Vision: Explainable Hidden Mental States as Influence Indicators

Brodie Mather,[1] **Ian Perera,**[1] **Vera Kazakova,**[1] **Daniel Capecci,**[2] **Muskan Garg,**[2] **Damon Woodard,**[2] **Bonnie J. Dorr**[2]

[1]Institute for Human and Machine Cognition, Pensacola, Florida; {bmather,iperera}@ihmc.org
[2]University of Florida, Gainesville, Florida; {dcapecci,muskangarg,dwoodard,bonniejdorr}@ufl.edu

Abstract

We posit that the next major thrust relevant to capturing dynamics for detecting and responding to information operations is *inference of hidden mental states through natural language processing and social computing techniques*. An important factor contributing to this vision is the need for *explainable representations*, e.g., propositions, to capture hidden mental states as indicators of influence campaigns. Hidden mental states under exploration include, for example, *belief*, *stance*, and *concern*. We view *explainability* not in terms of a "reason describer" for machine learning (ML) model output, but in terms of an inherently *interpretable* paradigm that leverages hidden mental states to produce both an explanation and a justification of output. The aim is to reap the benefits of both worlds: (1) breadth of coverage for features that are essential to the task at hand (e.g., embedding and attention models for extracting sentiment); (2) depth and transparency of representational formalisms for explaining system decisions (e.g., propositions that identify beliefs and attitudes).

Introduction

Influence campaigns are increasingly a global problem. Attacks, once of a physical nature, have shifted to a digital space. Identifying attacks effectively, efficiently, and in an **explainable** manner is crucial to thwarting them. We suggest a new paradigm for inferring hidden mental states in social interactions (social media) where information campaigns run rampant: hybridized natural language processing that combines symbolic representations with traditional ML approaches. If successful, such an endeavor would enable the development of an inherently *interpretable* paradigm that produces an explanation and justification of output. Examples below focus on the context of influence campaigns, but we expect our vision for an interpretable paradigm to generalize to a range of different areas, such as social engineering attacks and mental health diagnostics.

Vision

Our vision is illustrated in Figure 1.[3] Traditional ML approaches have historically lacked a reliable explanation for the output that they produce, as the inner workings of most ML are a black box. Recent approaches have attempted to take the output of ML algorithms and provide an explanation on top of it (see traditional XAI, upper portion of Figure 1). Although this is a step in the right direction it is inadequate and at times also inaccurate (Rudin 2019). Highly volatile situations such as detecting and countering influence operations demand a certain level of trust in the tools that are being used. Therefore, our vision for a reliable explanatory approach to inferring hidden mental states is a comprehensive interpretable model (see lower portion of Figure 1) that has explainability baked into the building of the model itself. These inherent characteristics of the paradigm itself leads to explainable output that boosts both the accuracy of the model as well as the user's trust in the model's output.

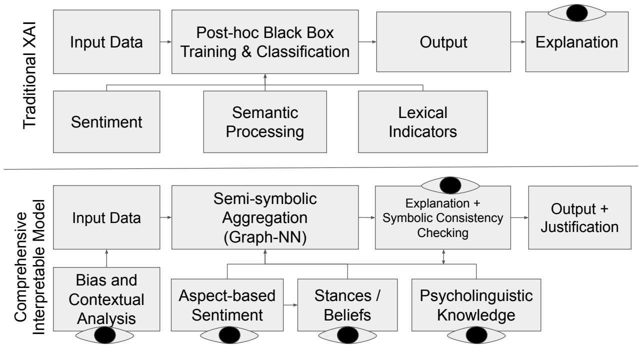
Traditional XAI
Input Data
Post-hoc Black Box Training & Classification
Output
Explanation
Sentiment
Semantic Processing
Lexical Indicators
Comprehensive Interpretable Model
Input Data
Semi-symbolic Aggregation (Graph-NN)
Explanation + Symbolic Consistency Checking
Output + Justification
Bias and Contextual Analysis
Aspect-based Sentiment
Stances / Beliefs
Psycholinguistic Knowledge

Figure 1: Traditional Explainable Artificial Intelligence (XAI) Approaches vs. Vision for the Future: Comprehensive Interpretable Model for Inferring Influence Indicators with potential implementation methods (e.g., Graph-NN)

How is it done today?

The most prevalent position among developers of *explainable AI models* for inferring mental states is one that yields general explanations from black-box output. This is achieved through use of *surface-level linguistic* and high-level *semantic* features for *post-hoc black box* training and classification to infer hidden states. Outputs from non-symbolic or sub-symbolic XAI approaches are not human-interpretable. However, promising new directions for ex-

[3] Mather&Dorr are supported by DARPA HR001121C0186; all opinions are those of the authors, not necessarily US goverment.

plainable AI have emerged in symbolic and sub-symbolic techniques (Calegari et al. 2020; Kazakova et al. 2019).

Existing black-box approaches focus on deriving the importance of the input feature set, returning a representation of the degree to which particular features play a role in the output prediction. These *sensitivities* or *saliency* measures are often just as brittle as the statistical associations they are trying to explain (Zhang et al. 2020; Ghorbani, Abid, and Zou 2019). Additionally, the task of mapping the *relationships* between model features to meaningful interpretations is daunting and ill-defined. Without these key elements, current XAI approaches cannot elucidate the way a model *reasons* about any particular decision. Fragile black-box approximations give a false sense of understanding and further muddy the waters of interpretation Even in the case where stake-holders have 'white box' access to internal model state at any given time, (e.g., in analyzing neural activation spaces), the problem has only shifted to a different high-dimensional hard-to-interpret space.

Consider the problem of influence indicator detection in the domain of French elections (Daignan 2017). In a tweet such as *Marine Le Pen LEADS in French poll as far left Jean-Luc Melenchon could ruin economy*, standard sentiment analysis uses ML algorithms such as Gardner et al. (2018) which—even with a more advanced 2021 RoBERTa model (Liu et al. 2019)—produces the following output: *somewhat confident negative*.

A traditional XAI approach may take lexical information into account to assign a sentiment score to each word (or ngram), using the words that contributed most heavily to the negative sentiment output to "explain" the output. However, because the model is not inherently interpretable, there is a high risk that the contributing weights of lexical items (or ngrams) in isolation do not match the way the output is computed in the first place (i.e., the *reason* the output is *negative* is unknown). The two processes are entirely independent. Moreover, trained models are prone to a lack of accuracy with respect to understanding deeper notions underlying the input sentence, e.g., that the target of negativity is *Jean-Luc Melenchon* (not, for example, the word *ruin* itself).

More recently, an explainable representation based on propositional analysis shows generalizability across domains, tasks, and languages (Mather et al. 2022). Three types of hidden mental states are extracted from Twitter data: *beliefs*, *stances*, and *concerns* (first steps toward intention). Each uses a propositional representation that lends itself to explainability: *ruin(jean-luc melenchon, economy)*. However, this work stops short of what an explanation would look like or what paradigm would support the generation of such an explanation. Moreover, this representation is built independently from ML models, thus missing an opportunity to benefit from broad feature coverage of ML models. The next step, then is to examine how one might benefit from the best of both worlds: (1) breadth of featural coverage (ML) and (2) depth/transparency of representational formalisms for explainability.

How would it be done tomorrow?

To go beyond traditional explainable artificial intelligence (XAI) approaches (Arrieta et al. 2019), we move away from mere explanations of black-box decisions (through exemplars, feature-relevance values, confidence values) and towards independently transparent and interpretable components that can be combined and verified to provide a justification alongside each decision. This entails constrained decisions that may be scrutinized throughout every step of system operation, from selection of data (and its inherent biases) to application of the system to novel situations (and corresonding verification of reliability). Approaches incorporating interpretable rules combined with the flexibility and generalizability of statistical approaches allow human knowledge and insight to be incorporated throughout the system (Perera et al. 2018), rather than just as a means to accept or discard a system's output.

This transparency further enhances the application of such systems to problems without a final ground truth answer. For example, returning to the information operations context: a shift in ideologies in an online group of people may never have a specific instance that signals the shift and there may not be a way to feed the system the "final" state of such a change; however assessing the magnitude and dimensions of that shift could be a vital capability of future systems. Additionally, a wider range of potential applications are possible without the need for ground truth when the independent components of the system are trustable and verifiable.

In contrast to traditional XAI systems, a comprehensive interpretive model ties together explanatory representations and components of a full system that aggregates knowledge, checks for symbolic consistency, identifies biases, and ultimately produces an output and its justification (see Figure 1). Returning to our French election example, instead of a single output (*negative*), the system extracts and outputs a proposition *ruin(jean-luc melenchon, economy)*, along with a belief type (*DESTROY*) and a belief-driven attitude (*negative*). The proposition and associated sentential elements are fed into a semi-symbolic aggregation process from which an explanation is produced: *The author's attitude toward Jean-Luc is negative due to the belief that Jean-Luc is ruining the economy.* Such mental states might predict further propositions by the same agent or those with similar beliefs.

Recommendations to Achieve Vision

We recommend the following steps to implement our vision for deriving hidden mental states as influence indicators:

1. Exploration of transparent representations (e.g., propositions) that support understanding of mental-state inference and enable the construction of trustworthy outputs with justifications.
2. Definition/implementation of novel ML paradigms that employ such representations during model building to facilitate detection and countering of influence operations.
3. Periodic human-in-the-loop feedback for continuous enrichments to the interpretable paradigm, as captured in the two-way arrow in Figure 1 (between symbolic consistency checking and input knowledge).