docsinpolitics https://t.co/7bp9xqWICy https://t.co/Uvimf2yPjg	pro-vax	0.18669486045837402	True	moral	authority/subversion	13	□
2983	2982	Thank You @POTUS! So productive having REAL leadership from the @WhiteHouse!!! #Biden #BuildBackBetter #COVID19 #COVID #vaccine https://t.co/moG0EiNesh	pro-vax	0.17249584197998047	True	moral	authority/subversion	13	□

Figure 6.2. Finding Similar Instances

Quality Assurance Operations

These operations allow users to evaluate the quality of the discovered clusters and the grounded instances. We describe each of them below.

Listing Themes and Grounded Examples

We allow users to browse the current list of themes and their grounded examples. Examples are ranked in order of "goodness", corresponding to the similarity in the embedding space to the theme representation. Users can choose explore from closest to most distant, or from most distant to closest. Screenshots of this functionality are provided in Figures 6.3 and 6.4.

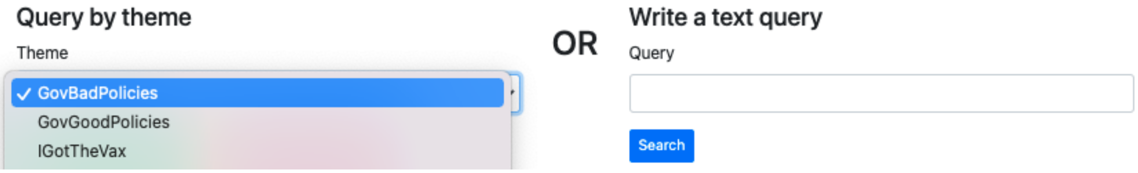
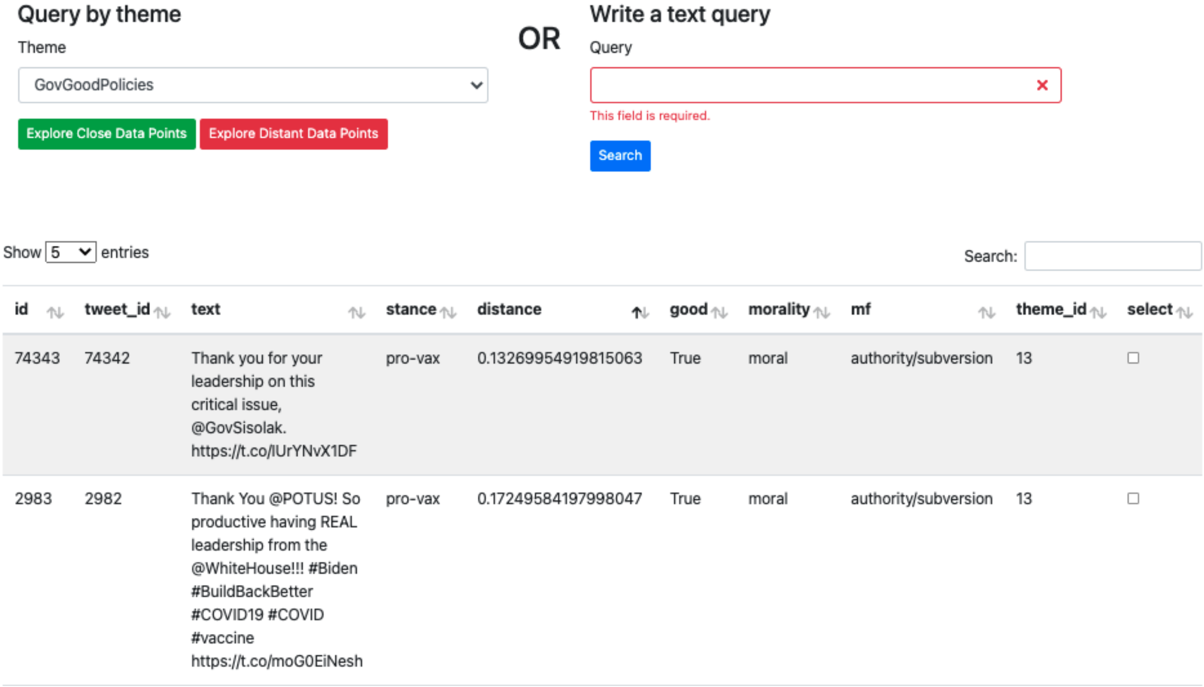


Figure 6.3. Querying Instances



id	tweet_id	text	stance	distance	good	morality	mf	theme_id	select
74343	74342	Thank you for your leadership on this critical issue, @GovSisolak. https://t.co/lUrYNvX1DF	pro-vax	0.13269954919815063	True	moral	authority/subversion	13	□
2983	2982	Thank You @POTUS! So productive having REAL leadership from the @WhiteHouse!!! #Biden #BuildBackBetter #COVID19 #COVID #vaccine https://t.co/moG0EiNesh	pro-vax	0.17249584197998047	True	moral	authority/subversion	13	□

Figure 6.4. Listing Instances From Closest to Most Distant

Visualizing Local Theme Explanations

Our tool allows users to visualize aggregated statistics and explanations for each of the grounded themes. To obtain these explanations, we aggregate all instances that have been identified as being associated with a theme. We experiment with different grounding techniques, which are further explained in Sections 6.4.1 and 6.4.2. We support the following explanations:

Word Clouds: We render a word cloud to visualize the most common linguistic patterns present in a given theme. To do this, we extract bigram and trigram TF-IDF features. TF-IDF stands for Term-Frequency Times Inverse Document-Frequency, and it is used to scale down the raw frequencies of lexical patterns that occur across many instances. Word clouds show high TF-IDF bigrams and tigrams in larger fonts. An example of a word cloud for the theme *The Vaccine Doesn't Work* can be observed in Figure 6.5.

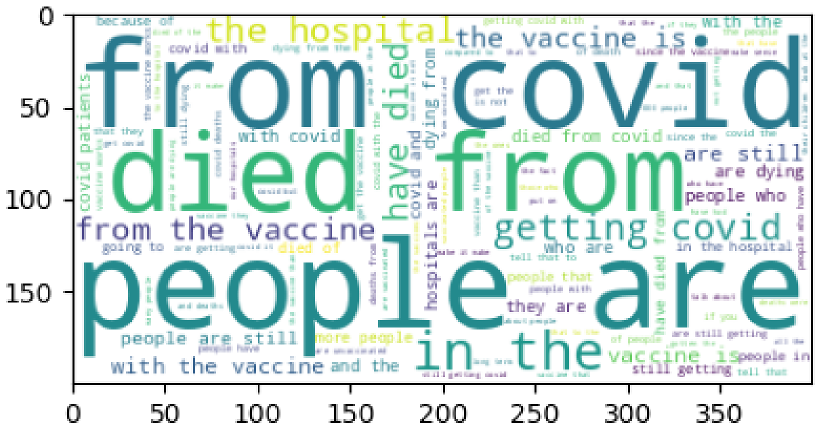
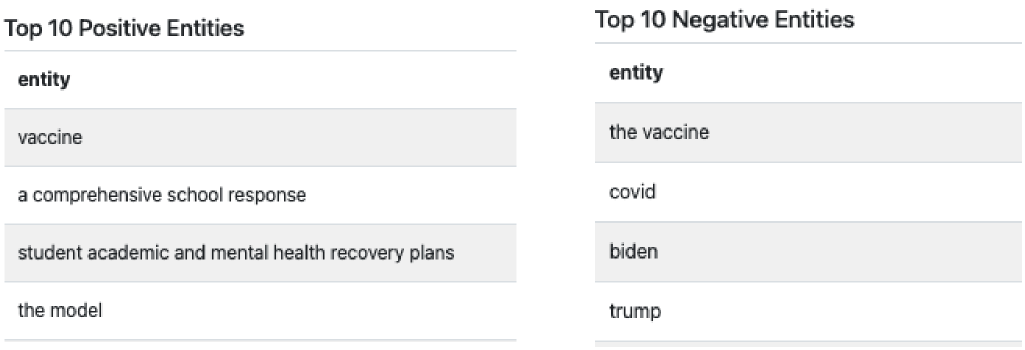


Figure 6.5. Word Cloud Example for *The Vaccine Doesn't Work*

Top Positive and Negative Entities: For each theme, we show the most frequent positive and negative entities. To group entities, we rely on exact lexical matching. An example of top positive and negative entities for the theme *Bad Governmental Policies* can be seen in Figure 6.6.



Top 10 Positive Entities

entity
vaccine
a comprehensive school response
student academic and mental health recovery plans
the model

Top 10 Negative Entities

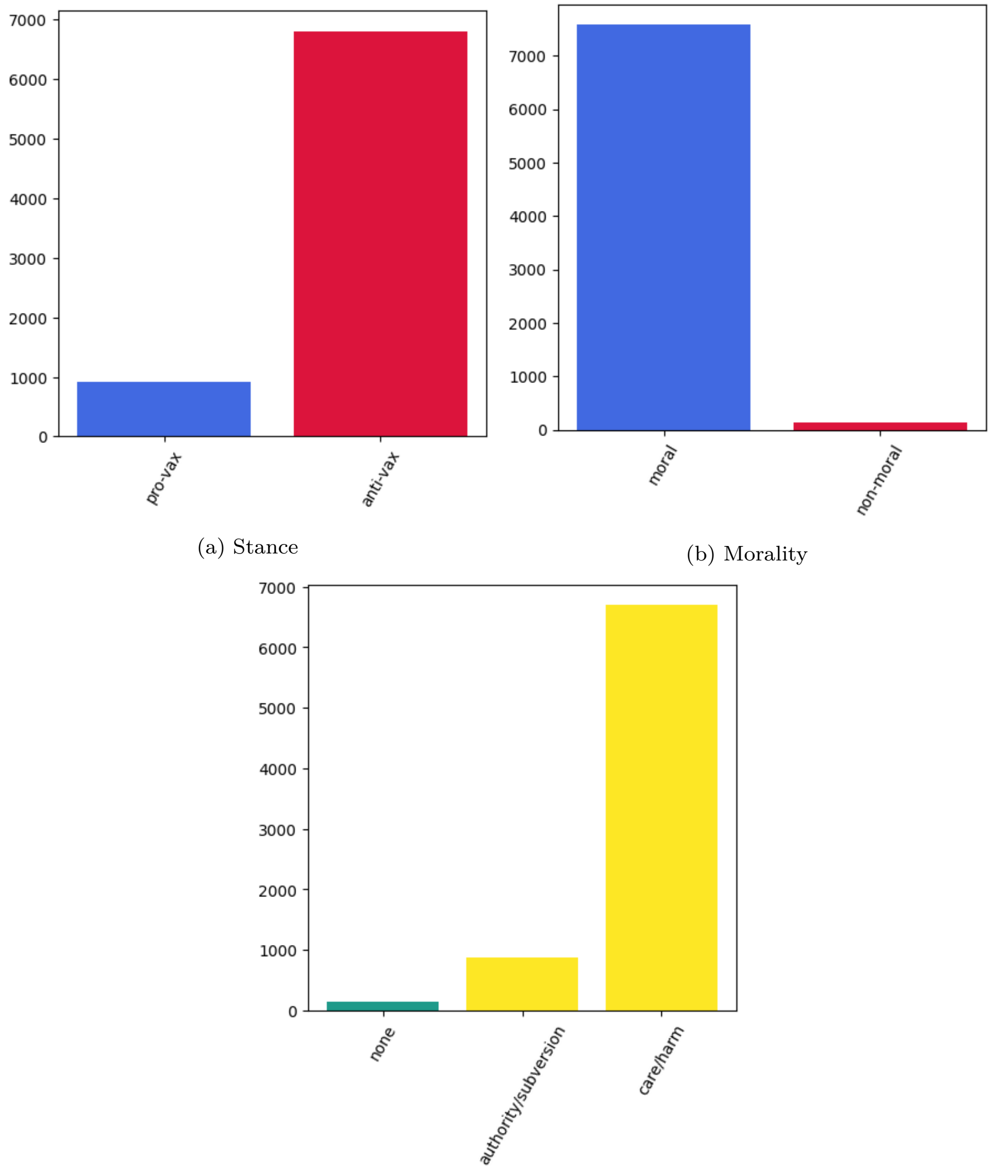
entity
the vaccine
covid
biden
trump

(a) Top Positive Entities

(b) Top Negative Entities

Figure 6.6. Most Frequent Positive and Negative Entities for Theme *Bad Governmental Policies*

Stance, Morality and Moral Foundation Distributions: We allow users to upload additional observed or predicted attributes for each textual instance. Given the COVID-19 domain, we focus on *stance* with respect to the vaccine (i.e. pro, anti), *morality* (i.e. moral, non-moral) and *moral foundations*. To initialize these attributes, we use the best performing models obtained in Chapter 5 and obtain predictions for the 85,000 unlabeled instances. This way, users can visualize the distribution of predicted attributes for each theme. Figure 6.7 shows examples of attribute distributions for the theme *The Vaccine Doesn't Work*.



(a) Stance

(b) Morality

(c) Moral Foundation

Figure 6.7. Stance, Morality and Moral Foundation Distribution for Theme *The Vaccine Doesn't Work*

Visualizing Global Explanations

Our tool allows users to visualize aggregated statistics and explanations for the global state of the system. To do this, we aggregate all instances in the database. We support the following explanations

Theme Distribution: We render a bar plot to visualize the number of instances that have been assigned to each theme. We include an entry *"Unknown"* for the instances that have not been assigned to any theme. An example of this visualization is presented in Figure 6.8.

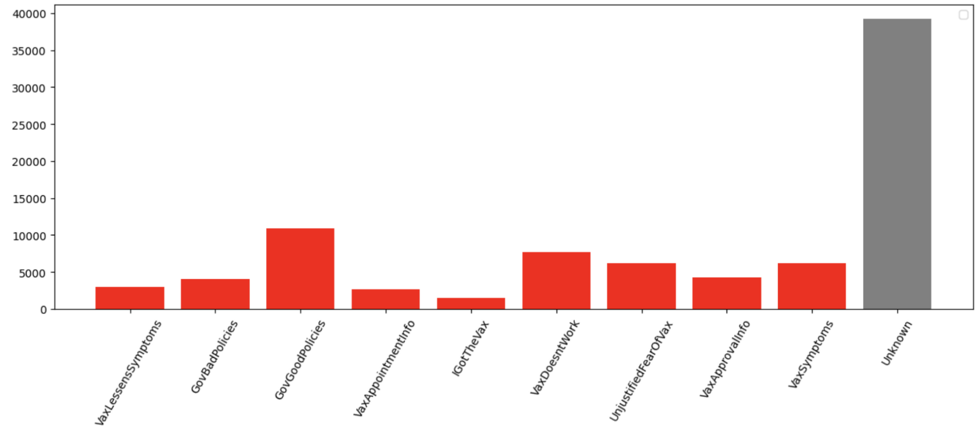


Figure 6.8. Example of the Theme Distribution

Coverage: We render a pie plot to better visualize the proportion of instances that have been assigned to a theme, in contrast to the proportion of instances that remain unassigned. An example of this visualization is presented in Figure 6.9

t-SNE Plot: We render a t-SNE plot to visualize all instance groundings in a 2-Dimensional plot [198]. To do this, t-SNE converts similarities between different high-dimensional data points to joint probabilities, and minimizes the KL divergence between the probabilities of a low-dimensional embedding and the high-dimensional data points. An example of this visualization is presented in Fig. 6.10.

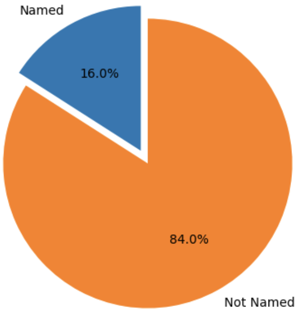


Figure 6.9. Example of Coverage Plot

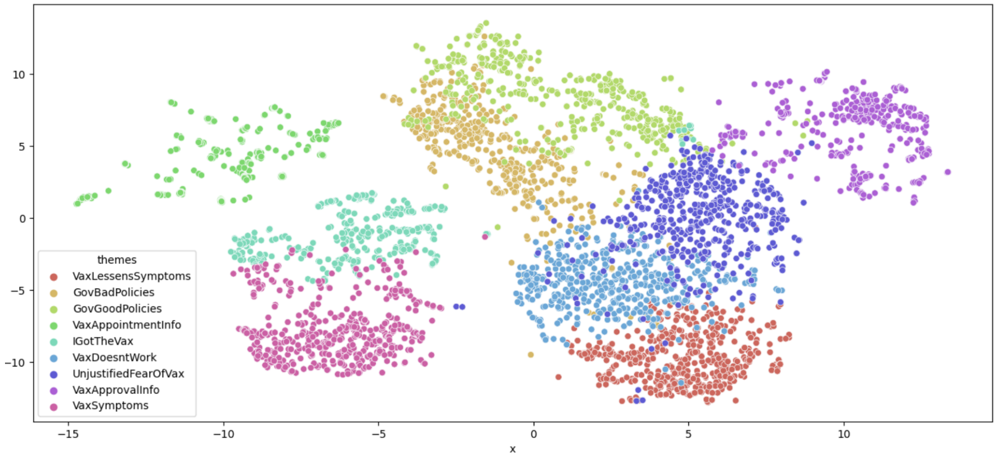


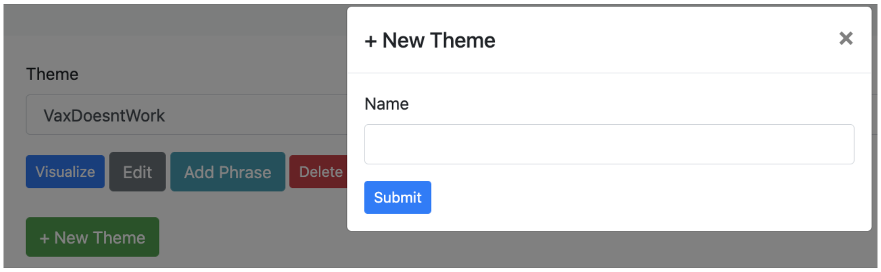
Figure 6.10. Example of 2D t-SNE Visualization

6.3.4 Intervention Operations

Intervention operations allow experts to introduce knowledge and judgements into the system to improve the discovery and grounding of emerging themes. We describe each of them below.

Adding and Removing Themes

We allow users to create and remove themes. The only requirement for creating a new theme is to give it a unique name. A screenshot of this functionality is provided in Figure 6.11. Similarly, themes can be removed at any point. If any instances are assigned to a theme being removed, they will be assigned to the *"Unknown"* theme.



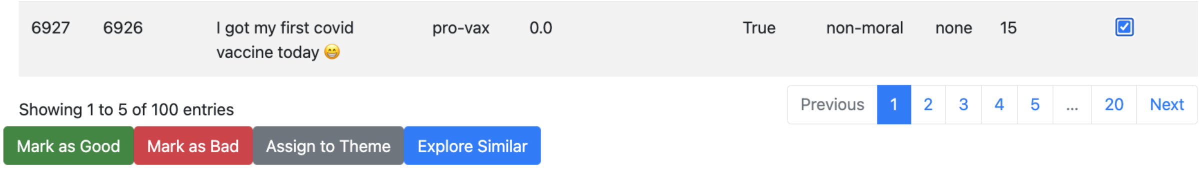
Theme: VaxDoesntWork

Figure 6.11. Adding New Themes

Adding and Removing Examples

We allow users to assign “good” and “bad” examples to existing themes. Good examples are instances that characterize the named theme. For example, for the theme *“The Vaccine Doesn’t Work”*, we could add an example stating something like *“People are dying every day with the vaccine, people are still getting COVID with the vaccine. Open your eyes!”*. In contrast, bad examples are instances that could have similar wording to a good example, but that have different meaning. For example, for the theme *“The Vaccine Doesn’t Work”*, we could add an example stating something like *“While it is true that you could still get COVID with the vaccine, it significantly reduces the chances of contracting it and dying from it.”* .

Users can add examples in two ways: they can mark grounded instances as “good” or “bad” (See Fig 6.12), or they can directly contribute example phrases (See Fig 6.13).

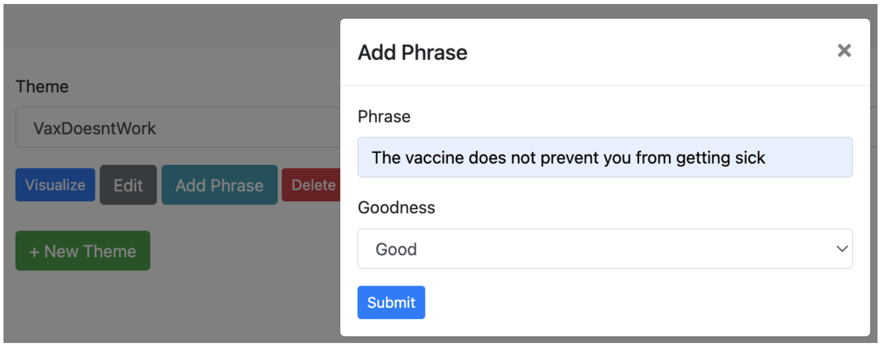


6927	6926	I got my first covid vaccine today 😁	pro-vax	0.0	True	non-moral	none	15	☑

Showing 1 to 5 of 100 entries

Figure 6.12. Marking Instances as *Good* or *Bad*

Similarly, “good” and “bad” phrases can be deleted. In the case of grounded instances, the “goodness” field is updated. In the case of contributed phrases, they are eliminated from the database.



Add Phrase

Theme

VaxDoesntWork

Phrase

The vaccine does not prevent you from getting sick

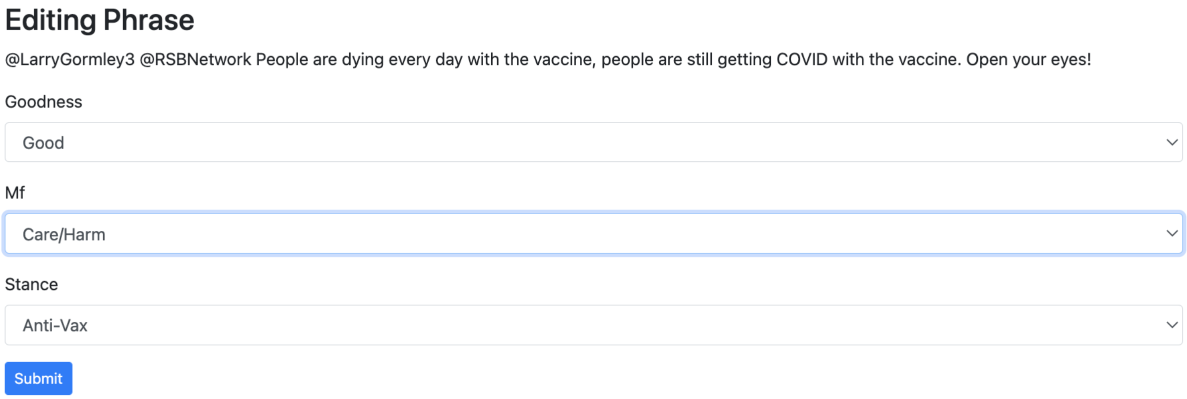
Goodness

Good

Figure 6.13. Adding *Good* or *Bad* Examples

Adding or Correcting Stances and Moral Foundations

As we saw in the exploratory operations, we allow users to upload additional observed or predicted attributes for each textual instance. For instances and phrases added as “good” and “bad” examples, we allow users to add or edit the values of this attributes. The intuition behind this operation is to collect holistic high-quality examples for learning to ground instances. A screenshot of this operation can be observed in Fig 6.14.



Editing Phrase

@LarryGormley3 @RSBNetwork People are dying every day with the vaccine, people are still getting COVID with the vaccine. Open your eyes!

Goodness

Good

Mf

Care/Harm

Stance

Anti-Vax

Figure 6.14. Correcting Stances and Moral Foundations

6.4 Interaction Stages

We recruit three human experts in Natural Language Processing and Computational Social Science and let them interact with our tool to find and ground themes emerging from the COVID-19 debate in the United States. To evaluate the usefulness of our tool, we perform a two-stage analysis. In the first stage, experts focus on the challenge of interactively grounding the latent themes. In the second stage, the experts focus on discovering the space of relevant themes from scratch. Below, we present each of these scenarios in detail and perform both qualitative and quantitative evaluations to assess the outcome of the interaction.

6.4.1 Stage 1: Interactively Assigning Statements to Themes

We first address the challenge of grounding known themes in large scale language resources. In this setting, we assume that we know what is the set of relevant themes. Our main goal is to improve the matching of instances (in this case, tweets) to the set of relevant themes. Essentially, we want to improve the assignments from tweets to themes. Given that we characterize themes as reasons used by people to accept or refuse the vaccine, we consider assignments to be better if they are more cohesive (e.g. if they are more strongly correlated with stance, moral foundations and entity roles). Below, we explain the interactive process in detail and present an evaluation of the results obtained.

Interactive Sessions

To evaluate this stage, we follow a simple protocol where three human coders use the operations above to explore an initial seed set of themes. To initialize the system, we use the 12 reasons suggested by Wawrzuta, Jaworski, Gotlib, *et al.*, 2021 [183], and represent them using the one-sentence explanation provided. Our main goal in this stage is to ground these reasons in a set of approximately 85,000 unlabeled tweets about the COVID-19 vaccine. To map tweets to reasons, we use the similarity between their SBERT embeddings [196]. Intuitively, the exploratory operations allow humans to diagnose how themes map to text, and the intervention operations allow them to act on the result of this diagnosis, by adding and removing reasons, and modifying the phrases characterizing each reason.

During the session, the three coders start by looking at the global picture: the themes distribution, the 2D visualizations [198] and the silhouette score [199]. Then, they query the

Table 6.1. Resulting Themes

Pro Vax	government distrust, vaccine dangerous, covid fake, vaccine oppression, pharma bad, natural immunity effective, vaccine against religion, vaccine does not work, vaccine not tested, bill gates' micro chip, vaccine tested on dogs, vaccine has fetal tissue, vaccine makes you sterile
Anti Vax	government trust, vaccine safe, covid real, vaccine not oppression, pharma good, natural immunity ineffective, vaccine not against religion, vaccine works, vaccine tested

themes one by one, looking at the word cloud (characterizing the distribution of short phrases over all texts assigned to the reason) and the 10 closest tweets to each reason. Following these observations, there is a discussion phase in which the coders follow a thematic analysis approach [200] to uncover the overarching themes that are not covered by the current set of themes, as well as the argumentation patterns that the method fails to identify. Then, they are allowed to add and remove themes, as well as explanatory phrases for them in natural language. Every time a reason or phrase is added or removed, all tweets are reassigned to their closest themes. This process was done over two one-hour sessions. The coders were NLP and Computational Social Science researchers, two female and one male, between the ages of 25 and 40.

In the first session, the coders focused on adding new themes and removing themes that were not prevalent in the data. For example, they noticed that the initial set of themes contained mostly anti-vaccine arguments, and added a positive theme for each negative theme (e.g. *government distrust* ⇒ *government trust*). In addition to this, they broke down the theme ”Conspiracy Theory” into specific conspiracy theories, such as *Bill Gates' micro chip*, *the vaccine contains fetal tissue*, and *the vaccine makes you sterile*. They also removed infrequent themes, such as *the swine flu vaccine*. The final set of themes can be observed in Tab. 6.1

In the second session, the coders focused on identifying the argumentative patterns that were not being captured by the original reason explanations, and came up with overarching patterns to create new examples to improve the representation of the themes. For example, in the case of the *government distrust* reason, the coders found that phrases with strong words were needed (e.g. *F the government*), as well as examples that suggested that the government was ”good at being bad” (e.g. *the government strong record of screwing things up*), and examples with explicit negations (e.g. *the government does not work logically*). The full list of uncovered patterns is presented in Tab. 6.2. Once patterns were identified, each coder contributed a set of 2-5 examples, which were added to the reason representation. Tables 6.3 and 6.4 enumerate the full list derived phrases, and specifically highlight which themes were added and removed during interaction.

Grounding Approach

In this scenario, we performed a fairly straightforward Nearest Neighbors grounding approach. Each tweet was assigned to its closest theme based on the embedding similarity between the tweet and theme explanatory phrases. Given that each theme may contain more than one phrase, the maximum similarity is considered. During the interaction, whenever an intervention operation was triggered, the examples were re-grounded.

To visualize the impact of interaction, we show the overall distribution of themes before and after interaction in Fig. 6.15. We can observe that interactively refining the theme representations by incorporating new phrases results in a better distributed assignment of instances to themes.

Evaluation

To evaluate the relevance of our refined themes, we perform a correlation test between the different dimensions of analysis in our small subset of annotated data. We calculate the Pearson correlation matrices and present them in Fig. 6.16. We compare the themes obtained interactively with the seed set of manual themes, and with a set of topics extracted using Latent Dirichlet Allocation (LDA) [201], a generative, unsupervised topic modeling technique that allows a set of textual instances (or documents) to be explained by unobserved groups of words that explain their similarity. In Figs 6.16 and 6.17, we can observe that our refined themes have higher correlations with both vaccination stance and moral foundations than both the original set of themes and the derived LDA topics.

Table 6.2. Overarching argumentation patterns uncovered by coders during interaction

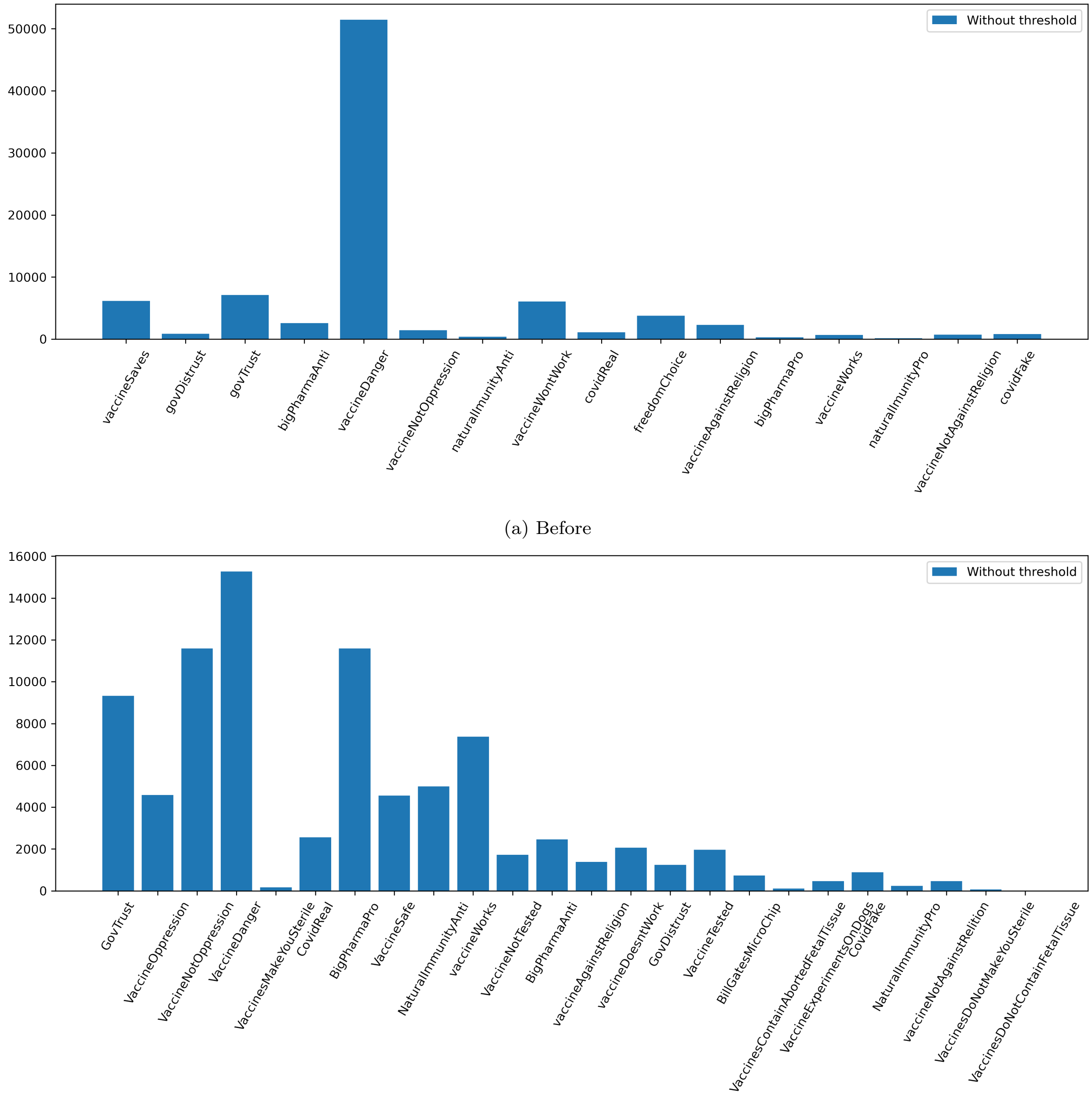
Themes	Overarching Patterns
GovDistrust	Add phrases with strong word for distrust "Good at being bad" Explicit negations
GovTrust	Hedging phrases (sort-of trust)
VaxDanger	Closer connection between vaccine words and danger words (related to sickness, bad effects) Explit negations Rhetorical questions Refusing the vaccine for medical reasons
VaxSafe	Explicit mentions of safety Explicit negations
CovidFake	Stronger relevant negative words (fake, scam, hoax) Explicit negations
CovidReal	Trust the science References to Covid hospitalization on the rise, explicit mentions of hospitals Explicit negations
VaxOppression	Legal language Explicit mentions of discrimination and oppression Sarcasm
VaxNotOppression	Justifying mandates Freedom to be protected Criticizing others using "you/people" language, focus freedom on me/my/I
BigPharmaAnti	Stronger words against pharmaceutical companies (corrupt, evil) Not accountable / irresponsible past behavior Mentions of negative side-effect of other products (cancer)
BigPharmaPro	Trust science/research and vaccine development process Language about intent, the vaccine was created to do something good, explicit names of companies
NaturalImmunityPro	The vaccine is not enough Explicit mentions to population immunity, herd immunity and antibodies
NaturalImmunityAnti	Emphasis on global look, collective entities, society Natural immunity characterized as dangerous or not effective Mentions of experts and trusting science
VaxAgainstReligion	I put it in god hands (god is deciding) Treating pro-vax as another religion
VaxNotAgainstReligion	"Religious" in quotes Bugus exemptions "Where is your faith" Call to action: get tested/get vaccinated/put a mask on (mentions of compassion) No religion ask members to refuse vaccine
VaxDoesntWork	Reference to "magic vaccine" "Never developed", "doesn't work" Questions: why are deaths high? Why is corona not going away? Why are vaccinated people dying?
VaxWorks	"ask a doctor", consult with an expert Research on the vaccine is good/has been going on for a long time Capture differences, e.g. "good trials" vs. rushed ones.
VaxNotTested	Language suggesting "rushed through trials" and "experimental vaccine"
VaxTested	trust the research and development process Testing can be confused with covid-test, use other language.

Table 6.3. AntiVax themes and phrases. Themes that were added during interaction are shown in blue. Themes that were removed are shown in red. Original explanatory phrases and examples are presented in bold.

Themes	Phrases
GovDistrust	**”lack of trust in the government”**, ”Fuck the government”, ”The government is a total failure”, ”Never trust the government”, ”Biden is a failure”, ”Biden lied people die”, ”The government and Fauci have been dishonest”, ”The government always lies”, ”The government has a strong record of screwing things up”, ”The government is good at screwing things up”, ”The government is screwing things up”, ”The government is lying”, ”The government only cares about money”, ”The government doesn’t work logically”, ”Do not trust the government”, ”The government doesn’t care about people’s health”, ”The government won’t tell you the truth about the vaccine”
VaxDanger	**”the vaccine will be dangerous to health”**, ”Covid vaccines can cause blood clots”, ”The vaccine is a greater danger to our children’s health than COVID itself”, ”The vaccine will kill you”, ”The experimental covid vaccine is a death jab”, ”The covid vaccine causes cancer”, ”The covid vaccine is harmful for pregnant women and kids”, ”The vaccine increases health risk”, ”The vaccine isn’t safe”, ”What are vaccines good for? Nothing, rather it increases risk”, ”I and many others have medical exemptions”, ”The vaccine is dangerous for people with medical conditions”, ”I won’t take the vaccine due to medical reasons”, ”The vaccine has dangerous side effects”
CovidFake	**”COVID-19 disease does not exist”**, ”Covid is fake”, ”covid is a hoax”, ”covid is a scam”, ”covid is propaganda”, ”the pandemic is a lie”, ”covid isn’t real”, ”I don’t think that covid is real”, ”I don’t buy that covid is real”, ”I don’t think there is a pandemic”, ”I don’t think the pandemic is real”, ”I don’t buy that there is a pandemic”
VaxOppression	**”I do not want to be vaccinated because I have freedom of choice”** ”Forcing people to take experimental vaccines is oppression”, ”The vaccine has nothing to do with Covid-19, it’s about the vaccine passport and tyranny”, ”The vaccine mandate is unconstitutional”, ”I choose not to take the vaccine”, ”My body my choice”, ”I’m not against the vaccine but I am against the mandate”, ”I have freedom to choose not to take the vaccine”, ”I am free to refuse the vaccine”, ”It is not about covid, it is about control”, ”Medical segregation based on vaccine mandates is discrimination”, ”The vaccine mandate violates my rights”, ”Falsely labeling the injection as a vaccine is illegal”, ”Firing over vaccine mandates is oppression”, ”Vaccine passports are medical tyranny”, ”I won’t let the government tell me what I should do with my body”, ”I won’t have the government tell me what to do”
BigPharmaAnti	**”the vaccine was created only for the profit of pharmaceutical companies”**, ”We are the subjects of massive experiments for the Moderna and Pfizer vaccines”, ”Pharmaceutical companies are corrupt”, ”The pharmaceutical industry is rotten”, ”Big Pharma is evil”, ”How would you trust big pharma with the COVID vaccine? They haven’t been liable for vaccine harm in the past”, ”Covid vaccines are not doing what the pharmaceutical companies promised”, ”Pharmaceutical companies have a history of irresponsible behavior”, ”I don’t trust Johnson & Johnson after knowing their baby powder caused cancer for decades”
NatImmunityPro	**”natural methods of protection against the disease are better than vaccines”**, ”Herd immunity is broad, protective, and durable”, ”Natural immunity has higher level of protection than the vaccine”, ”Embrace population immunity”, ”I trust my immune system”, ”I have antibodies I do not need the vaccine”, ”Natural immunity is effective”
VaxAgainstReligion	”The vaccine is against my religion”, ”The vaccines are the mark of the beast”, ”The vaccine is a tool of Satan”, ”The vaccine is haram”, ”The vaccine is not halal”, ”I will protect my body from a man made vaccine”, ”I put it all in God’s hands”, ”God will decide our fate”, ”The vaccine contains bovine, which conflicts with my religion”, ”The vaccine contains aborted fetal tissue which is against my religion”, ”The vaccine contains pork, muslims can’t take the vaccine”, ”Jesus will protect me”, ”The vaccine doesn’t protect you from getting or spreading Covid, God does”, ”The covid vaccine is another religion”
VaxDoesntWork	**”the vaccine does not work”**, ”covid vaccines do not stop the spread”, ”If the vaccine works, why are deaths so high?”, ”Why are vaccinated people dying?”, ”If the vaccine works, why is covid not going away?”
VaxNotTested	**”the vaccine is not properly tested, it has been developed too quickly”**, ”Covid-19 vaccines have not been through the same rigorous testing as other vaccines”, ”The Covid vaccine is experimental”, ”The covid vaccine was rushed through trials”, ”The approval of the experimental vaccine was rushed”, ”How was the vaccine developed so quickly?”
VaxExperimentDogs	”Fauci tortures dogs and puppies”, ”Animal shelters are empty because Dr Fauci allowed experimenting of various Covid vaccines/drugs on dogs and other domestic pets”
BillGatesMicroChip	”The covid vaccine is a ploy to microchip people”, ”Bill Gates wants to use vaccines to implant microchips in people”, ”Globalists support a covert mass chip implantation through the covid vaccine”
VaxFetalTissue	”There is aborted fetal tissue in the Covid Vaccines”, ”the Covid vaccines contain aborted fetal cells”
VaxMakeYouSterile	”The covid vaccine will make you sterile”, ”Covid vaccine will affect your fertility”
NoResponsibility	**no one is responsible for the potential side effects of the vaccine**
SwineFluVax	**mentioning the past development of the swine flu vaccine**
VaxResistance	**the vaccine has existed before the COVID-19 epidemic, now there is too much resistance**
ConspiracyTheories	**conspiracy theories, hidden vaccine effects (e.g., chips)**

Table 6.4. ProVax themes and phrases. Themes that were added during interaction are shown in blue. Themes that were removed are shown in red. Original explanatory phrases and examples are presented in bold.

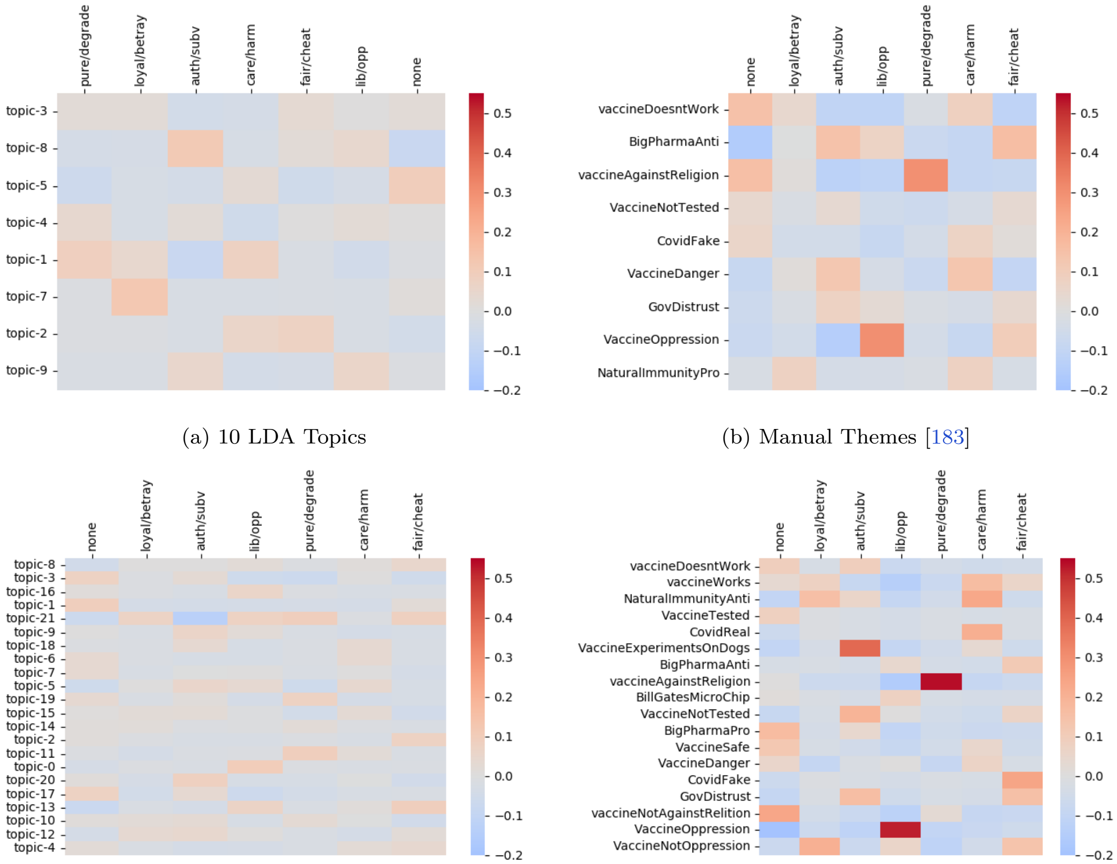
Themes	**Phrases**
GovTrust	”We trust the government”, ”The government cares for people”, ”We are thankful to the government for the vaccine availability”, ”Hats off to the government for tackling the pandemic”, ”It is a good thing to be skeptical of the government, but they are right about the covid vaccine”, ”It is a good thing to be skeptical of the government, but they haven’t lied about the covid vaccine”, ”The government can be corrupt, but they are telling the truth about the covid vaccine”, ”The government can be corrupt, but they are not lying about the covid vaccine”
VaxSafe	”The vaccine is safe”, ”Millions have been vaccinated with only mild side effects”, ”Millions have been safely vaccinated against covid”, ”The benefits of the vaccine outweigh its risks”, ”The vaccine has benefits”, ”The vaccine is safe for women and kids”, ”The vaccine won’t make you sick”, ”The vaccine isn’t dangerous”, ”The vaccine won’t kill you”, ”The covid vaccine isn’t a death jab”, ”The covid vaccine doesn’t harm women and kids”
CovidReal	”Covid is real”, ”I trust science”, ”Covid death is real”, ”The science doesn’t lie about covid”, ”Scientist know what they are doing”, ”Scientist know what they are saying”, ”Covid hospitalizations are on the rise”, ”Covid hospitalizations are climbing as fourth stage surge continues”, ”Covid’s death toll has grown faster”, ”Covid is not a hoax”, ”The pandemic is not a lie”, ”The pandemic is not a lie, hospitalizations are on the rise”
VaxNotOppression	”The vaccine mandate is not oppression because vaccines lower hospitalizations and death rates”, ”The vaccine mandate is not oppression because it will help to end this pandemic”, ”The vaccine mandate will help us end the pandemic”, ”We need a vaccine mandate to end this pandemic”, ”I support vaccine mandates”, ”If you don’t get the vaccine based on your freedom of choice, don’t come crawling to the emergency room when you get COVID”, ”If you refuse a free FDA-approved vaccine for non-medical reasons, then the government shouldn’t continue to give you free COVID tests”, ”You are free not to take the vaccine, businesses are also free to deny you entry”, ”You are free not to take the vaccine, businesses are free to protect their customers and employees”, ”If you choose not to take the vaccine, you have to deal with the consequences”, ”If it is your body your choice, then insurance companies should stop paying for your hospitalization costs for COVID”
BigPharmaPro	”I trust the science and pharmaceutical research”, ”Pharmaceutical companies are not hiding anything”, ”The research behind covid vaccines is public”, ”The Pfizer vaccine is saving lives”, ”The Moderna vaccines are helping stop the spread of covid”, ”The Johnson and Johnson vaccine was created to stop covid”, ”Pharmaceutical companies are seeking FDA approval”, ”Pharmaceutical companies are following standard protocols”
NatImmunityAnti	”Only the vaccine will end the pandemic”, ”Vaccines will allow us to defeat covid without death and sickness”, ”The vaccine has better long term protection than to natural immunity”, ”Natural immunity is not effective”, ”Natural immunity would require a lot of people getting sick”, ”Experts recommend the vaccine over natural immunity”
VaxReligionOk	”The vaccine is not against religion, get the vaccine”, ”No religion ask members to refuse the vaccine”, ”Religious exemptions are bogus”, ”When turning in your religious exemption forms for the vaccine, remember ignorance is not a religion”, ”Disregard for others’ lives isn’t part of your religion”, ”Jesus is trying to protect us from covid by divinely inspiring scientists to create vaccines”
VaxWorks	”The vaccine works”, ”Vaccines do work, ask a doctor or consult with an expert”, ”The covid vaccine helps to stop the spread”, ”Unvaccinated people are dying at a rapid rate from COVID-19”, ”There is a lot of research supporting that vaccines work”, ”The research on the covid vaccine has been going on for a long time”
VaxTested	”Covid vaccine research has been going on for a while”, ”Plenty of research has been done on the covid vaccine”, ”The technologies used to develop the COVID-19 vaccines have been in development for years to prepare for outbreaks of infectious viruses”, ”The testing processes for the vaccines were thorough didn’t skip any steps”, ”The vaccine received FDA approval”
ProVax	**positive attitude**



(a) Before

(b) After

Figure 6.15. Theme grounding before and after interaction.



(a) 10 LDA Topics

(b) Manual Themes [183]

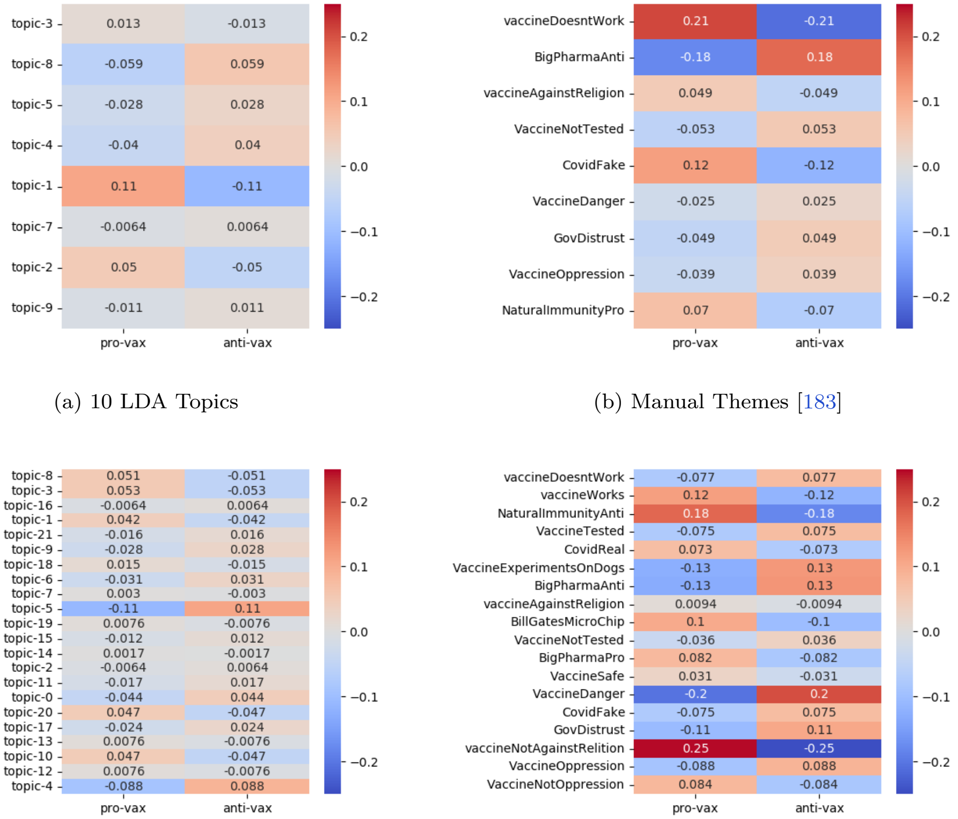
(c) 20 LDA Topics

(d) After Interaction

Figure 6.16. Correlations between themes and moral foundations

Given that our themes are named, they are easier to interpret than LDA topics. We can interpret themes as distributions over moral foundations and stances (and vice-versa). This analysis provides a useful way to explain each of these dimensions. For example, we can see that *care/harm* is strongly correlated with themes such as *covid is real*, *the vaccine works*, and *natural immunity is ineffective*. Other expected trends emerge, such as *purity/degradation* being highly correlated with *vaccine against religion*.

In Tab. 6.5 we show the top four themes for *fairness/cheating*. We choose this moral foundation given that is evenly split among stances and is active for different themes. We show the top two (E,R,P) tuples for each theme. We can appreciate that while this moral foundation is used by people on both sides, the reasons offered and entities used vary. On the anti-vax side, authority figures and vaccine trials are portrayed as negative actors, while



(a) 10 LDA Topics

(b) Manual Themes [183]

(c) 20 LDA Topics

(d) After Interaction

Figure 6.17. Correlations between themes and stance

women and children are portrayed as targets. On the pro-vax side, COVID and unvaccinated people are portrayed as negative actors, and the general public is portrayed as a target.

Lastly, we perform a quantitative analysis of the impact of the interactive themes in the modeling framework for COVID-19 introduced in Chapter 5, Section 5.2.3. As a reminder, we used DRaiL probabilistic rules to capture the dependencies between the repeating themes used as reasons to accept or refuse the vaccine, and stance and moral foundation prediction. Table 6.6 shows the impact of explicitly modeling themes as reasons (Eq. 6.4). We show the performance for the initial themes proposed by Wawrzuta, Jaworski, Gotlib, *et al.*, 2021 [183], which are all from the anti-vaccine perspective, and the impact of our two rounds of interaction, expanding and refining themes (round 1) and augmenting argumentative themes (round 2).

Table 6.5. Top 4 reasons for **Fairness/Cheating**, and their most frequent opinions and entity roles

VaxNotOppression	**VaxDanger**
70% Pro-Vax (responsible people, target, neg) (un-vax people, actor, neg)	**60% Anti-Vax** (pregnant women, target, neg) (trial vax, actor, neg)
GovDistrust	**VaxWorks**
75% Anti-Vax (children, target, neg) (Fauci, actor, neg)	**75% Pro-Vax** (people, target, neg) (COVID, actor, neg)

Table 6.6. Contribution of themes at different interaction rounds (Weighted F1)

Model	**MF**	**Vax. Stance**
ALL (-Themes)	60.07	67.72
+ Themes-Original	61.51	72.62
+ Themes-Interaction-1	61.21	**73.83**
+ Themes-Interaction-2	**62.27**	72.53

6.4.2 Stage 2: Interactively Discovering Latent Explanations

In this Section, we address the challenge of discovering the space of themes that emerge in a large language resource. In this setting, we make no assumptions about the number of relevant themes or what they ought to be. Our main goal is to allow human experts to leverage both computational and qualitative techniques to explore the data and identify relevant patterns.

The main challenge in this stage is to obtain a set of themes that accounts for as many tweets as possible, while maintaining the cohesiveness of the partitions and the tweet to theme assignments. Below, we explain the interactive process in detail and present an evaluation of the results obtained.

Interactive Sessions

In this stage, we follow a simple protocol where three human coders use the operations above to discover themes from scratch. Unlike the scenario presented before, coders do not have access to an initial set of themes, and they must leverage the operations provided to find them in the data. To initialize the system, the coders start by using the clustering operation to find 10 initial clusters of roughly the same size. Our goal in this stage is to find an initial broad set of themes that accounts for a large portion of our data. The process was done over two one-hour sessions. The coders were NLP and Computational Social Science researchers, two female and one male, between the ages of 25 and 40.

During the first session, the coders inspected the clusters one by one by looking at the examples closest to the centroid. This followed a discussion phase, in which the coders follow a thematic analysis approach [200]. When a pattern was identified, the coders created a new theme, named it, and marked a set of good example instances that helped in characterizing the named theme. Table 6.7 shows each initial cluster, the patterns identified, and the named chosen by the coders during the discussion. When a pattern was not obvious, coders explored similar instances to the different statements found. Whenever the similarity search resulted in a new pattern, the coders created a new theme, named it, and marked a set of good example instances that helped in characterizing the named theme.

During the second session, the coders looked at the local theme explanations and repeated a process similar to the first scenario, enhancing each theme with additional phrases. Note that each theme already contained a small set of representative tweets, which were marked as "good" in the previous session. In addition to contributing "good" example phrases, coders also contributed some "bad" example phrases to push the representation of the theme away from statements that have high lexical overlap with the good examples, but different meaning (e.g. the vaccine works *vs.* the vaccine does not work). Finally, coders examined each exemplary tweet and phrase for the additional analysis dimensions (e.g. stance, morality, moral foundation). In cases where the initial prediction was perceived as wrong, the coders corrected it. Table 6.8 contains a subset of the exemplary tweets, with their corrected judgements. Table 6.9 contains the additional contributed phrases and their judgements.

Table 6.7. Patterns Identified in Initial Clusters and Resulting Themes

Cluster	Pattern	Named Theme
K-Means 0	Discusses what the vaccine can and cannot do. Emphasis in reducing COVID-19 symptoms in case of infection ("like a bad cold"). Contains tweets with both stances.	VaxLessensSymptoms
K-Means 1	A lot of mentions to political entities. Politicians get in the way of public safety	GovBadPolicies
K-Means 2	A lot of tweets with mentions and links. Not a lot of textual context. Some examples thanking and praising governmental policies. **Theme added upon inspecting similar tweets**	GovGoodPolicies
K-Means 3	Overarching theme related to vaccine rollout. Mentions to pharmacies that can distribute, distribution in certain states, places with unfulfilled vax appointments. **Too broad to create a theme**	-
K-Means 4	Broadcast of vaccine appointments. Which places you can get vaccine appointments at.	VaxAppointments
K-Means 5	"I got my vaccine" type tweets	GotTheVax
K-Means 6	Mixed cluster, not a clear theme in centroid. Two prominent flavors: the vaccine not working and people complaining about those who are scared of vaccine.	VaxDoesntWork UnjustifiedFearOfVax
K-Means 7	Tweets look the same as K-Means 5	-
K-Means 8	Tweets about development and approval of vaccines	VaxApproval
K-Means 9	Tweets related to common vaccine side-effects	VaxSideEffects

Table 6.8. Examples of “Good” Tweets for each Theme.

Theme	Good Tweets	Stance	Morality	MF
VaxLessens Symptoms	The vaccine doesn’t prevent you from getting Covid people!! All it does is lessen the symptoms. Plenty of vaccinated people are still getting Covid! Who needs it with a 99% survival rate anyhow?	Anti-Vax	Moral	Care/Harm
	Just a heads up. The vaccine doesn’t prevent you from catching COVID. It prevents you from dying of COVID. Breakthrough cases are rare but do happen and are usually mild.	Pro-Vax	Non-moral	None
	We SHOULD worry about others The vaccine doesn’t guarantee u won’t get COVID but does vastly improve odds u won’t get it. Also vastly improves odds symptoms are minor, so improves odds u survive and limits likelihood you’ll spread virus while helping prevent future mutations.	Pro-Vax	Moral	Care/Harm
GovBadPolicies	Your Demagogue FASCIST downplayed the deployment of the COVID vaccine. He never undertook any substantial effort to promote vaccination. No events focused on it. NO mention of the vaccine publicly, he was so focused on injecting his election-fraud nonsense into his CULT.	Pro-Vax	Moral	Auth/Subv
	The Biden administration, is the misguided, medical disinformation, the political motivated danger to its own citizens health. People are dying at a rate of 30+ a day from covid vaccine. The risk from the vaccine, are worse than contracting covid for those younger than 50yrs	Anti-Vax	Moral	Auth/Subv
	Worst Covid s of the year and @GovDunleavy issues a presser that doesn’t even contain the word “vaccine” Meanwhile he caters to conspiracy theorists by giving interviews to a Florida-based blogger who spreads Covid misinformation	Pro-Vax	Moral	Auth/Subv
GovGoodPolicies	Thank you for your leadership on this critical issue, @GovSisolak.	Pro-Vax	Moral	Auth/Subv
	Thank You @POTUS! So productive having REAL leadership from the @WhiteHouse!!!	Pro-Vax	Moral	Auth/Subv
	Thank You @GovernorTomWolf for helping to making the protection of the Lives of Pennsylvanians a priority, unlike your peers to the South.	Pro-Vax	Moral	Care/Harm
VaxAppointments	COVID19 vaccine appointments are also available for eligible individuals at local pharmacies.	Neutral	Non-moral	None
	A new Covid vaccine site is open in NationalCity You have to make an appointment before going. You can call 211 or make an appointment online. For more information go to ...	Neutral	Non-moral	None
	New COVID vaccine appointments available for 04020. Go to Walgreen Drug Store, 151 MAPLE STREET, CORNISH, ME	Neutral	Non-moral	None
GotTheVax	I got my first covid vaccine today	Pro-Vax	Non-moral	None
	Today I got my vaccine — take that #covid_19! #moderna #firstround #immunize #flattenthecurve #grateful #pandemiclife	Pro-Vax	Non-moral	None
	Getting my vaccine @CityofDetroit #COVID19 #CovidVaccine #COVID	Pro-Vax	Non-moral	None
VaxDoesntWork	People are dying every day with the vaccine, people are still getting COVID with the vaccine. Open your eyes!	Anti-Vax	Moral	Care/Harm
	The vaccine is not keeping people out of the hospital. People are still getting Covid after vaccination and dying.	Anti-Vax	Moral	Care/Harm
	People are still dying from the vaccine and, the covid even if they are vaccinated	Anti-Vax	Moral	Care/Harm
UnjustifiedFear OfVax	To all the people afraid of the vaccine. I say be afraid of #COVID more and choose to face covid on your own terms.	Pro-Vax	Moral	Care/Harm
	There are a lot of people out there passing on their opportunity at receiving the COVID-19 #vaccine. I want to try to eliminate fear surrounding it. #Skepticism about the COVID vaccine is troubling with a quarter of people indicating they will never take the vaccine. #vaccinehesitancy definitely threatens herd immunity which is said to be up to 80% #RAC #healthequity	Pro-Vax	Moral	Fair/Cheat
VaxApproval	There it is! A whole vaccine approved by the FDA to combat Covid! #Pfizer	Pro-Vax	Non-moral	None
	Johnson Johnson Covid Vaccine Shots Future Depends On Cautious Vaccine Experts	Pro-Vax	Non-moral	None
	How about getting that shot now my Twitter scientists? CNBC: FDA grants full approval to Pfizer-BioNTech’s Covid shot, clearing path to more vaccine mandates.	Pro-Vax	Non-moral	None
VaxSideEffects	This COVID vaccine is f*ing me up more than when I actually had COVID back in January	Neutral	Non-moral	None
	Well these covid vaccine symptoms ain’t no joke, I had a high temp two days in a row, chills, I brake one temp after another, hopefully I’m finally finished with the symptoms	Neutral	Non-moral	None
	Covid Vaccine absolutely crushing me, feel worse now than I did with COVID	Neutral	Non-moral	None

Table 6.9. Good and Bad Contributed Phrases

Theme	Type Phrases	Phrases	Stance	Morality	MF
GovBadPolicies	Good	Politicians don't know what they are doing	Neutral	Moral	Auth/Subv
	Bad	Politicians have a handle on COVID	Neutral	Moral	Auth/Subv
GovGoodPolicies	Good	Politicians have a handle on COVID	Neutral	Moral	Auth/Subv
	Bad	Politicians don't know what they are doing	Neutral	Moral	Auth/Subv
VaxDoesntWork	Good	You can still die from COVID with the vaccine	Anti-Vax	Moral	Care/Harm
	Bad	The vaccine keeps you safe	Pro-Vax	Moral	Care/Harm
UnjustifiedFearOfVax	Good	Don't be scared of the vaccine	Pro-Vax	Moral	Care/Harm
	Bad	So many people are hurt from the vaccine, I am afraid to take it	Anti-Vax	Moral	Care/Harm

Grounding Approaches

In this scenario, we do not assume that we know the full space of latent themes. For this reason, we do not try to assign a theme to each instance. We expect that the set of themes introduced by the human experts at each round of interaction will cover a subset of the total instances available. We evaluate two alternative methods to assign instances to themes:

Nearest Neighbors Approach

In this approach, each tweet was assigned to its closest theme *if and only if* the distance to the closest theme was less than or equal to the distance to its previous cluster *and* the distance to the closest theme was less than or equal to the distance to the theme's bad examples and phrases. The pseudo-code for this process is outlined in Algorithm 4.

Algorithm 4 *Nearest Neighbors Grounding Approach*

```
for instance i ∈Dataset do
    Let $\text{assignment}_\text{i}$ be instance i's previous cluster assignment
    for theme $t$ ∈Themes do
        if $\text{dist(instance, theme)} \leq \text{dist(instance, assignment}_\text{i})$
        $\wedge$ $\text{dist(instance, theme)} \leq \text{dist(instance, bad\_ex}_t)$ then
            $\text{assignment}_\text{i} \leftarrow \text{theme}$
        end if
    end for
end for=0
```

DRaiL Model

In this approach, we use the contributed good and bad examples to generate training data for each tweet and train a DRaiL model. Our model uses the following rules:

Analysis Dimensions: Following Chapter 5, we define a rule for each analysis dimension. Here, we score whether a tweet $\mathtt{t_i}$ is moral, has moral foundation $\mathtt{m}$ and has stance $\mathtt{s}$. Additionally, we score whether entity $\mathtt{e_i}$ has role $\mathtt{r}$ and polarity $\mathtt{p}$.

$$\begin{aligned}
&r_0 : \mathtt{Tweet(t_i)} \Rightarrow \mathtt{IsMoral(t_i)}\\
&r_1 : \mathtt{Tweet(t_i)} \Rightarrow \mathtt{HasMF(t_i, m)}\\
&r_2 : \mathtt{Tweet(t_i)} \Rightarrow \mathtt{VaxStance(t_i, s)}\\
&r_3 : \mathtt{Mentions(t_i, e_i)} \Rightarrow \mathtt{HasRole(e_i, r)}\\
&r_4 : \mathtt{Mentions(t_i, e_i)} \Rightarrow \mathtt{EntPolarity(e_i, p)}
\end{aligned} \tag{6.1}$$

Dependency between roles and moral foundations and stances: The way an entity is portrayed in a tweet can be highly indicative of its moral and stance. Following Chapter 5, we explicitly model the dependency between an entity, its moral role, and the MF and stance.

$$\begin{aligned}
&r_5 : \mathtt{Mentions(t_i, e_j)} \wedge \mathtt{HasRole(e_i, r)} \wedge \mathtt{EntPolarity(e_i, p)} \Rightarrow \mathtt{HasMf(t_i, m)}\\
&r_6 : \mathtt{Mentions(t_i, e_j)} \wedge \mathtt{HasRole(e_i, r)} \wedge \mathtt{EntPolarity(e_i, p)} \Rightarrow \mathtt{HasStance(t_i, s)}
\end{aligned} \tag{6.2}$$

Dependency between stances and moral foundations: There is a significant correlation between the stance of a tweet with respect to the vaccine debate, and its moral foundation. Following Chapter 5, we model the dependency between the stance of a tweet and its MF.

$$r_7 : \mathtt{VaxStance(t_i, s)} \Rightarrow \mathtt{HasMf(t_i, m)} \tag{6.3}$$

Dependency between themes and moral foundations/stances: Explicitly modeling the dependency between repeating themes and other decisions can help us add inductive bias into our model, potentially simplifying the task. Following Chapter 5, we add two rules to capture this dependency, one between themes and moral foundations, and one between themes and stances.

$$\begin{aligned} &r_8 : \mathtt{Mentions(t_i, r)} \Rightarrow \mathtt{HasMf(t_i, m)} \\ &r_9 : \mathtt{Mentions(t_i, r)} \Rightarrow \mathtt{VaxStance(t_i, s)} \end{aligned} \tag{6.4}$$

Theme Predictions: We define rules for each theme $\mathtt{r}$, to score whether a tweet $\mathtt{t_i}$ mentions a theme $\mathtt{r}$, given each of the analysis dimensions. These are the main rules that differ from the model introduced in Chapter 5, and they will allow us to change the theme assignments.

$$\begin{aligned} &r_10 : \mathtt{Tweet(t_i)} \wedge \mathtt{IsMoral(t_i)} \wedge \mathtt{HasMF(t_i, m)} \wedge \mathtt{VaxStance(t_i, s)} \wedge \mathtt{IsTheme(r)} \\ &\Rightarrow \mathtt{Mentions(t_i, r)} \\ &r_{11} : \mathtt{Mentions(t_i, e_i)} \wedge \mathtt{HasRole(e_i, r)} \wedge \mathtt{EntPolarity(e_i, p)} \wedge \mathtt{IsTheme(r)} \\ &\Rightarrow \mathtt{Mentions(t_i, r)} \end{aligned} \tag{6.5}$$

Hard constraints: Following Chapter 5, we enforce that, if a tweet is predicted to be moral, then it needs to also be associated to a specific moral foundation. Likewise, if a tweet is not moral, then no MF should be assigned to it. Lastly, we include a constraint that limits the number of assigned themes for a tweet to be at most one. Note that our model allows for tweets to not be assigned to any theme, which is crucial to our design.

$$\begin{aligned} &c_0 : \mathtt{IsMoral(t_i)} \Rightarrow \neg\mathtt{HasMf(t_i, none)} \\ &c_1 : \neg\mathtt{IsMoral(t_i)} \Rightarrow \mathtt{HasMf(t_i, none)} \\ &c_2 : \mathtt{IsTheme(r_1)} \wedge \mathtt{IsTheme(r_2)} \wedge \mathtt{Mentions(t_i, r_1)} \Rightarrow \neg\mathtt{Mentions(t_i, r_2)} \end{aligned} \tag{6.6}$$

This DRaiL model allows us to assign tweets to themes. To learn this model, we generate K positive and negative examples using the good and bad examples for each theme. Namely, we find the K closest tweet to *any* of the phrases. The generated examples will inherit the phrase's theme, as well as all of their other judgements (i.e. moral foundation, stance, etc). We generate additional negative examples by drawing tweets from other themes. The pseudo-code for this procedure is outlined in Algorithm 5. Once the model has been learned, we can run inference over the full dataset and obtain the final assignments. Note that this will also update the previous predictions for all the analysis dimensions (e.g. stance, moral foundation). To prevent over-fitting other analysis dimensions to the current subset of themes, we only update them for the tweets that were assigned to one of the themes, and leave the rest untouched.

Evaluation

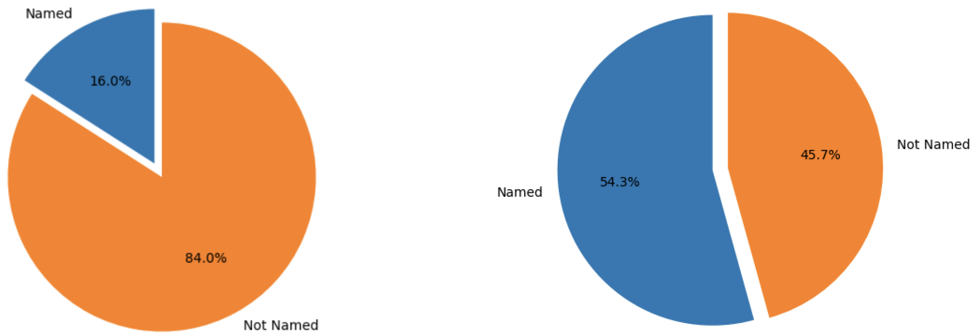
To evaluate the impact of the interactive protocol and grounding methodology, we look at the proportion of tweets that we are able to ground after our two rounds of interaction. Figure 6.18 shows that using the Nearest Neighbors approach, we are able to account for 16% of the space of tweets, while using our DRaiL model, we are able to account for 54% of the space of tweets. This results suggest that leveraging inference and the dependencies between the different analysis dimensions allows us to find more relevant tweets.

To measure whether the gain in coverage affects the grounding performance, we perform a human evaluation and present it in Table 6.10. From the set of assigned tweets, we sample a set of 100 tweets uniformly from the tweets that are closest to the centroid of the theme (i.e. those tweets that are closest to the good examples and phrases), to the tweets that are furthest from the centroid. We create 4 buckets, from closest to furthest, and sample accordingly. Then, for each (theme, tweet) pair, we generate a negative example by randomly sampling a tweet from a different theme. We annotate the pairs manually, using three annotators (K-alpha agreement can be seen in Tab. 6.10). We consider a pair as a positive pair if at least two annotators annotated it as such. Then, we measure the binary precision, recall and F1. While the numbers are not directly comparable across methods, it

Algorithm 5 *DRAIL Grounding Approach*

```
pos_train ← $\emptyset$
neg_train ← $\emptyset$
for theme $t$ ∈Themes do
    pos_train$_t$ ← $\emptyset$
    neg_train$_t$ ← $\emptyset$
    for phrase $p$ ∈good_ex$_t$ do
        pos_train$_t$ ← pos_train$_t$ $\bigcup$ closest(Dataset, phrase, K)
    end for
    for phrase $p$ ∈bad_ex$_t$ do
        neg_train$_t$ ← neg_trainv $\bigcup$ closest(Dataset, phrase, K)
    end for
    for theme $t'$ ∈Themes do
        if $t' \neq t$ then
            for phrase $p'$ ∈good_ex$_t'$ do
                neg_train$_t$ ← neg_train$_t$ $\bigcup$ closest(Dataset, phrase, K)
            end for
        end if
    end for
    pos_train ← pos_train$_t$
    neg_train ← neg_train$_t$
end for
rule_groundings ← get_rule_groundings(pos_train, neg_train)
model ← train_drail(rule_groundings)
for instance i ∈Dataset do
    assignment$_{\mathrm{i}}$ ← model(instance)
end for
=0
```

can give us an idea of the quality of our theme assignments. We confirm that our DRaiL model increases coverage without compromising on the quality of the assignments.



(a) Nearest Neighbors

(b) DRaiL

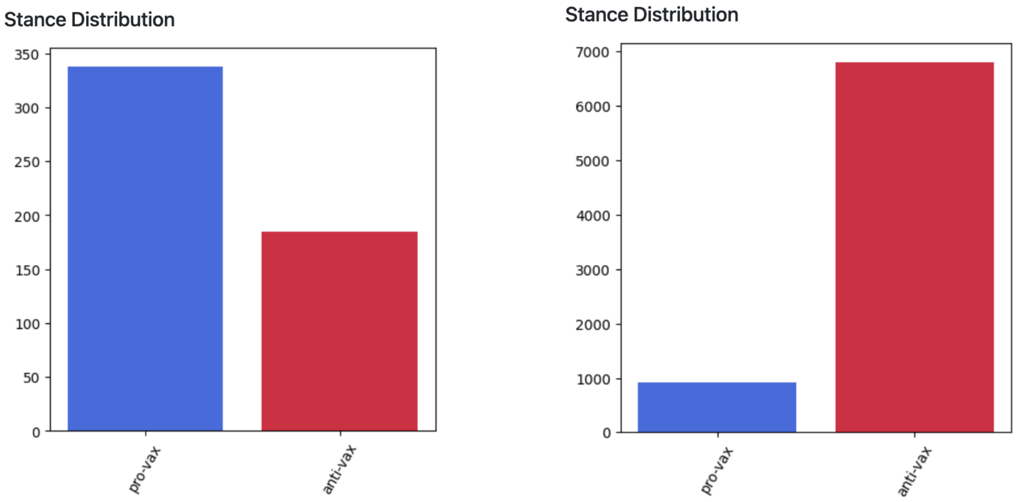
Figure 6.18. Theme coverage after two rounds of interaction

Table 6.10. Human Evaluation of Grounding Method

Grounding Method	K-Alpha	Precision	Recall	F1
Nearest Neighbors	0.7520	0.8617	**0.9878**	**0.9205**
DRaiL	0.7809	**0.8630**	0.9844	0.9197

Lastly, we inspect the theme visualizations to contrast the distribution of the supporting analysis dimensions. While we do not perform an exhaustive analysis, we can observe that the overall distribution of stances and moral foundations is less noisy when predictions are updated using the DRaiL model. For example, Fig 6.20 shows the distribution of stances for the theme *"The Vaccine Doesn't Work"*. We find that after updating predictions using DRaiL, the distribution goes from a slight majority of *pro-vaccine* to a solid majority of *anti-vaccine*, which more closely resembles the expected outcome.

We see a similar trend for the distribution of moral foundations for the theme *"Vaccine Appointments"*. This theme was introduced to represent tweets that advertised available appointments, which are generally informative tweets that do not state any opinions or judgements. We find that after updating predictions using DRaiL, the distribution goes from a majority of morally-charged tweets, to a majority of non-moral tweets, which more closely resembles the expected outcome.



(a) Nearest Neighbors

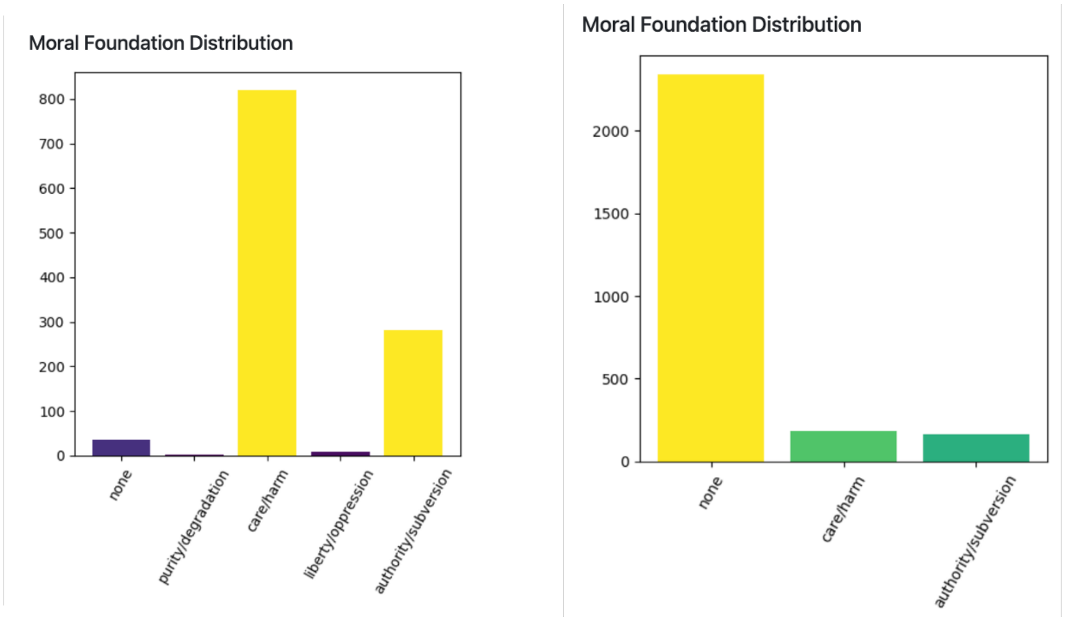
(b) DRAIL

Figure 6.19. Distribution of stances for the theme *“The Vaccine Doesn’t Work”*

6.5 Limitations and Open Challenges

In this Chapter we presented a first step towards leveraging insights from human experts to interactively discover latent themes from language data. To do this, we presented the experts with other analysis dimensions (e.g. opinions, sentiment and moral framing) to aid them in grounding and explaining the discovered themes. In our exploratory analysis, we divided the interaction into two distinct stages: leveraging human expertise to improve grounding, and leveraging human expertise to discover the space of relevant themes. The main limitation in our analysis is that we did not close the loop from discovery to grounding, and we did not evaluate the effect of multiple rounds of human interaction. Our current work tries to bridge this gap.

This research directions presents numerous challenges from multiple perspectives, including modeling and design, learning and grounding, and evaluation. First, we need to decide how to present information to experts so that they can effectively explore the space of language data. Given the large space of examples, this is not trivial. Moreover, the way we present information to the users will directly affect the outcome of the interaction. Coming up with strategies that strike a balance between exploitation and exploration is key to get the most out of the interaction, and to avoid inadvertently biasing the outcome.



(a) Nearest Neighbors

(b) DRaiL

Figure 6.20. Distribution of moral foundations for the theme *"Vaccine Appointments"*

Given the unsupervised nature of the task, we are presented with several modeling and learning challenges. In previous Chapters, we made the case for neural-symbolic and structured approaches in low supervision settings. Given the amount of data generated daily about COVID, there are broader opportunities for exploiting these resources than what we explored in this dissertation. We provided a preliminary analysis of the correlation between stances, reasons and morality, and exploited them in our grounding approach using distant supervision and some level of interaction. However, we would like to empower experts to instantiate and ground new concepts, as well as to provide us with high-level inference rules.

As we discussed in Chapter 4, the high computational cost of constrained inference makes structured approaches prohibitive. This challenge is exacerbated when dealing with massive datasets. Making these approaches fast and scalable is particularly important if we want to allow users to arbitrarily add additional analysis dimensions and dependencies.

Lastly, it is not clear how to best evaluate the interaction and its outcome. In this Chapter, we relied on a mix of qualitative and quantitative analyses, and combined intrinsic and extrinsic evaluations (e.g. how cohesive were the theme groundings, and whether they helped us make better stance predictions). The main challenge that arises is that the space of relevant labels is unknown. Moreover, it changes at each step of interaction. Our current work looks at designing evaluation protocols that measure the effectiveness of the interaction from both a user-centered and machine learning perspective.

7. SUMMARY

In this dissertation, we presented a comprehensive survey of existing neural-symbolic frameworks and identified the main challenges and opportunities that we face when trying to model complex language domains using existing approaches. To address these challenges, we introduced DRaiL, a declarative framework for combining neural and symbolic representations designed designed to support a variety of natural language scenarios. We showed the flexibility of DRaiL by modeling a diverse set of complex language scenarios dealing with rich linguistic and contextual structures. Further, we showed the advantages of DRaiL's modeling approach with respect to end-to-end neural networks, graph neural networks, traditional statistical relational learning approaches, as well as competitive neural-symbolic approaches.

One of the main reasons that precludes researchers and practitioners from incorporating constrained symbolic inference in their models is its high computational cost. In this dissertation, we showed that we can learn DRaiL models efficiently by leveraging approximate and randomized inference procedures, while still being able to incorporate arbitrary domain constraints. In addition to this, we motivated neural-symbolic models to deal with low-supervision scenarios. We showed that DRaiL can be used to encode knowledge and expectations about the language domain in a declarative way, and guide the learning process when direct supervision is not available. We showed that explicitly representing this information considerably reduces the amount of direct supervision needed to obtain competitive performance.

One of the main advantages of symbolic representations is their inherent interpretability. In this dissertation, we showed that we can leverage latent neural-symbolic variables to learn to explain higher-level decisions. By making latent variables discrete, we can easily interpret their meaning, introduce inductive bias into our models, and debug the learning process. By learning distributed representations for them, we are able to ground them in the textual data. We showed the advantages of this combined representation by contrasting it with latent variables in traditional graphical models, as well as with end-to-end neural networks.

Lastly, we presented a first step towards leveraging neural-symbolic representations to assist human experts in the process of making sense of large amounts of language data.

We showed that we can involve users in the process of discovering emerging themes from large language resources, as well as grounding them in the unlabeled data in an efficient, interactive way. We presented preliminary results using the COVID-19 vaccination debate, and included a detailed discussion on the open challenges and potential future directions.

REFERENCES

[1] R. Manhaeve, S. Dumancic, A. Kimmig, T. Demeester, and L. De Raedt, “Deepproblog: Neural probabilistic logic programming,” in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., Curran Associates, Inc., 2018, pp. 3749–3759. [Online]. Available: http://papers.nips.cc/paper/7632-deepproblog-neural-probabilistic-logic-programming.pdf.

[2] S. M. Kazemi and D. Poole, “Relnn: A deep neural model for relational learning,” in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, S. A. McIlraith and K. Q. Weinberger, Eds., AAAI Press, 2018, pp. 6367–6375. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16233.

[3] T. Rocktäschel and S. Riedel, “End-to-end differentiable proving,” in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017, pp. 3788–3800. [Online]. Available: http://papers.nips.cc/paper/6969-end-to-end-differentiable-proving.pdf.

[4] W. W. Cohen, F. Yang, and K. Mazaitis, “Tensorlog: A probabilistic database implemented using deep-learning infrastructure,” *J. Artif. Intell. Res.*, vol. 67, pp. 285–325, 2020. DOI: 10.1613/jair.1.11944. [Online]. Available: https://doi.org/10.1613/jair.1.11944.

[5] J. Haidt and J. Graham, “When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize,” *Social Justice Research*, vol. 20, no. 1, pp. 98–116, 2007.

[6] S. Roy, M. L. Pacheco, and D. Goldwasser, “Identifying morality frames in political tweets using relational learning,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 9939–9958. DOI: 10.18653/v1/2021.emnlp-main.783. [Online]. Available: https://aclanthology.org/2021.emnlp-main.783.

[7] M. Zamani, H. A. Schwartz, J. Eichstaedt, S. C. Guntuku, A. Virinchipuram Ganesan, S. Clouston, and S. Giorgi, “Understanding weekly COVID-19 concerns through dynamic content-specific LDA topic modeling,” in *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, Online: Association for Computational Linguistics, Nov. 2020, pp. 193–198. DOI: 10.18653/v1/2020.nlpcss-1.21. [Online]. Available: https://aclanthology.org/2020.nlpcss-1.21.

[8] N. Reimers, B. Schiller, T. Beck, J. Daxenberger, C. Stab, and I. Gurevych, "Classification and clustering of arguments with contextualized word embeddings," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 567–578. DOI: 10.18653/v1/P19-1054. [Online]. Available: https://aclanthology.org/P19-1054.

[9] K. S. Hasan and V. Ng, "Why are you taking this stance? identifying and classifying reasons in ideological debates," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 751–762. DOI: 10.3115/v1/D14-1083. [Online]. Available: https://aclanthology.org/D14-1083.

[10] M. L. Pacheco and D. Goldwasser, "Modeling content and context with deep relational learning," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 100–119, 2021. DOI: 10.1162/tacl_a_00357. [Online]. Available: https://aclanthology.org/2021.tacl-1.7.

[11] M. Widmoser, M. L. Pacheco, J. Honorio, and D. Goldwasser, "Randomized deep structured prediction for discourse-level processing," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online: Association for Computational Linguistics, Apr. 2021, pp. 1174–1184. DOI: 10.18653/v1/2021.eacl-main.100. [Online]. Available: https://aclanthology.org/2021.eacl-main.100.

[12] A. Jain, M. L. Pacheco, S. Lancette, M. Goindani, and D. Goldwasser, "Identifying collaborative conversations using latent discourse behaviors," in *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 1st virtual meeting: Association for Computational Linguistics, Jul. 2020, pp. 74–78. [Online]. Available: https://aclanthology.org/2020.sigdial-1.10.

[13] A. Jain, "Using latent discourse indicators to identify goodness in online conversations," An optional note, M.S. thesis, Purdue University, Purdue University Graduate School, Jan. 2020.

[14] M. L. Pacheco, T. Islam, M. Mahajan, A. Shor, M. Yin, L. Ungar, and D. Goldwasser, "A holistic framework for analyzing the covid-19 vaccine debate," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, 2022.

[15] B. Milch, B. Marthi, S. Russell, D. Sontag, D. L. Ong, and A. Kolobov, "Blog: Probabilistic models with unknown objects," in *(IJCAI '05) Nineteenth International Joint Conference on Artificial Intelligence*, Jul. 2005. [Online]. Available: https://www.microsoft.com/en-us/research/publication/blog-probabilistic-models-unknown-objects-2/.

[16] N. D. Goodman, V. K. Mansinghka, D. M. Roy, K. Bonawitz, and J. B. Tenenbaum, “Church: A language for generative models,” in *UAI 2008, Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence, Helsinki, Finland, July 9-12, 2008*, 2008, pp. 220–229. [Online]. Available: https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1%5C&smnu=2%5C&article%5C_id=1346%5C&proceeding%5C_id=24.

[17] M. Richardson and P. M. Domingos, “Markov logic networks,” *Mach. Learn.*, vol. 62, no. 1-2, pp. 107–136, 2006. DOI: 10.1007/s10994-006-5833-1. [Online]. Available: https://doi.org/10.1007/s10994-006-5833-1.

[18] S. H. Bach, M. Broecheler, B. Huang, and L. Getoor, “Hinge-loss markov random fields and probabilistic soft logic,” *Journal of Machine Learning Research (JMLR)*, vol. 18, pp. 1–67, 2017. [Online]. Available: https://github.com/stephenbach/bach-jmlr17-code.

[19] A. McCallum, K. Schultz, and S. Singh, “Factorie: Probabilistic programming via imperatively defined factor graphs,” in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds., Curran Associates, Inc., 2009, pp. 1249–1257. [Online]. Available: http://papers.nips.cc/paper/3654-factorie-probabilistic-programming-via-imperatively-defined-factor-graphs.pdf.

[20] N. Rizzolo and D. Roth, “Learning based Java for rapid development of NLP systems,” in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, Valletta, Malta: European Language Resources Association (ELRA), May 2010. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/747_Paper.pdf.

[21] P. Kordjamshidi, D. Roth, and H. Wu, “Saul: Towards declarative learning based programming,” in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, 2015, pp. 1844–1851. [Online]. Available: http://ijcai.org/Abstract/15/262.

[22] W. Y. Wang, K. Mazaitis, and W. W. Cohen, “Programming with personalized pagerank: A locally groundable first-order probabilistic logic,” in *22nd ACM International Conference on Information and Knowledge Management, CIKM’13, San Francisco, CA, USA, October 27 - November 1, 2013*, Q. He, A. Iyengar, W. Nejdl, J. Pei, and R. Rastogi, Eds., ACM, 2013, pp. 2129–2138. DOI: 10.1145/2505515.2505573. [Online]. Available: https://doi.org/10.1145/2505515.2505573.

[23] B. Perozzi, R. Al-Rfou, and S. Skiena, “Deepwalk: Online learning of social representations,” in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’14, New York, New York, USA: ACM, 2014, pp. 701–710, ISBN: 978-1-4503-2956-9. DOI: 10.1145/2623330.2623732. [Online]. Available: http://doi.acm.org/10.1145/2623330.2623732.

[24] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, “LINE: large-scale information network embedding,” in *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*, 2015, pp. 1067–1077. DOI: 10.1145/2736277.2741093. [Online]. Available: https://doi.org/10.1145/2736277.2741093.

[25] S. Pan, J. Wu, X. Zhu, C. Zhang, and Y. Wang, “Tri-party deep network representation,” in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, 2016, pp. 1895–1901. [Online]. Available: http://www.ijcai.org/Abstract/16/271.

[26] A. Grover and J. Leskovec, “Node2vec: Scalable feature learning for networks,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 2016, pp. 855–864. DOI: 10.1145/2939672.2939754. [Online]. Available: https://doi.org/10.1145/2939672.2939754.

[27] C. Tu, H. Liu, Z. Liu, and M. Sun, “CANE: Context-aware network embedding for relation modeling,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1722–1731. DOI: 10.18653/v1/P17-1158. [Online]. Available: https://www.aclweb.org/anthology/P17-1158.

[28] H. Xiao, M. Huang, L. Meng, and X. Zhu, “SSP: semantic space projection for knowledge graph embedding with text descriptions,” in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, 2017, pp. 3104–3110. [Online]. Available: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14306.

[29] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, “Translating embeddings for modeling multi-relational data,” in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2013, pp. 2787–2795. [Online]. Available: http://papers.nips.cc/paper/5071-translating-embeddings-for-modeling-multi-relational-data.pdf.

[30] Z. Wang, J. Zhang, J. Feng, and Z. Chen, “Knowledge graph embedding by translating on hyperplanes,” in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada*, 2014, pp. 1112–1119. [Online]. Available: http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8531.

[31] T. Trouillon, J. Welbl, S. Riedel, E. Gaussier, and G. Bouchard, “Complex embeddings for simple link prediction,” in *Proceedings of The 33rd International Conference on Machine Learning*, M. F. Balcan and K. Q. Weinberger, Eds., ser. Proceedings of Machine Learning Research, vol. 48, New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 2071–2080. [Online]. Available: http://proceedings.mlr.press/v48/trouillon16.html.

[32] Z. Sun, Z.-H. Deng, J.-Y. Nie, and J. Tang, "Rotate: Knowledge graph embedding by relational rotation in complex space," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. [Online]. Available: https://openreview.net/forum?id=HkgEQnRqYQ.

[33] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. [Online]. Available: https://openreview.net/forum?id=SJU4ayYgl.

[34] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017, pp. 1024–1034. [Online]. Available: http://papers.nips.cc/paper/6703-inductive-representation-learning-on-large-graphs.pdf.

[35] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[36] G. Sourek, V. Aschenbrenner, F. Zelezný, S. Schockaert, and O. Kuzelka, "Lifted relational neural networks: Efficient learning of latent relational structures," *J. Artif. Intell. Res.*, vol. 62, pp. 69–100, 2018. DOI: 10.1613/jair.1.11203. [Online]. Available: https://doi.org/10.1613/jair.1.11203.

[37] I. Donadello, L. Serafini, and A. d'Avila Garcez, "Logic tensor networks for semantic image interpretation," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 1596–1602. DOI: 10.24963/ijcai.2017/221. [Online]. Available: https://doi.org/10.24963/ijcai.2017/221.

[38] F. Yang, Z. Yang, and W. W. Cohen, "Differentiable learning of logical rules for knowledge base reasoning," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017, pp. 2319–2328. [Online]. Available: http://papers.nips.cc/paper/6826-differentiable-learning-of-logical-rules-for-knowledge-base-reasoning.pdf.

[39] A. Sadeghian, M. Armandpour, P. Ding, and D. Z. Wang, "Drum: End-to-end differentiable rule mining on knowledge graphs," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019, pp. 15 347–15 357. [Online]. Available: http://papers.nips.cc/paper/9669-drum-end-to-end-differentiable-rule-mining-on-knowledge-graphs.pdf.

[40] H. Dong, J. Mao, T. Lin, C. Wang, L. Li, and D. Zhou, “Neural logic machines,” in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. [Online]. Available: https://openreview.net/forum?id=B1xY-hRctX.

[41] H. Wang and H. Poon, “Deep probabilistic logic: A unifying framework for indirect supervision,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 1891–1902. DOI: 10.18653/v1/D18-1215. [Online]. Available: https://www.aclweb.org/anthology/D18-1215.

[42] G. Marra, F. Giannini, M. Diligenti, and M. Gori, “Integrating learning and reasoning with deep logic models,” in *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2019, Würzburg, Germany, September 16-20, 2019, Proceedings, Part II*, 2019, pp. 517–532. DOI: 10.1007/978-3-030-46147-8_31. [Online]. Available: https://doi.org/10.1007/978-3-030-46147-8%5C_31.

[43] D. Chen and C. Manning, “A fast and accurate dependency parser using neural networks,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 740–750. DOI: 10.3115/v1/D14-1082. [Online]. Available: https://www.aclweb.org/anthology/D14-1082.

[44] D. Weiss, C. Alberti, M. Collins, and S. Petrov, “Structured training for neural network transition-based parsing,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 323–333. DOI: 10.3115/v1/P15-1032. [Online]. Available: https://www.aclweb.org/anthology/P15-1032.

[45] X. Ma and E. Hovy, “End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1064–1074. DOI: 10.18653/v1/P16-1101. [Online]. Available: https://www.aclweb.org/anthology/P16-1101.

[46] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, “Neural architectures for named entity recognition,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 260–270. DOI: 10.18653/v1/N16-1030. [Online]. Available: https://www.aclweb.org/anthology/N16-1030.

[47] E. Kiperwasser and Y. Goldberg, “Simple and accurate dependency parsing using bidirectional LSTM feature representations,” *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 313–327, 2016. DOI: 10.1162/tacl_a_00101. [Online]. Available: https://www.aclweb.org/anthology/Q16-1023.

[48] C. Malaviya, M. R. Gormley, and G. Neubig, “Neural factor graph models for cross-lingual morphological tagging,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2653–2663. DOI: 10.18653/v1/P18-1247. [Online]. Available: https://www.aclweb.org/anthology/P18-1247.

[49] I. Beltagy, K. Erk, and R. Mooney, “Probabilistic soft logic for semantic textual similarity,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, Maryland: Association for Computational Linguistics, Jun. 2014, pp. 1210–1219. DOI: 10.3115/v1/P14-1114. [Online]. Available: https://www.aclweb.org/anthology/P14-1114.

[50] D. Sridhar, J. Foulds, B. Huang, L. Getoor, and M. Walker, “Joint models of disagreement and stance in online debate,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 116–125. DOI: 10.3115/v1/P15-1012. [Online]. Available: https://www.aclweb.org/anthology/P15-1012.

[51] S. Liu, Y. Chen, S. He, K. Liu, and J. Zhao, “Leveraging FrameNet to improve automatic event detection,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 2134–2143. DOI: 10.18653/v1/P16-1201. [Online]. Available: https://www.aclweb.org/anthology/P16-1201.

[52] S. Subramanian, T. Cohn, T. Baldwin, and J. Brooke, “Joint sentence-document model for manifesto text analysis,” in *Proceedings of the Australasian Language Technology Association Workshop 2017*, Brisbane, Australia, Dec. 2017, pp. 25–33. [Online]. Available: https://www.aclweb.org/anthology/U17-1003.

[53] Q. Ning, Z. Feng, H. Wu, and D. Roth, “Joint reasoning for temporal and causal relations,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2278–2288. DOI: 10.18653/v1/P18-1212. [Online]. Available: https://www.aclweb.org/anthology/P18-1212.

[54] V. Niculae, J. Park, and C. Cardie, “Argument mining with structured SVMs and RNNs,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 985–995. DOI: 10.18653/v1/P17-1091. [Online]. Available: https://www.aclweb.org/anthology/P17-1091.

[55] R. Han, Q. Ning, and N. Peng, “Joint event and temporal relation extraction with shared representations and structured prediction,” 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)*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 434–444. DOI: 10.18653/v1/D19-1041. [Online]. Available: https://www.aclweb.org/anthology/D19-1041.

[56] Y. Zhang, T. Lei, R. Barzilay, and T. S. Jaakkola, “Greed is good if randomized: New inference for dependency parsing,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, 2014, pp. 1013–1024. DOI: 10.3115/v1/d14-1109. [Online]. Available: https://doi.org/10.3115/v1/d14-1109.

[57] Y. Zhang, C. Li, R. Barzilay, and K. Darwish, “Randomized greedy inference for joint segmentation, POS tagging and dependency parsing,” in *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, 2015, pp. 42–52. DOI: 10.3115/v1/n15-1005. [Online]. Available: https://doi.org/10.3115/v1/n15-1005.

[58] J. Honorio and T. S. Jaakkola, “Structured prediction: From gaussian perturbations to linear-time principled algorithms,” in *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence, UAI 2016, June 25-29, 2016, New York City, NY, USA*, 2016. [Online]. Available: http://auai.org/uai2016/proceedings/papers/79.pdf.

[59] C. Ma, F. A. R. R. Chowdhury, A. Deshwal, M. R. Islam, J. R. Doppa, and D. Roth, “Randomized greedy search for structured prediction: Amortized inference and learning,” in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, 2019, pp. 5130–5138. DOI: 10.24963/ijcai.2019/713. [Online]. Available: https://doi.org/10.24963/ijcai.2019/713.

[60] L. van der Maaten, M. Welling, and L. Saul, “Hidden-unit conditional random fields,” in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, G. Gordon, D. Dunson, and M. Dudík, Eds., ser. Proceedings of Machine Learning Research, vol. 15, Fort Lauderdale, FL, USA: JMLR Workshop and Conference Proceedings, Nov. 2011, pp. 479–488. [Online]. Available: http://proceedings.mlr.press/v15/maaten11b.html.

[61] S. Kok and P. Domingos, “Statistical predicate invention,” in *Proceedings of the 24th International Conference on Machine Learning*, ser. ICML ’07, Corvalis, Oregon, USA: Association for Computing Machinery, 2007, pp. 433–440, ISBN: 9781595937933. DOI: 10.1145/1273496.1273551. [Online]. Available: https://doi.org/10.1145/1273496.1273551.

[62] S. H. Bach, B. Huang, J. Boyd-Graber, and L. Getoor, “Paired-dual learning for fast training of latent variable hinge-loss mrfs,” in *International Conference on Machine Learning (ICML)*, Stephen Bach and Bert Huang contributed equally., 2015.

[63] S. Bach, B. Huang, and L. Getoor, “Learning latent groups with hinge-loss markov random fields,” in *ICML Workshop on Inferning: Interactions between Inference and Learning*, 2013. [Online]. Available: https://linqs.soe.ucsc.edu/sites/default/files/papers/bach-inferning13.pdf.

[64] Y. Ji, G. Haffari, and J. Eisenstein, “A latent variable recurrent neural network for discourse-driven language models,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 332–342. DOI: 10.18653/v1/N16-1037. [Online]. Available: https://www.aclweb.org/anthology/N16-1037.

[65] M. Zaheer, A. Ahmed, and A. J. Smola, “Latent LSTM allocation: Joint clustering and non-linear dynamic modeling of sequence data,” in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., ser. Proceedings of Machine Learning Research, vol. 70, International Convention Centre, Sydney, Australia: PMLR, Jun. 2017, pp. 3967–3976. [Online]. Available: http://proceedings.mlr.press/v70/zaheer17a.html.

[66] X. Zheng, M. Zaheer, A. Ahmed, Y. Wang, A. J. Smola, and E. Xing, “State space lstm models with particle mcmc inference,” 2017.

[67] P. Lertvittayakumjorn and F. Toni, “Explanation-based human debugging of NLP models: A survey,” *CoRR*, vol. abs/2104.15135, 2021. arXiv: 2104.15135. [Online]. Available: https://arxiv.org/abs/2104.15135.

[68] M. T. Ribeiro, S. Singh, and C. Guestrin, “”why should i trust you?”: Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’16, San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 1135–1144, ISBN: 9781450342322. DOI: 10.1145/2939672.2939778. [Online]. Available: https://doi.org/10.1145/2939672.2939778.

[69] P. Lertvittayakumjorn, L. Specia, and F. Toni, "FIND: Human-in-the-Loop Debugging Deep Text Classifiers," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 332–348. DOI: 10.18653/v1/2020.emnlp-main.24. [Online]. Available: https://aclanthology.org/2020.emnlp-main.24.

[70] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, e0130140, Jul. 2015. DOI: 10.1371/journal.pone.0130140. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0130140.

[71] H. Zylberajch, P. Lertvittayakumjorn, and F. Toni, "HILDIF: Interactive debugging of NLI models using influence functions," in *Proceedings of the First Workshop on Interactive Learning for Natural Language Processing*, Online: Association for Computational Linguistics, Aug. 2021, pp. 1–6. DOI: 10.18653/v1/2021.internlp-1.1. [Online]. Available: https://aclanthology.org/2021.internlp-1.1.

[72] H. Yao, Y. Chen, Q. Ye, X. Jin, and X. Ren, "Refining neural networks with compositional explanations," *CoRR*, vol. abs/2103.10415, 2021. arXiv: 2103.10415. [Online]. Available: https://arxiv.org/abs/2103.10415.

[73] A. Ratner, C. D. Sa, S. Wu, D. Selsam, and C. Ré, "Data programming: Creating large training sets, quickly," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16, Barcelona, Spain: Curran Associates Inc., 2016, pp. 3574–3582, ISBN: 9781510838819.

[74] A. Ratner, S. H. Bach, H. R. Ehrenberg, J. A. Fries, S. Wu, and C. Ré, "Snorkel: Rapid training data creation with weak supervision," *Proc. VLDB Endow.*, vol. 11, no. 3, pp. 269–282, 2017. DOI: 10.14778/3157794.3157797. [Online]. Available: http://www.vldb.org/pvldb/vol11/p269-ratner.pdf.

[75] I. Habernal, H. Wachsmuth, I. Gurevych, and B. Stein, "The argument reasoning comprehension task: Identification and reconstruction of implicit warrants," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 1930–1940. DOI: 10.18653/v1/N18-1175. [Online]. Available: https://aclanthology.org/N18-1175.

[76] J. Lawrence and C. Reed, "Argument mining: A survey," *Computational Linguistics*, vol. 45, no. 4, pp. 765–818, Dec. 2019. DOI: 10.1162/coli_a_00364. [Online]. Available: https://aclanthology.org/J19-4006.

[77] C. Stab and I. Gurevych, “Parsing argumentation structures in persuasive essays,” *Comput. Linguist.*, vol. 43, no. 3, pp. 619–659, Sep. 2017, ISSN: 0891-2017. DOI: 10.1162/COLI_a_00295. [Online]. Available: https://doi.org/10.1162/COLI_a_00295.

[78] A. F. T. Martins, M. A. T. Figueiredo, P. M. Q. Aguiar, N. A. Smith, and E. P. Xing, “Ad3: Alternating directions dual decomposition for map inference in graphical models,” *Journal of Machine Learning Research*, vol. 16, no. 16, pp. 495–545, 2015. [Online]. Available: http://jmlr.org/papers/v16/martins15a.html.

[79] S. Somasundaran and J. Wiebe, “Recognizing stances in ideological on-line debates,” in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Los Angeles, CA: Association for Computational Linguistics, Jun. 2010, pp. 116–124. [Online]. Available: https://www.aclweb.org/anthology/W10-0214.

[80] Q. Sun, Z. Wang, Q. Zhu, and G. Zhou, “Stance detection with hierarchical attention network,” in *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, 2018, pp. 2399–2409. [Online]. Available: https://www.aclweb.org/anthology/C18-1203/.

[81] M. Walker, P. Anand, R. Abbott, and R. Grant, “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*, Montréal, Canada: Association for Computational Linguistics, Jun. 2012, pp. 592–596. [Online]. Available: https://www.aclweb.org/anthology/N12-1072.

[82] K. S. Hasan and V. Ng, “Stance classification of ideological debates: Data, models, features, and constraints,” in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, Nagoya, Japan: Asian Federation of Natural Language Processing, Oct. 2013, pp. 1348–1356. [Online]. Available: https://www.aclweb.org/anthology/I13-1191.

[83] C. Li, A. Porco, and D. Goldwasser, “Structured representation learning for online debate stance prediction,” in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 3728–3739. [Online]. Available: https://www.aclweb.org/anthology/C18-1316.

[84] C. Xu, C. Paris, S. Nepal, and R. Sparks, “Cross-target stance classification with self-attention networks,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 778–783.

[85] I. Augenstein, T. Rocktäschel, A. Vlachos, and K. Bontcheva, “Stance detection with bidirectional conditional encoding,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas: Association for Computational Linguistics, 2016, pp. 876–885.

[86] D. Pomerleau and D. Rao, *The fake news challenge*, http://www.fakenewschallenge.org/, 2017.

[87] A. Hanselowski, A. PVS, B. Schiller, F. Caspelherr, D. Chaudhuri, C. M. Meyer, and I. Gurevych, “A retrospective analysis of the fake news challenge stance-detection task,” in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018, pp. 1859–1874.

[88] B. Ghanem, P. Rosso, and F. Rangel, “Stance detection in fake news a combined feature representation,” in *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 66–71.

[89] J. Zhang, R. Kumar, S. Ravi, and C. Danescu-Niculescu-Mizil, “Conversational flow in oxford-style debates,” in *Proceedings of NAACL-HLT*, 2016, pp. 136–141.

[90] P. Potash and A. Rumshisky, “Towards debate automation: A recurrent model for predicting debate winners,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2465–2475.

[91] L. Wang, N. Beauchamp, S. Shugars, and K. Qin, “Winning on the merits: The joint effects of content and style on debate outcomes,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 219–232, 2017.

[92] E. Durmus and C. Cardie, “Exploring the role of prior beliefs for argument persuasion,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, vol. 1, 2018, pp. 1035–1045.

[93] J. Haidt and C. Joseph, “Intuitive ethics: How innately prepared intuitions generate culturally variable virtues,” *Daedalus*, vol. 133, no. 4, pp. 55–66, 2004.

[94] R. M. Entman, “Framing: Toward clarification of a fractured paradigm,” *Journal of communication*, vol. 43, no. 4, pp. 51–58, 1993.

[95] D. Chong and J. N. Druckman, “Framing theory,” *Annu. Rev. Polit. Sci.*, vol. 10, pp. 103–126, 2007.

[96] A. Boydstun, D. Card, J. H. Gross, P. Resnik, and N. A. Smith, “Tracking the development of media frames within and across policy issues,” 2014.

[97] J. Hoover, K. Johnson, R. Boghrati, J. Graham, M. Dehghani, and M. B. Donnellan, “Moral framing and charitable donation: Integrating exploratory social media analyses and confirmatory experimentation,” *Collabra: Psychology*, vol. 4, no. 1, 2018.

[98] M. Mooijman, J. Hoover, Y. Lin, H. Ji, and M. Dehghani, “Moralization in social networks and the emergence of violence during protests,” *Nature human behaviour*, vol. 2, no. 6, pp. 389–396, 2018.

[99] M. Dehghani, K. Johnson, J. Hoover, E. Sagi, J. Garten, N. J. Parmar, S. Vaisey, R. Iliev, and J. Graham, “Purity homophily in social networks.,” *Journal of Experimental Psychology: General*, vol. 145, no. 3, p. 366, 2016.

[100] O. Tsur, D. Calacci, and D. Lazer, “A frame of mind: Using statistical models for detection of framing and agenda setting campaigns,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 1629–1638. DOI: 10.3115/v1/P15-1157. [Online]. Available: https://aclanthology.org/P15-1157.

[101] E. Baumer, E. Elovic, Y. Qin, F. Polletta, and G. Gay, “Testing and comparing computational approaches for identifying the language of framing in political news,” in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado: Association for Computational Linguistics, May 2015, pp. 1472–1482. DOI: 10.3115/v1/N15-1171. [Online]. Available: https://aclanthology.org/N15-1171.

[102] D. Card, A. E. Boydstun, J. H. Gross, P. Resnik, and N. A. Smith, “The media frames corpus: Annotations of frames across issues,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 438–444. DOI: 10.3115/v1/P15-2072. [Online]. Available: https://aclanthology.org/P15-2072.

[103] A. Field, D. Kliger, S. Wintner, J. Pan, D. Jurafsky, and Y. Tsvetkov, “Framing and agenda-setting in Russian news: A computational analysis of intricate political strategies,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 3570–3580. DOI: 10.18653/v1/D18-1393. [Online]. Available: https://aclanthology.org/D18-1393.

[104] D. Demszky, N. Garg, R. Voigt, J. Zou, J. Shapiro, M. Gentzkow, and D. Jurafsky, “Analyzing polarization in social media: Method and application to tweets on 21 mass shootings,” 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)*, 2019, pp. 2970–3005.

[105] L. Fan, M. White, E. Sharma, R. Su, P. K. Choubey, R. Huang, and L. Wang, "In plain sight: Media bias through the lens of factual reporting," 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)*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6343–6349. DOI: 10.18653/v1/D19-1664. [Online]. Available: https://aclanthology.org/D19-1664.

[106] S. Roy and D. Goldwasser, "Weakly supervised learning of nuanced frames for analyzing polarization in news media," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 7698–7716. DOI: 10.18653/v1/2020.emnlp-main.620. [Online]. Available: https://aclanthology.org/2020.emnlp-main.620.

[107] D. Fulgoni, J. Carpenter, L. Ungar, and D. Preoţiuc-Pietro, "An empirical exploration of moral foundations theory in partisan news sources," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 3730–3736. [Online]. Available: https://aclanthology.org/L16-1591.

[108] M. Dehghani, K. Sagae, S. Sachdeva, and J. Gratch, "Analyzing political rhetoric in conservative and liberal weblogs related to the construction of the "ground zero mosque"," *Journal of Information Technology & Politics*, vol. 11, no. 1, pp. 1–14, 2014.

[109] W. J. Brady, J. A. Wills, J. T. Jost, J. A. Tucker, and J. J. Van Bavel, "Emotion shapes the diffusion of moralized content in social networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 28, pp. 7313–7318, 2017.

[110] K. Johnson and D. Goldwasser, "Classification of moral foundations in microblog political discourse," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 720–730. DOI: 10.18653/v1/P18-1067. [Online]. Available: https://aclanthology.org/P18-1067.

[111] J. Hoover, G. Portillo-Wightman, L. Yeh, S. Havaldar, A. M. Davani, Y. Lin, B. Kennedy, M. Atari, Z. Kamel, M. Mendlen, *et al.*, "Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment," *Social Psychological and Personality Science*, vol. 11, no. 8, pp. 1057–1071, 2020.

[112] Y. Lin, J. Hoover, G. Portillo-Wightman, C. Park, M. Dehghani, and H. Ji, "Acquiring background knowledge to improve moral value prediction," in *2018 ieee/acm international conference on advances in social networks analysis and mining (asonam)*, IEEE, 2018, pp. 552–559.

[113] J. Y. Xie, R. F. P. Junior, G. Hirst, and Y. Xu, "Text-based inference of moral sentiment change," 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)*, 2019, pp. 4646–4655.

[114] J. Garten, R. Boghrati, J. Hoover, K. M. Johnson, and M. Dehghani, "Morality between the lines: Detecting moral sentiment in text," in *Proceedings of IJCAI 2016 workshop on Computational Modeling of Attitudes*, 2016.

[115] J. Graham, J. Haidt, and B. Nosek, *Graham, Haidt, & Nosek (2009): Liberals and conservatives rely on different sets of moral foundations*, version V1, 2009. DOI: 10.7910/DVN/SJTRBI. [Online]. Available: https://doi.org/10.7910/DVN/SJTRBI.

[116] S. Roy and D. Goldwasser, "Analysis of nuanced stances and sentiment towards entities of US politicians through the lens of moral foundation theory," in *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, Online: Association for Computational Linguistics, Jun. 2021, pp. 1–13. DOI: 10.18653/v1/2021.socialnlp-1.1. [Online]. Available: https://aclanthology.org/2021.socialnlp-1.1.

[117] L. Deng and J. Wiebe, "Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 179–189. DOI: 10.18653/v1/D15-1018. [Online]. Available: https://aclanthology.org/D15-1018.

[118] A. Field and Y. Tsvetkov, "Entity-centric contextual affective analysis," *arXiv preprint arXiv:1906.01762*, 2019.

[119] C. Y. Park, X. Yan, A. Field, and Y. Tsvetkov, "Multilingual contextual affective analysis of lgbt people portrayals in wikipedia," *arXiv preprint arXiv:2010.10820*, 2020.

[120] S. Loomba, A. de Figueiredo, S. J. Piatek, K. de Graaf, and H. J. Larson, "Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa," *Nature human behaviour*, vol. 5, no. 3, pp. 337–348, 2021.

[121] G. K. Shahi, A. Dirkson, and T. A. Majchrzak, "An exploratory study of covid-19 misinformation on twitter," *Online social networks and media*, vol. 22, p. 100 104, 2021.

[122] J. V. Lazarus, S. C. Ratzan, A. Palayew, L. O. Gostin, H. J. Larson, K. Rabin, S. Kimball, and A. El-Mohandes, "A global survey of potential acceptance of a covid-19 vaccine," *Nature medicine*, vol. 27, no. 2, pp. 225–228, 2021.

[123] W. Ahmed, J. Vidal-Alaball, J. Downing, F. L. Seguí, *et al.*, "Covid-19 and the 5g conspiracy theory: Social network analysis of twitter data," *Journal of medical internet research*, vol. 22, no. 5, e19458, 2020.

[124] M. A. Weinzierl and S. M. Harabagiu, "Automatic detection of covid-19 vaccine misinformation with graph link prediction," *Journal of biomedical informatics*, vol. 124, p. 103 955, 2021.

[125] Y. Bang, E. Ishii, S. Cahyawijaya, Z. Ji, and P. Fung, "Model generalization on covid-19 fake news detection," *arXiv preprint arXiv:2101.03841*, 2021.

[126] J. C. M. Serrano, O. Papakyriakopoulos, and S. Hegelich, "Nlp-based feature extraction for the detection of covid-19 misinformation videos on youtube," in *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, 2020.

[127] M. S. Al-Rakhami and A. M. Al-Amri, "Lies kill, facts save: Detecting covid-19 misinformation in twitter," *Ieee Access*, vol. 8, pp. 155 961–155 970, 2020.

[128] F. M. Alliheibi, A. Omar, and N. Al-Horais, "Opinion mining of saudi responses to covid-19 vaccines on twitter," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, pp. 72–78, 2021.

[129] J. C. Lyu, E. Le Han, and G. K. Luli, "Covid-19 vaccine–related discussion on twitter: Topic modeling and sentiment analysis," *Journal of medical Internet research*, vol. 23, no. 6, e24435, 2021.

[130] M. Skeppstedt, A. Kerren, and M. Stede, "Vaccine hesitancy in discussion forums: Computer-assisted argument mining with topic models," in *Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth*, IOS Press, 2018, pp. 366–370.

[131] H. Sha, M. A. Hasan, G. Mohler, and P. J. Brantingham, "Dynamic topic modeling of the covid-19 twitter narrative among us governors and cabinet executives," *arXiv preprint arXiv:2004.11692*, 2020.

[132] M. Zamani, H. A. Schwartz, J. Eichstaedt, S. C. Guntuku, A. V. Ganesan, S. Clouston, and S. Giorgi, "Understanding weekly covid-19 concerns through dynamic content-specific lda topic modeling," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, NIH Public Access, vol. 2020, 2020, p. 193.

[133] L. Wang and C. Cardie, "A piece of my mind: A sentiment analysis approach for online dispute detection," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Baltimore, Maryland: Association for Computational Linguistics, Jun. 2014, pp. 693–699.

[134] T. Althoff, K. Clark, and J. Leskovec, “Large-scale analysis of counseling conversations: An application of natural language processing to mental health,” *Transactions of the Association for Computational Linguistics*, vol. 4, p. 463, 2016.

[135] V. Niculae and C. Danescu-Niculescu-Mizil, “Conversational markers of constructive discussions,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 568–578. DOI: 10.18653/v1/N16-1070. [Online]. Available: https://aclanthology.org/N16-1070.

[136] C. Napoles, A. Pappu, and J. Tetreault, “Automatically identifying good conversations online (yes, they do exist!)” In *Proceedings of the International AAAI Conference on Web and Social Media*, 2017.

[137] C. Napoles, J. Tetreault, E. Rosata, B. Provenzale, and A. Pappu, “Finding good conversations online: The yahoo news annotated comments corpus,” in *Proceedings of The 11th Linguistic Annotation Workshop*, Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 13–23.

[138] C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts, “A computational approach to politeness with application to social factors,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2013, pp. 250–259.

[139] C. Tan, V. Niculae, C. Danescu-Niculescu-Mizil, and L. Lee, “Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions,” in *Proceedings of the 25th international conference on world wide web*, International World Wide Web Conferences Steering Committee, 2016, pp. 613–624.

[140] Z. Wei, Y. Liu, and Y. Li, “Is this post persuasive? ranking argumentative comments in online forum,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, 2016, pp. 195–200.

[141] I. Habernal and I. Gurevych, “What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas: Association for Computational Linguistics, 2016, pp. 1214–1223.

[142] S. Lukin, P. Anand, M. Walker, and S. Whittaker, “Argument strength is in the eye of the beholder: Audience effects in persuasion,” in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, vol. 1, 2017, pp. 742–753.

[143] S. Chaturvedi, D. Goldwasser, and H. Daumé III, "Predicting instructor's intervention in mooc forums," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2014, pp. 1501–1511.

[144] A. X. Zhang, B. Culbertson, and P. Paritosh, "Characterizing online discussion using coarse discourse sequences," in *Proceedings of the Eleventh International Conference on Web and Social Media*, 2017.

[145] A. Misra and M. Walker, "Topic independent identification of agreement and disagreement in social media dialogue," in *Proceedings of the SIGDIAL 2013 Conference*, 2013, pp. 41–50.

[146] J. J. Gumperz, "Contextualization revisited," pp. 39–54, 1992.

[147] A. Duranti and C. Goodwin, "Rethinking context: An introduction," in *Rethinking Context: Language as an Interactive Phenomenon*, A. Duranti and C. Goodwin, Eds., Cambridge: Cambridge University Press, 1992, ch. 1, pp. 1–42. [Online]. Available: http://www.sscnet.ucla.edu/anthro/faculty/duranti/reprints/rethco.pdf.

[148] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 632–642. DOI: 10.18653/v1/D15-1075. [Online]. Available: https://www.aclweb.org/anthology/D15-1075.

[149] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. [Online]. Available: https://www.aclweb.org/anthology/N18-1202.

[150] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. [Online]. Available: https://www.aclweb.org/anthology/N19-1423.

[151] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[152] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu, “The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision,” 2019. [Online]. Available: https://openreview.net/forum?id=rJgMlhRctm.

[153] I. Gurobi Optimization, *Gurobi optimizer reference manual*, 2015. [Online]. Available: http://www.gurobi.com.

[154] H. Zhou, Y. Zhang, S. Huang, and J. Chen, “A neural probabilistic structured-prediction model for transition-based dependency parsing,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 1213–1222. DOI: 10.3115/v1/P15-1117. [Online]. Available: https://www.aclweb.org/anthology/P15-1117.

[155] D. Andor, C. Alberti, D. Weiss, A. Severyn, A. Presta, K. Ganchev, S. Petrov, and M. Collins, “Globally normalized transition-based neural networks,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 2442–2452. DOI: 10.18653/v1/P16-1231. [Online]. Available: https://www.aclweb.org/anthology/P16-1231.

[156] I.-T. Lee and D. Goldwasser, “Multi-relational script learning for discourse relations,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4214–4226. DOI: 10.18653/v1/P19-1413. [Online]. Available: https://www.aclweb.org/anthology/P19-1413.

[157] X. Zhang and D. Goldwasser, “Sentiment tagging with partial labels using modular architectures,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 579–590. DOI: 10.18653/v1/P19-1055. [Online]. Available: https://www.aclweb.org/anthology/P19-1055.

[158] I. Turc, M.-W. Chang, K. Lee, and K. Toutanova, “Well-read students learn better: On the importance of pre-training compact models,” *arXiv preprint arXiv:1908.08962v2*, 2019.

[159] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. [Online]. Available: https://www.aclweb.org/anthology/D14-1162.

[160] T. Kuribayashi, H. Ouchi, N. Inoue, P. Reisert, T. Miyoshi, J. Suzuki, and K. Inui, "An empirical study of span representations in argumentation structure parsing," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4691–4698. DOI: 10.18653/v1/P19-1464. [Online]. Available: https://www.aclweb.org/anthology/P19-1464.

[161] P. Potash, A. Romanov, and A. Rumshisky, "Here's my point: Joint pointer architecture for argument mining," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 1364–1373. DOI: 10.18653/v1/D17-1143. [Online]. Available: https://www.aclweb.org/anthology/D17-1143.

[162] Y. Ji and J. Eisenstein, "Representation learning for text-level discourse parsing," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, 2014, pp. 13–24. DOI: 10.3115/v1/p14-1002. [Online]. Available: https://doi.org/10.3115/v1/p14-1002.

[163] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2015, pp. 1529–1537. DOI: 10.1109/ICCV.2015.179. [Online]. Available: https://doi.org/10.1109/ICCV.2015.179.

[164] A. F. T. Martins and M. S. C. Almeida, "Priberam: A turbo semantic parser with second order features," in *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*, 2014, pp. 471–476. DOI: 10.3115/v1/s14-2082. [Online]. Available: https://doi.org/10.3115/v1/s14-2082.

[165] D. Goldwasser and X. Zhang, "Understanding satirical articles using common-sense," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 537–549, 2016. DOI: 10.1162/tacl_a_00116. [Online]. Available: https://www.aclweb.org/anthology/Q16-1038.

[166] H. Poon and P. Domingos, "Unsupervised semantic parsing," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore: Association for Computational Linguistics, Aug. 2009, pp. 1–10. [Online]. Available: https://www.aclweb.org/anthology/D09-1001.

[167] R. Samdani, K.-W. Chang, and D. Roth, "A discriminative latent variable model for online clustering," in *Proceedings of the 31st International Conference on Machine Learning*, E. P. Xing and T. Jebara, Eds., ser. Proceedings of Machine Learning Research, vol. 32, Bejing, China: PMLR, 22–24 Jun 2014, pp. 1–9. [Online]. Available: http://proceedings.mlr.press/v32/samdani14.html.

[168] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," English (US), 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015, Jan. 2015.

[169] C.-N. J. Yu and T. Joachims, "Learning structural svms with latent variables," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09, Montreal, Quebec, Canada: Association for Computing Machinery, 2009, pp. 1169–1176, ISBN: 9781605585161. DOI: 10.1145/1553374.1553523. [Online]. Available: https://doi.org/10.1145/1553374.1553523.

[170] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," in *Proceedings of the 26th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, 2017, pp. 1391–1399.

[171] J. Zhang, J. Chang, C. Danescu-Niculescu-Mizil, L. Dixon, Y. Hua, D. Taraborelli, and N. Thain, "Conversations gone awry: Detecting early signs of conversational failure," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 1350–1361.

[172] D. Jurafsky, R. Ranganath, and D. McFarland, "Extracting social meaning: Identifying interactional style in spoken conversation," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, 2009, pp. 638–646.

[173] S. Greiff, "From interactive to collaborative problem solving: Current issues in the programme for international student assessment," *Review of psychology*, vol. 19, no. 2, pp. 111–121, 2012.

[174] M. Flor, S.-Y. Yoon, J. Hao, L. Liu, and A. von Davier, "Automated classification of collaborative problem solving interactions in simulated science tasks," in *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 2016, pp. 31–41.

[175] J. Hao, L. Liu, A. von Davier, P. Kyllonen, and C. Kitchen, "Collaborative problem solving skills versus collaboration outcomes: Findings from statistical analysis and data mining.," in *Proceedings of the 9th International Conference on Educational Data Mining*, 2016.

[176] R. Barzilay and M. Elhadad, "Using lexical chains for text summarization," in *In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, 1997, pp. 10–17.

[177] F. Tagliabue, L. Galassi, and P. Mariani, "The "pandemic" of disinformation in covid-19," *SN comprehensive clinical medicine*, vol. 2, no. 9, pp. 1287–1289, 2020.

[178] I. Montagni, K. Ouazzani-Touhami, A. Mebarki, N. Texier, S. Schück, C. Tzourio, *et al.*, "Acceptance of a covid-19 vaccine is associated with ability to detect fake news and health literacy," *Journal of public health (Oxford, England)*, 2021.

[179] T. Hossain, R. L. Logan IV, A. Ugarte, Y. Matsubara, S. Young, and S. Singh, "COVIDLies: Detecting COVID-19 misinformation on social media," in *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online: Association for Computational Linguistics, Dec. 2020. DOI: 10.18653/v1/2020.nlpcovid19-2.11. [Online]. Available: https://aclanthology.org/2020.nlpcovid19-2.11.

[180] F. Alam, F. Dalvi, S. Shaar, N. Durrani, H. Mubarak, A. Nikolov, G. Da San Martino, A. Abdelali, H. Sajjad, K. Darwish, *et al.*, "Fighting the covid-19 infodemic in social media: A holistic perspective and a call to arms," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 15, 2021, pp. 913–922.

[181] M. Weinzierl, S. Hopfer, and S. M. Harabagiu, "Misinformation adoption or rejection in the era of covid-19," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 15, 2021, pp. 787–795.

[182] K. Glandt, S. Khanal, Y. Li, D. Caragea, and C. Caragea, "Stance detection in COVID-19 tweets," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 1596–1611. DOI: 10.18653/v1/2021.acl-long.127. [Online]. Available: https://aclanthology.org/2021.acl-long.127.

[183] D. Wawrzuta, M. Jaworski, J. Gotlib, and M. Panczyk, "What arguments against covid-19 vaccines run on facebook in poland: Content analysis of comments," *Vaccines*, vol. 9, no. 5, p. 481, 2021.

[184] J. Hoover, G. Portillo-Wightman, L. Yeh, S. Havaldar, A. Davani, Y. Lin, B. Kennedy, M. Atari, Z. Kamel, M. Mendlen, G. Moreno, C. Park, T. Chang, J. Chin, C. Leong, J. Leung, A. Mirinjian, and M. Dehghani, "Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment," *Social Psychological and Personality Science*, vol. 11, no. 8, pp. 1057–1071, 2020.

[185] L. Deng and J. Wiebe, "MPQA 3.0: An entity/event-level sentiment corpus," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado: Association for Computational Linguistics, May 2015, pp. 1323–1328. DOI: 10.3115/v1/N15-1146. [Online]. Available: https://aclanthology.org/N15-1146.

[186] G. Muric, Y. Wu, and E. Ferrara, “COVID-19 Vaccine Hesitancy on Social Media: Building a Public Twitter Dataset of Anti-vaccine Content, Vaccine Misinformation and Conspiracies,” May 2021. arXiv: 2105.05134. [Online]. Available: http://arxiv.org/abs/2105.05134.

[187] K. Krippendorff, “Measuring the reliability of qualitative text analysis data,” *Quality and quantity*, vol. 38, pp. 787–800, 2004.

[188] S. Stumpf, V. Rajaram, L. Li, W.-K. Wong, M. Burnett, T. Dietterich, E. Sullivan, and J. Herlocker, “Interacting meaningfully with machine learning systems: Three experiments,” *Int. J. Hum.-Comput. Stud.*, vol. 67, no. 8, pp. 639–662, Aug. 2009, ISSN: 1071-5819. DOI: 10.1016/j.ijhcs.2009.03.004. [Online]. Available: https://doi.org/10.1016/j.ijhcs.2009.03.004.

[189] S. Teso and K. Kersting, “Explanatory interactive machine learning,” in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES ’19, Honolulu, HI, USA: Association for Computing Machinery, 2019, pp. 239–245, ISBN: 9781450363242. DOI: 10.1145/3306618.3314293. [Online]. Available: https://doi.org/10.1145/3306618.3314293.

[190] M. Hanafi, A. Abouzied, L. Chiticariu, and Y. Li, “Synthesizing extraction rules from user examples with seer,” English (US), in *SIGMOD 2017 - Proceedings of the 2017 ACM International Conference on Management of Data*, ser. Proceedings of the ACM SIGMOD International Conference on Management of Data, Publisher Copyright: © 2017 ACM. Copyright: Copyright 2018 Elsevier B.V., All rights reserved.; 2017 ACM SIGMOD International Conference on Management of Data, SIGMOD 2017 ; Conference date: 14-05-2017 Through 19-05-2017, Association for Computing Machinery, May 2017, pp. 1687–1690. DOI: 10.1145/3035918.3056443.

[191] B. Settles and M. Craven, “An analysis of active learning strategies for sequence labeling tasks,” in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii: Association for Computational Linguistics, Oct. 2008, pp. 1070–1079. [Online]. Available: https://www.aclweb.org/anthology/D08-1112.

[192] T. Rogers and J. L. McClelland, “Semantic cognition: A parallel distributed processing approach,” 2004.

[193] D. Navon, “Forest before trees: The precedence of global features in visual perception,” *Cognitive Psychology*, vol. 9, no. 3, pp. 353–383, 1977, ISSN: 0010-0285. DOI: https://doi.org/10.1016/0010-0285(77)90012-3. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0010028577900123.

[194] R. H. Johnson, “Gilbert harman change in view: Principles of reasoning (cambridge, ma: Mit press 1986). pp. ix 147.,” *Canadian Journal of Philosophy*, vol. 18, no. 1, pp. 163–178, 1988. DOI: 10.1080/00455091.1988.10717172.

[195] P. Sowa, Ł. Kiszkiel, P. P. Laskowski, M. Alimowski, Ł. Szczerbiński, M. Paniczko, A. Moniuszko-Malinowska, and K. Kamiński, "Covid-19 vaccine hesitancy in poland—multifactorial impact trajectories," *Vaccines*, vol. 9, no. 8, p. 876, 2021.

[196] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," 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)*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. DOI: 10.18653/v1/D19-1410. [Online]. Available: https://aclanthology.org/D19-1410.

[197] X. Jin and J. Han, "K-means clustering," in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, 2010, pp. 563–564, ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_425. [Online]. Available: https://doi.org/10.1007/978-0-387-30164-8_425.

[198] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008. [Online]. Available: http://www.jmlr.org/papers/v9/vandermaaten08a.html.

[199] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, 1987, ISSN: 0377-0427. DOI: http://dx.doi.org/10.1016/0377-0427(87)90125-7. [Online]. Available: http://portal.acm.org/citation.cfm?id=38772.

[200] V. Braun and V. Clarke, "Thematic analysis.," in. Jan. 2012, pp. 57–71, ISBN: 978-1-4338-1003-9.

[201] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

VITA

Maria Leonor Pacheco was born in 1989 in Caracas, Venezuela. She received her Bachelor of Science in Computer Science and Engineering from the Universidad Simon Bolivar in 2013. Before joining Purdue, she worked as Software Engineer and Data Scientist for various startups in Caracas, Venezuela. She joined the Department of Computer Science at Purdue University in the fall of 2015 and began working with Dr. Dan Goldwasser. Her research has focused on neural-symbolic methods to model natural language discourse scenarios. Her work has been published in top Natural Language Processing conferences and journals, such as TACL, ACL, NAACL, EMNLP, EACL and SIGDIAL. She is a recipient of the 2021 Microsoft Research Dissertation Grant. During her graduate studies, Maria has served as Diversity Coordinator for the Computer Science Graduate Student Association, Computer Science Representative for the graduate chapter of the Purdue Women in Science Programs, and Global Ambassador for the Purdue Graduate School. In the broader research community, she has served as Social Media Co-Chair for SIGDIAL 2021, Student Chair for the NAACL 2022 Student Research Workshop, a main organizer of the LatinX in AI and Queer in AI events in NLP conferences, and a reviewer for various NLP and AI conferences. Upon graduation, she will spend one year as a Postdoctoral Researcher at Microsoft Research NYC. In the fall of 2023, she will join the Department of Computer Science at the University of Colorado Boulder as an Assistant Professor.

PUBLICATIONS

M. L. Pacheco, T. Islam, M. Mahajan, A. Shor, M. Yin, L. Ungar, and D. Goldwasser, "A Holistic Framework for Analyzing the COVID-19 Vaccine Debate", in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2022.

N. Mehta, M. L. Pacheco, and D. Goldwasser, "Tackling Fake News Detection by Continually Improving Social Context Representations using Graph Neural Networks", in *Proceedings of the 60th Anual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

M. L. Pacheco, M. von Hippel, B. Weintraub, D. Goldwasser, and C. Nita-Rotaru, "Automated Attack Synthesis by Extracting Finite State Machines from Protocol Specification Documents", in *43rd IEEE Symposium on Security and Privacy, (S&P Oakland)*, 2022.

S. Roy, M. L. Pacheco, and D. Goldwasser, "Identifying Morality Frames in Political Tweets using Relational Learning", in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

M. L. Pacheco, and D. Goldwasser, "Modeling Content and Context with Deep Relational Learning", in *Trasactions of the Association for Computational Linguistics (TACL)*, 2021.

I. T. Lee, M. L. Pacheco, and D. Goldwasser, "Modeling Human Mental States with an Entity-based Narrative Graph", in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2021.

M. Widmoser[1], M. L. Pacheco[1], and D. Goldwasser, “Randomized Deep Structured Prediction for Discourse-level Processing”, in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2021.

I. T. Lee, M. L. Pacheco, and D. Goldwasser, “Weakly-Supervised Modeling of Contextualized Event Embedding for Discourse Relations”, in *Findings of the Association for Computational Linguistics: EMNLP 2020.*

A. Jain, M. L. Pacheco, S. Lancette, M. Goindani, and D. Goldwasser, “Identifying Collaborative Conversations using Latent Discourse Behaviors”, in *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2020.

S. Jero, M. L. Pacheco, D. Goldwasser, and C. Nita-Rotaru, “Leveraging Textual Specifications for Grammar-Based Fuzzing of Network Protocols”, in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

I. T. Lee, M. Goindani, C. Li, D. Jin, K. Johnson, X. Zhang, M. L. Pacheco, D. Goldwasser, “PurdueNLP at SemEval-2017 Task 1: Predicting Semantic Textual Similarity with Paraphrase and Event Embeddings”, in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval)*, 2017.

X. Zhang[1], M. L. Pacheco[1], C, Li, and D. Goldwasser, “Introducing DRaiL: A Step Towards Declarative Deep Relational Learning”, in *Proceedings of the Workshop on Structured Prediction for NLP*, 2016.

M. L. Pacheco, I. T. Lee, X. Zhang, A. K. Zehady, P. Daga, D. Jin, A. Parolia, and D. Goldwasser, “Adapting Event Embedding for Implicit Discourse Relation Recognition”, in *Proceedings of the CoNLL-16 shared task*, 2016.

[1] ↑Equal contribution