Public Wisdom Matters! Discourse-Aware Hyperbolic Fourier Co-Attention for Social-Text Classification

Karish Grover[1] Phaneendra Angara[2] Md. Shad Akhtar[1] Tanmoy Chakraborty[3]

[1]IIIT Delhi, India [2]LinkedIn, India [3]IIT Delhi, India

Github

INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

Paper

Overview

Overview

Motivation

- Social media has become the fulcrum of all forms of communication. Oftentimes, the **public wisdom** expressed through the social-media discourse acts as a surrogate of crowd-sourced view and provides us with complementary signals.

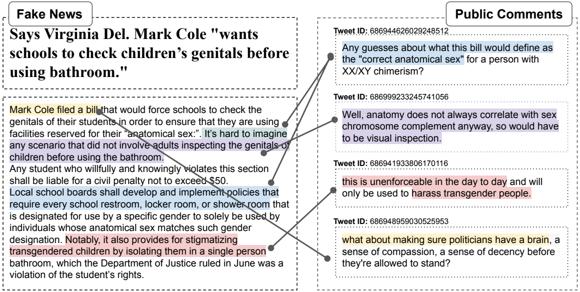
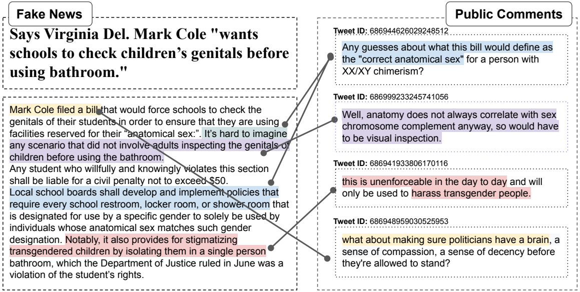


Figure 1: User comments act as **analogous evidence** for a fake news article.

Motivation



Problem Statement

- Classifying social texts such as fake news, rumour, sarcasm, etc. has gained significant attention. **Public discourse** can be used in unison with the source posts to enhance these tasks.
- We hypothesise that such public discourse carries **complementary and rich latent signals** (public wisdom, worldly knowledge, fact busting, opinions, emotions, etc.), which would otherwise be difficult to obtain from just standalone source-post analysis.

Overview

Public-discourse Encoder

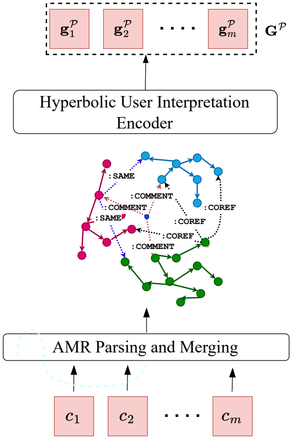
- We parse the user comments into an AMR (**Abstract Meaning Representation**) graph.
- The individual comment-level AMR graphs are merged to form a macro-AMR, representing the global public wisdom and latent *frequencies* in the discourse.

Why abstract meaning representation?

AMR abstracts away the **syntactic representation** i.e. sentences with similar semantics will be assigned similarly structured AMRs - helpful to capture the **structural context**.

What are frequencies here?

Frequencies *isomorphically* represent **rich latent signals** (public wisdom, worldly knowledge, fact busting, opinions, emotions, etc.) present in the public discourse.

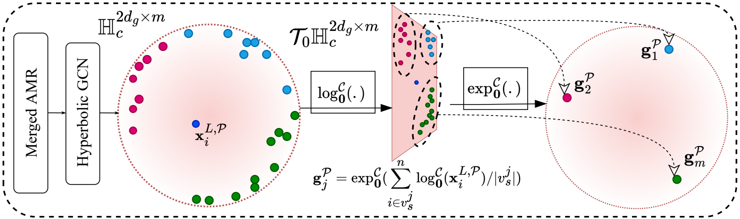


Public-discourse Encoder

- We learn representations of the macro-AMR using a **HGCN** (Hyperbolic Graph Conv. Network).
- Our graph encoder stacks multiple hyperbolic GCN layers to perform message passing.
- The message passing in a **HGCN** layer can be shown as:

$$\mathbf{x}_i^{\ell,\mathcal{H}} = \sigma^{\otimes_{K_{\ell-1},K_\ell}}\left(\mathrm{AGG}^{K_{\ell-1}}((W^\ell \otimes_{K_{\ell-1}} \mathbf{x}_i^{\ell-1,\mathcal{H}}) \oplus_{K_{\ell-1}} \mathbf{b}^\ell)\right)$$

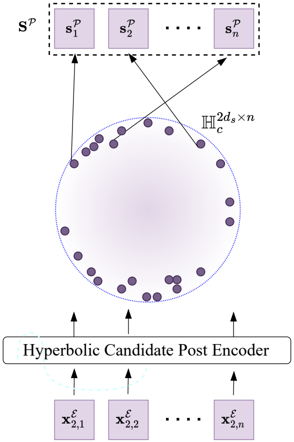
- We take the mean of the node embeddings for nodes present in a subgraph to yield the aggregated subgraph embedding.



Hyperbolic Candidate Post Encoder

- We learn the source post content representations through a hierarchical attention network in the hyperbolic space.
- **Word-level** and **sentence-level** attention weights help in determining the veracity of the news sentences, which are later used for explainability.
- The word-level attention weights are then computed as α_{it} by:

$$\alpha_{it} = \exp(-\beta_w d_{\mathcal{L}}(\mathbf{c}_w^{\mathcal{L}}, \mathbf{h}_{it}^{\mathcal{L}'}) - c_w)$$

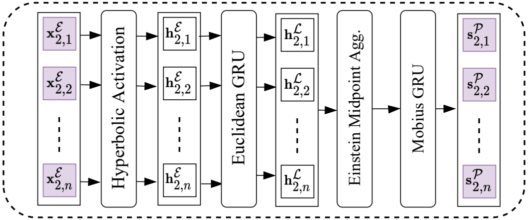


Hyperbolic Candidate Post Encoder

- After capturing the hyperbolic attention weights, the semantic meaning of words appearing in the same sentences is aggregated via Einstein midpoint:

$$\mathbf{s}_i^{\mathcal{K}w} = \sum_t \left[\frac{\alpha_{it}\gamma\left(\mathbf{h}_{it}^{\mathcal{K}}\right)}{\sum_l \alpha_{il}\gamma\left(\mathbf{h}_{it}^{\mathcal{K}}\right)} \right]; \quad \gamma\left(\mathbf{h}_{it}^{\mathcal{K}}\right) = \frac{1}{\sqrt{1 - ||\mathbf{h}_{it}^{\mathcal{K}}||^2}} = \frac{1}{\sqrt{1 - \frac{\sinh^2(r_{it})}{\cosh^2(r_{it})}}}$$

- Similar to the word-level encoder, *Mobiüs*-GRU units are utilized with aggregation using **Einstein midpoint** to encode each sentence in the source post, yielding the final document level representation.

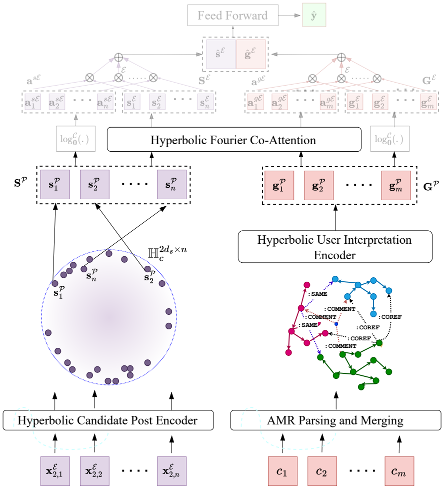


Bringing it all together!

- Next, we develop a hyperbolic Fourier co-attention mechanism to model the mutual influence between the source social media post (i.e., $\mathbf{S}^{\mathcal{P}}$) and user comments (interpretation) embeddings (i.e., $\mathbf{G}^{\mathcal{P}}$), to learn pairwise attention scores.

Why Hyperbolic Geometry?

The resultant Macro-AMR is highly **hierarchical**. We wish to utilize powerful **hierarchical representation abilities** of hyperbolic geometry.



Hyperbolic Fourier co-attention

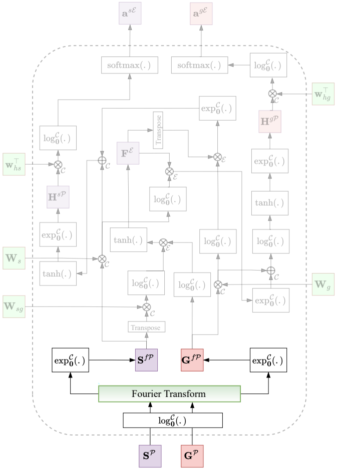
- The Fourier sublayer applies a 2D DFT to its (sequence length, hidden dimension) embedding input – one 1D DFT along the sequence dimension, $\mathcal{F}_{\text{seq}}$, and one 1D DFT along the hidden dimension, $\mathcal{F}_{\text{h}}$.

$$\mathbf{S}^{f\mathcal{P}} = \exp_{\mathbf{0}}^{c}\left(\mathcal{F}_{\text{seq}}\left(\mathcal{F}_{\text{h}}\left(\log_{\mathbf{0}}^{c}(\mathbf{S}^{\mathcal{P}})\right)\right)\right)$$

$$\mathbf{G}^{f\mathcal{P}} = \exp_{\mathbf{0}}^{c}\left(\mathcal{F}_{\text{seq}}\left(\mathcal{F}_{\text{h}}\left(\log_{\mathbf{0}}^{c}(\mathbf{G}^{\mathcal{P}})\right)\right)\right)$$

Why Fourier transform?

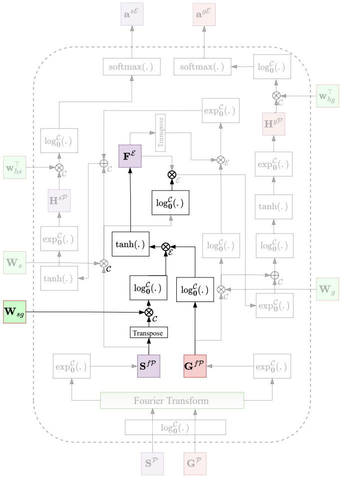
Finding the dominant *frequencies* is within a signal, i.e. the most common user interpretations and opinions from discourse.



Hyperbolic Fourier co-attention

- The affinity (proximity) matrix $\mathbf{F}^{\mathcal{E}}$ can be thought to transform the user-interpretation attention space to the candidate post attention space, and vice versa for its transpose $\mathbf{F}^{\mathcal{E}\top}$.

$$\mathbf{F}^{\mathcal{E}} = \tanh\left(\log_{\mathbf{0}}^{c}\left(\mathbf{S}^{f\mathcal{P}\top} \otimes_{c} \mathbf{W}_{sg}\right) \otimes_{\mathcal{E}} \log_{\mathbf{0}}^{c}\left(\mathbf{G}^{f\mathcal{P}}\right)\right)$$

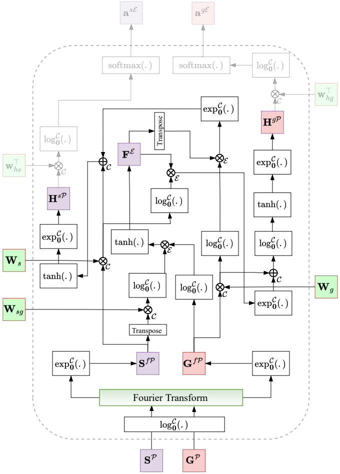


Hyperbolic Fourier co-attention

- By treating the affinity matrix as a feature, we can learn to predict candidate post and user interpretation attention maps $\mathbf{H}^{s\mathcal{P}} \in \mathbb{H}_c^{k\times n}$ and $\mathbf{H}^{g\mathcal{P}} \in \mathbb{H}_c^{k\times m}$.

$$\mathbf{H}^{s\mathcal{P}} = \exp_{\mathbf{0}}^{c}(\tanh(\log_{\mathbf{0}}^{c}(\mathbf{W}_s \otimes_c \mathbf{S}^{f\mathcal{P}} \oplus_c \exp_{\mathbf{0}}^{c}(\log_{\mathbf{0}}^{c}(\mathbf{W}_g \otimes_c \mathbf{G}^{f\mathcal{P}})) \otimes_{\mathcal{E}} \mathbf{F}^{\mathcal{E}\top}))))$$

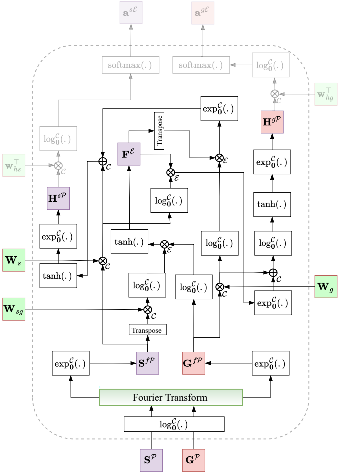
$$\mathbf{H}^{g\mathcal{P}} = \exp_{\mathbf{0}}^{c}(\tanh(\log_{\mathbf{0}}^{c}(\mathbf{W}_g \otimes_c \mathbf{G}^{f\mathcal{P}} \oplus_c \exp_{\mathbf{0}}^{c}(\log_{\mathbf{0}}^{c}(\mathbf{W}_s \otimes_c \mathbf{S}^{f\mathcal{P}})) \otimes_{\mathcal{E}} \mathbf{F}^{\mathcal{E}}))))$$



Hyperbolic Fourier co-attention

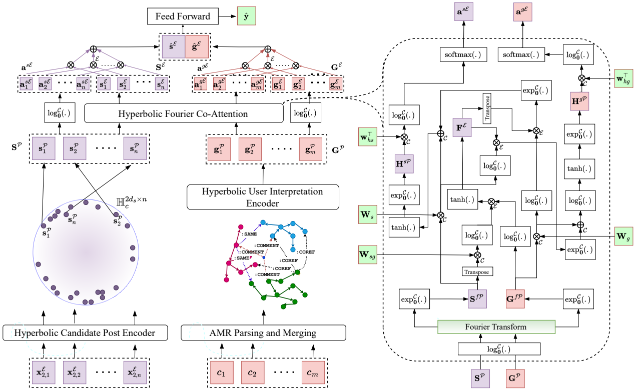
- We can then generate the attention weights of source words and interaction users through the Softmax function.
- $\mathbf{a}^{s\mathcal{E}} \in \mathbb{R}^{1\times m}$ and $\mathbf{a}^{g\mathcal{E}} \in \mathbb{R}^{1\times n}$ are the vectors of attention probabilities for each sentence in the source story and each user comment.

$$\mathbf{a}^{s\mathcal{E}} = \mathrm{softmax}(\log_{\mathbf{0}}^{c}(\mathbf{w}_{hs}^{\top} \otimes_c \mathbf{H}^{s\mathcal{P}}))$$
$$\mathbf{a}^{g\mathcal{E}} = \mathrm{softmax}(\log_{\mathbf{0}}^{c}(\mathbf{w}_{hg}^{\top} \otimes_c \mathbf{H}^{g\mathcal{P}}))$$



Hyphen

- Finally, we generate the attention vectors through weighted sum of the derived attention weights given by $\hat{\mathbf{s}}^{\mathcal{E}} = \sum_{i=1}^{n} \mathbf{a}_i^{s\mathcal{E}} \mathbf{s}_i^{\mathcal{E}}$ and $\hat{\mathbf{g}}^{\mathcal{E}} = \sum_{j=1}^{m} \mathbf{a}_j^{g\mathcal{E}} \mathbf{g}_j^{\mathcal{E}}$.



Overview

Experimentation

Task	Dataset	Data-specific Baselines			General Neural Baseline				Hyphen	
		Baseline1	Baseline2	Baseline3	HAN	dEFEND	BERT	RoBERTa	Euclidean	Hyperbolic
Fake News Detection	Politifact	0.827	0.881	0.871	0.902	0.928	0.905	0.906	0.941	0.967
	Gossipcop	0.603	0.673	0.682	0.778	0.755	0.762	0.772	0.781	0.816
	ANTiVax	0.872	0.865	0.908	0.833	0.935	0.942	0.939	0.937	0.945
Hate Speech Detection	HASOC	0.591	0.634	0.698	0.614	0.657	0.641	0.648	0.702	0.713
Rumour Detection	Pheme	0.782	0.801	0.844	0.799	0.841	0.861	0.852	0.844	0.875
	Twitter15	0.811	0.825	0.941	0.881	0.848	0.891	0.908	0.936	0.967
	Twitter16	0.778	0.759	0.919	0.889	0.887	0.919	0.892	0.937	0.938
	RumourEval	0.598	0.667	0.695	0.518	0.573	0.533	0.595	0.697	0.712
Sarcasm Detection	FigLang Twitter	0.681	0.741	0.752	0.721	0.757	0.797	0.801	0.812	0.822
	FigLang Reddit	0.585	0.667	0.678	0.665	0.631	0.677	0.689	0.698	0.701

r

Hyphen-hyperbolic outperforms all the baselines!

The performance improvement over baselines is significant (4%; $p < 0.005$) on datasets like Politifact, Gossipcop, and Twitter15. By incorporating public comments along with the social post, Hyphen shows significant performance improvement over all the baselines.

Experimentation

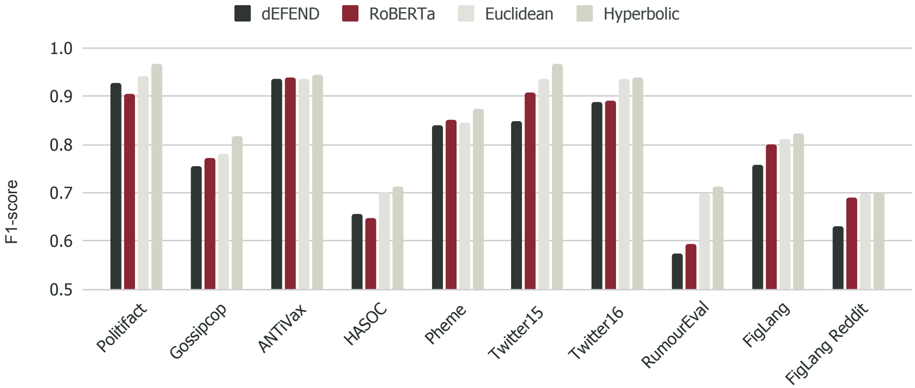


Figure 2: Performance comparisons (F1 score) of various baselines against `Hyphen`.

Ablation study

We perform ablations with two variants of our model, namely `Hyphen-hyperbolic` and `Hyphen-euclidean`, in which Hyperbolic and Euclidean represent the underlying manifold.

Effect of Fourier transform layer!

Including the **Fourier transform layer** to capture the most prominent user opinions about the source post and the most common (latent) messages conveyed by the source post, boosts the overall performance of `Hyphen`.

Effect of hyperbolic space!

`Hyphen-hyperbolic` outperforms `Hyphen-euclidean` across all datasets, owing to the **latent hierarchies** in the graph modalities used and the ability of Hyperbolic geometry to learn better hierarchical representations.

Effect of public wisdom!

We consider only source post, get rid of the co-attention block for this analysis, and keep the Fourier transform layer to **capture the latent messages in the post**. The performance degrades on removing public discourse.

Ablation study

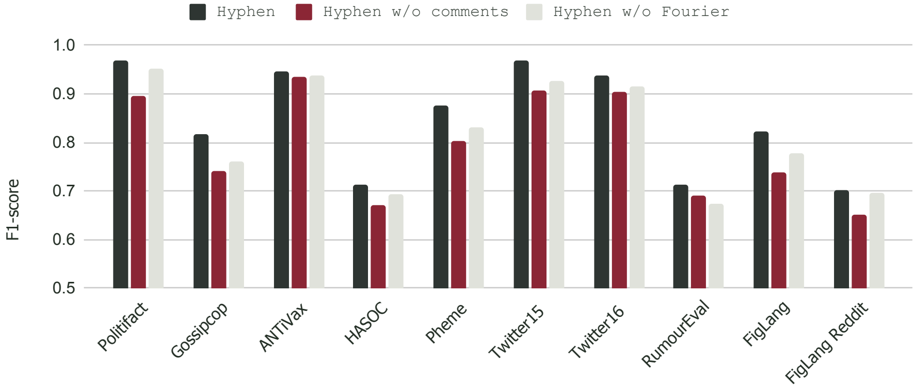


Figure 3: Ablation study for `Hyphen-hyperbolic`.

Ablation study

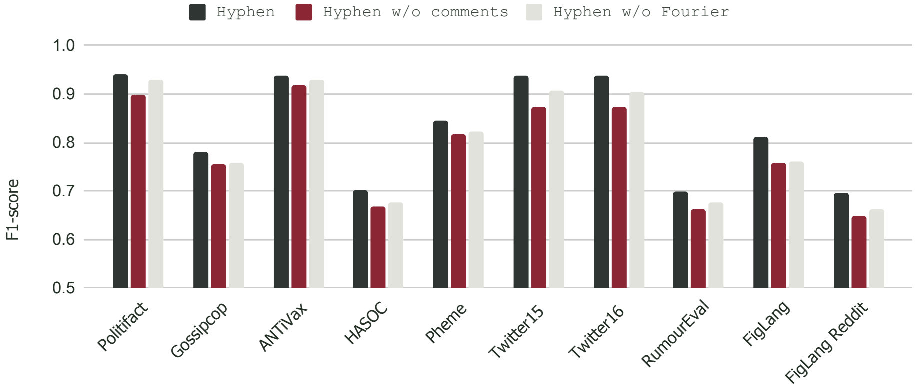


Figure 4: Ablation study for `Hyphen-euclidean`.

Overview

Explainability

- Using the hyperbolic co-attention weights, Hyphen generates an implicit rank list of sentences present in the source post, in the order of their relevance.
- Considering the Politifact dataset, we manually annotate sentences present in the source post, based on their **veracity** and rank them w.r.t to their **relevance**.
- Finally, to evaluate the performance, we compare the **correlation** between the two rank lists.

Performance comparison!

The explanations produced by dEFEND have almost no correlation to the annotated rank list. On the contrary, Hyphen shows a high positive correlation.

Model	Kendall's τ	Spearman's ρ
dEFEND	0.0231 ± 0.053	0.0189 ± 0.012
Hyphen-euclidean	0.4013 ± 0.072	0.4236 ± 0.072
Hyphen-hyperbolic	$\mathbf{0.4983 \pm 0.055}$	$\mathbf{0.5532 \pm 0.045}$

Model augmentation for Explainability!

We rule out the **Fourier sub-layer** from Hyphen for providing explainability. We cannot assert an ordered mapping from the spectral domain to the embedding space.

Thank You!

Github

Paper