Discovering Implicit Social Norms in Media Corpora via Norm Manifolds, Evaluative Stance, and Semantic Entropy

Abstract

This document specifies a step by step method for discovering implicit social norms in large media or social media corpora by combining three ingredients: (i) a corpus of explicit normative statements that defines a *norm manifold*, (ii) evaluative stance representations predicted by large language models, and (iii) semantic entropy estimates obtained from both representation geometry and masked language model probing. The method formalizes distance from explicit norms in a joint semantic and evaluative space and uses entropy as a confidence and contestation signal to identify implicit norm instances in unstructured text.

1 Problem setting

We assume access to:

- A corpus of *explicit normative text* containing statements such as codes of conduct, policies, etiquette descriptions and moral rules.
- A large *target corpus* of media or social media text such as news articles or posts, where explicit norms are rare and most normative content is implicit.
- A large language model or a set of models that provide:
  - sentence or span embeddings,
  - evaluative stance predictions,
  - token level probabilities for masked language modeling.

The goal is to construct a procedure that maps each sentence or event in the target corpus to:

1. a proximity score to one or more explicit norm types, and
2. a confidence and contestation measure based on semantic entropy,

and to use these to identify sentences that realize or presuppose social norms, even when the norms are never explicitly stated.

1.1 Notation

- Let $\mathcal{C}^{(E)} = \{n_j\}_{j=1}^{N_E}$ denote the explicit norm corpus.
- Let $\mathcal{C}^{(T)} = \{x_i\}_{i=1}^{N_T}$ denote the target media corpus.

- Let $f_e : \text{Text} \to \mathbb{R}^d$ be a sentence embedding function.
- Let $f_s : \text{Text} \to [0,1]^K$ be a stance classifier that outputs a probability vector over K evaluative stance dimensions.
- Let $f_p : \text{MaskedText} \times \mathcal{V} \to [0,1]$ be a masked language model that maps a masked input and a vocabulary $\mathcal{V}$ to token probabilities.

For a text span z we write:

$$\boldsymbol{e}(z) = f_e(z) \in \mathbb{R}^d, \tag{1}$$
$$\boldsymbol{s}(z) = f_s(z) \in [0,1]^K. \tag{2}$$

2 Representation of sentences and norms

2.1 Evaluative stance space

Let the stance dimensions be indexed by $k \in \{1, \ldots, K\}$ and interpreted, for example, as:

$$\begin{aligned} k &= 1 : \text{approval}, \\ k &= 2 : \text{blame}, \\ k &= 3 : \text{outrage}, \\ k &= 4 : \text{praise}, \\ k &= 5 : \text{perceived obligation}, \\ k &= 6 : \text{perceived prohibition}, \\ &\ldots \end{aligned}$$

For any span z, the stance vector $\boldsymbol{s}(z) = (s_1(z), \ldots, s_K(z))$ is a probability distribution:

$$\forall k \; s_k(z) \geq 0, \quad \sum_{k=1}^{K} s_k(z) = 1. \tag{3}$$

In practice f_s is implemented by fine tuning a model on labeled evaluative data with a cross entropy objective, but the method does not depend on the exact training procedure.

2.2 Semantic representation and joint feature space

For each span z, define a basic semantic representation:

$$\boldsymbol{e}(z) = f_e(z). \tag{4}$$

We will embed both explicit norms n_j and target sentences x_i into a joint representation:

$$\phi(z) = [\boldsymbol{e}(z); \boldsymbol{s}(z); H_{\text{repr}}(z); H_{\text{mask}}(z)] \in \mathbb{R}^{d+K+2}, \tag{5}$$

where H_{repr} and H_{mask} are two semantic entropy measures defined below.

3 Semantic entropy

We introduce two complementary notions of semantic entropy.

3.1 Representation neighborhood entropy

For each span z, consider its neighborhood in embedding space. Let $\mathcal{N}_M(z)$ be the set of M nearest neighbors of $\boldsymbol{e}(z)$ among the representations of a reference corpus $\mathcal{R}$, which may be $\mathcal{C}^{(E)} \cup \mathcal{C}^{(T)}$.

Let the similarity between $\boldsymbol{e}(z)$ and a neighbor $\boldsymbol{e}(z')$ be denoted by:

$$\mathrm{sim}(\boldsymbol{e}(z), \boldsymbol{e}(z')) = \frac{\boldsymbol{e}(z) \cdot \boldsymbol{e}(z')}{\|\boldsymbol{e}(z)\| \|\boldsymbol{e}(z')\|}. \tag{6}$$

Define:

$$\tilde{p}_m(z) = \exp\left(\tau \cdot \mathrm{sim}(\boldsymbol{e}(z), \boldsymbol{e}(z_m))\right), \tag{7}$$

where $\tau > 0$ is a temperature parameter and $z_m \in \mathcal{N}_M(z)$ is the mth neighbor.

Normalize to obtain a discrete distribution over neighbors:

$$p_m(z) = \frac{\tilde{p}_m(z)}{\sum_{\ell=1}^{M} \tilde{p}_\ell(z)}. \tag{8}$$

Then the representation neighborhood entropy is:

$$H_{\mathrm{repr}}(z) = -\sum_{m=1}^{M} p_m(z) \log p_m(z). \tag{9}$$

Low $H_{\mathrm{repr}}(z)$ indicates that z resides in a dense and coherent semantic cluster. High $H_{\mathrm{repr}}(z)$ indicates a more diffuse or ambiguous neighborhood.

3.2 Masked expectation entropy

For masked probes we define a set of masking templates $\mathcal{T}$ and associated masking operations. For a given span z and template $t \in \mathcal{T}$ we produce a masked input $z^{(t)}$ that hides a behavior, reaction or evaluation token.

Let $\mathcal{V}_t \subseteq \mathcal{V}$ be the subset of vocabulary items regarded as meaningful completions for template t (for example, evaluative adjectives for an evaluative mask).

The model f_p defines for each t a distribution:

$$p_t(w \mid z^{(t)}) = f_p(z^{(t)}, w), \quad w \in \mathcal{V}_t. \tag{10}$$

Normalize to the restricted vocabulary:

$$q_t(w \mid z^{(t)}) = \frac{p_t(w \mid z^{(t)})}{\sum_{u \in \mathcal{V}_t} p_t(u \mid z^{(t)})}. \tag{11}$$

Define the entropy for template t:

$$H_t(z) = -\sum_{w \in \mathcal{V}_t} q_t(w \mid z^{(t)}) \log q_t(w \mid z^{(t)}). \tag{12}$$

Aggregate over templates, for example by averaging:

$$H_{\text{mask}}(z) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} H_t(z). \tag{13}$$

Low $H_{\text{mask}}(z)$ indicates that the model has a sharp expectation about reactions or evaluations implied by z, which is an important signal for norm stability.

4 Building the explicit norm manifold

4.1 Embedding explicit norms

For each explicit norm $n_j \in \mathcal{C}^{(E)}$ compute:

$$\boldsymbol{e}_j^{(E)} = \boldsymbol{e}(n_j) = f_e(n_j), \tag{14}$$

$$\boldsymbol{s}_j^{(E)} = \boldsymbol{s}(n_j) = f_s(n_j), \tag{15}$$

$$H_{\text{repr}}^{(E)}(n_j) = H_{\text{repr}}(n_j), \tag{16}$$

$$H_{\text{mask}}^{(E)}(n_j) = H_{\text{mask}}(n_j). \tag{17}$$

Collect into joint vectors:

$$\phi^{(E)}(n_j) = \left[\boldsymbol{e}_j^{(E)}; \boldsymbol{s}_j^{(E)}; H_{\text{repr}}^{(E)}(n_j); H_{\text{mask}}^{(E)}(n_j) \right]. \tag{18}$$

4.2 Clustering explicit norms into norm types

Apply a clustering algorithm (for example, Gaussian mixture models or k means) to the set $\{\phi^{(E)}(n_j)\}_{j=1}^{N_E}$ to obtain K_N clusters, which we treat as norm types.

Let the cluster assignment for n_j be $c(j) \in \{1, \ldots, K_N\}$. Let cluster k contain the index set:

$$\mathcal{I}_k = \{j \mid c(j) = k\}. \tag{19}$$

Define the centroid of cluster k in each subspace:

$$\boldsymbol{\mu}_k^{(e)} = \frac{1}{|\mathcal{I}_k|} \sum_{j \in \mathcal{I}_k} \boldsymbol{e}_j^{(E)}, \tag{20}$$

$$\boldsymbol{\mu}_k^{(s)} = \frac{1}{|\mathcal{I}_k|} \sum_{j \in \mathcal{I}_k} \boldsymbol{s}_j^{(E)}, \tag{21}$$

$$\mu_k^{(\text{repr})} = \frac{1}{|\mathcal{I}_k|} \sum_{j \in \mathcal{I}_k} H_{\text{repr}}^{(E)}(n_j), \tag{22}$$

$$\mu_k^{(\text{mask})} = \frac{1}{|\mathcal{I}_k|} \sum_{j \in \mathcal{I}_k} H_{\text{mask}}^{(E)}(n_j). \tag{23}$$

The explicit norm manifold is then approximated by the set of centroids:

$$\mathcal{M}_{\text{norm}} = \left\{ \left[\boldsymbol{\mu}_k^{(e)}; \boldsymbol{\mu}_k^{(s)}; \mu_k^{(\text{repr})}; \mu_k^{(\text{mask})} \right] \mid k = 1, \ldots, K_N \right\}. \tag{24}$$

Each k corresponds to a norm type such as honesty, fairness or respect for authority.

5 Processing the target media corpus

For each sentence or event $x_i \in \mathcal{C}^{(T)}$ we compute:

$$\boldsymbol{e}_i^{(T)} = \boldsymbol{e}(x_i), \tag{25}$$

$$\boldsymbol{s}_i^{(T)} = \boldsymbol{s}(x_i), \tag{26}$$

$$H_{\text{repr}}^{(T)}(x_i) = H_{\text{repr}}(x_i), \tag{27}$$

$$H_{\text{mask}}^{(T)}(x_i) = H_{\text{mask}}(x_i), \tag{28}$$

and joint representation:

$$\phi^{(T)}(x_i) = \left[\boldsymbol{e}_i^{(T)}; \boldsymbol{s}_i^{(T)}; H_{\text{repr}}^{(T)}(x_i); H_{\text{mask}}^{(T)}(x_i)\right]. \tag{29}$$

6 Distances to norm types

6.1 Subspace distances

We define distances between an arbitrary target span x and a norm type k in each subspace.

Semantic distance. For embedding vectors we use cosine distance:

$$d_e(x,k) = 1 - \frac{\boldsymbol{e}(x) \cdot \boldsymbol{\mu}_k^{(e)}}{\|\boldsymbol{e}(x)\|\|\boldsymbol{\mu}_k^{(e)}\|}. \tag{30}$$

Stance distance. Treat stance vectors as distributions and use for example squared Euclidean distance:

$$d_s(x,k) = \|\boldsymbol{s}(x) - \boldsymbol{\mu}_k^{(s)}\|_2^2, \tag{31}$$

or Kullback Leibler divergence if desired:

$$d_s^{\text{KL}}(x,k) = \sum_{j=1}^{K} s_j(x) \log \frac{s_j(x)}{\mu_{k,j}^{(s)}}. \tag{32}$$

Entropy distance. We treat entropy as scalar features. For each entropy type:

$$d_{\text{repr}}(x,k) = \left|H_{\text{repr}}(x) - \mu_k^{(\text{repr})}\right|, \tag{33}$$

$$d_{\text{mask}}(x,k) = \left|H_{\text{mask}}(x) - \mu_k^{(\text{mask})}\right|. \tag{34}$$

6.2 Composite norm distance

Define a composite distance:

$$D_k(x) = \alpha\, d_e(x,k) + \beta\, d_s(x,k) + \gamma\, d_{\text{repr}}(x,k) + \delta\, d_{\text{mask}}(x,k), \tag{35}$$

where $\alpha, \beta, \gamma, \delta \geq 0$ are hyperparameters.

The smaller $D_k(x)$, the closer x is to norm type k in joint semantic, evaluative and entropy space.

We can convert distances to unnormalized scores:

$$R_k(x) = \exp\left(-D_k(x)\right), \tag{36}$$

and define a distribution over norm types:

$$\pi_k(x) = \frac{R_k(x)}{\sum_{\ell=1}^{K_N} R_\ell(x)}. \tag{37}$$

The index of the closest norm type is then:

$$k^{(x)=\arg\max_k \pi_k(x).} \tag{38}$$

7 Implicit norm identification criteria

A target span x is considered a candidate implicit realization of norm type k if:

$$k = k^{(x),} \tag{39}$$

$$D_k(x) \leq \tau_D, \tag{40}$$

$$H_{\text{mask}}(x) \leq \tau_H, \tag{41}$$

where τ_D and τ_H are thresholds chosen by validation or inspection.

The first condition selects the nearest norm type. The second condition requires that the distance is small enough to be meaningful. The third condition uses masked semantic entropy to enforce that the model has a sharp expectation about reactions or evaluations associated with x, which suggests a stable norm rather than a purely idiosyncratic event.

8 Masked probing for norm consistency

In addition to $H_{\text{mask}}(x)$ used as a scalar, masked probes can be used to test consistency with a given norm type.

Suppose each norm type k is associated with a set of prototypical completions $\mathcal{W}_k$ in specific templates. For instance, a fairness norm type might be associated with completions such as "unfair", "unjust", "unequal" in an evaluative mask.

For a template t and masked input $x^{(t)}$, we compute:

$$q_t(w \mid x^{(t)}) = \frac{p_t(w \mid x^{(t)})}{\sum_{u \in \mathcal{V}_t} p_t(u \mid x^{(t)})}. \tag{42}$$

Define a norm consistency score for type k:

$$C_k(x) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \sum_{w \in \mathcal{W}_k} q_t(w \mid x^{(t)}). \tag{43}$$

High $C_k(x)$ indicates that the masked completions for x resemble those expected for explicit instances of norm type k.

This can be integrated into the distance function via an additional term or used as a post hoc filter:

$$C_k(x) \geq \tau_C. \tag{44}$$

9 Algorithm

9.1 Overall pipeline

Algorithm 1 summarizes the entire method.

Algorithm 1 Implicit norm discovery via norm manifold and semantic entropy

Require: Explicit norm corpus $\mathcal{C}^{(E)}$, target corpus $\mathcal{C}^{(T)}$, models f_e, f_s, f_p, hyperparameters $\alpha, \beta, \gamma, \delta, \tau_D, \tau_H, \tau_C$, number of norm clusters K_N.

Embed explicit norms
for each $n_j \in \mathcal{C}^{(E)}$ **do**
  Compute $\boldsymbol{e}_j^{(E)} = f_e(n_j)$
  Compute $\boldsymbol{s}_j^{(E)} = f_s(n_j)$
  Compute $H_{\text{repr}}^{(E)}(n_j)$ from neighborhood in $\mathcal{C}^{(E)}$
  Compute $H_{\text{mask}}^{(E)}(n_j)$ from masked probes
end for
Form joint vectors $\phi^{(E)}(n_j)$
Cluster explicit norms
Cluster $\{\phi^{(E)}(n_j)\}$ into K_N clusters to obtain $\mathcal{I}_k$
Compute centroids $\boldsymbol{\mu}_k^{(e)}, \boldsymbol{\mu}_k^{(s)}, \mu_k^{(\text{repr})}, \mu_k^{(\text{mask})}$
Process target corpus
for each $x_i \in \mathcal{C}^{(T)}$ **do**
  Compute $\boldsymbol{e}_i^{(T)} = f_e(x_i)$
  Compute $\boldsymbol{s}_i^{(T)} = f_s(x_i)$
  Compute $H_{\text{repr}}^{(T)}(x_i)$ from neighborhood in $\mathcal{C}^{(T)}$
  Compute $H_{\text{mask}}^{(T)}(x_i)$ from masked probes
  for each norm type $k = 1, \ldots, K_N$ **do**
    Compute $d_e(x_i, k)$, $d_s(x_i, k)$, $d_{\text{repr}}(x_i, k)$, $d_{\text{mask}}(x_i, k)$
    Compute $D_k(x_i) = \alpha d_e + \beta d_s + \gamma d_{\text{repr}} + \delta d_{\text{mask}}$
    Compute $R_k(x_i) = \exp(-D_k(x_i))$
  end for
  Normalize $R_k(x_i)$ over k to get $\pi_k(x_i)$
  Let $k^*(x_i) = \arg\max_k \pi_k(x_i)$
  Optionally compute $C_{k^{(x_i)}(x_i)}$ via masked consistency
  if $D_{k^*(x_i)}(x_i) \leq \tau_D$ and $H_{\text{mask}}^{(T)}(x_i) \leq \tau_H$ and $C_{k^*(x_i)}(x_i) \geq \tau_C$ **then**
    Mark x_i as a candidate implicit instance of norm type $k^{(x_i)}$
  end if
end for
Output set of candidate implicit norm instances with associated norm type and scores

10 Interpretation and analysis

The procedure described above yields, for each sentence or event in the target media corpus, a soft association with a set of explicit norm types, together with semantic entropy based confidence measures.

- Low $D_k(x)$ and low $H_{\text{mask}}(x)$ identify clear implicit applications of well established norms.
- Low $D_k(x)$ and high $H_{\text{mask}}(x)$ identify contested or ambiguous applications that may highlight norm conflict or emerging norms.
- High $D_k(x)$ for all k suggests the sentence is not strongly norm governed in the sense defined by the explicit corpus.

The method can be extended to track diachronic change in norms by applying the pipeline to time slices of the target corpus and analyzing changes in the distribution of implicit instances per norm type.

11 Conclusion

This document formalizes a method that combines explicit normative texts, evaluative stance modeling and semantic entropy to discover implicit social norms in large unstructured corpora such as news and social media. The key technical idea is to construct a norm manifold from explicit norms in a joint semantic and evaluative space and to measure distances from this manifold for each sentence in the target corpus, while using semantic entropy, especially from masked probing, as a measure of norm stability and contestation. The resulting pipeline provides a principled way to map the implicit normative landscape of real world discourse.