Model Details

Semantic Similarity Model (SS)

The purpose of the semantic similarity model is to quantify the similarity between a pair of sentences. We use BERT (Devlin et al., 2019) finetuned on the semantic textual similarity task[1]. The Semantic Textual Similarity Benchmark (STS-B) dataset is a collection of sentence pairs drawn from news headlines and other sources (Cer et al., 2017). Each pair is annotated with a score from 1 to 5 denoting how similar the two sentences are in terms of semantic meaning. Our finetuned model achieves a pearson correlation of 89.76 which is similar to the number reported by Devlin et al. (2019) for STS-B.

Graph Attention Network (GAT)

Velikovi et al. (2018) introduced graph attention networks to address various shortcomings of GCNs. Most importantly, it enables nodes to attend over their neighborhoods features without depending on the graph structure upfront. The key idea is to compute the hidden representations of each node in the graph, by attending over its neighbors, following a self-attention (Vaswani et al., 2017) strategy. Let $Z^l = \{z_1, z_2, ..., z_n\}$ be the input node features and $Z^{l+1} = \{z'_1, z'_2, ..., z'_n\}$ be the output node features. Here n is the number of nodes in the graph and $z_i \in \mathcal{R}^F$, $z'_i \in \mathcal{R}^{F'}$, where F is the dimension of the input feature and F' is the dimension of the output feature. The attention coefficients are computed by,

$$e_{ij} = f\left(W'[Wh_i \| Wh_j]\right) \tag{1}$$

Here, W is a shared linear transformation $\in \mathcal{R}^{F \times F'}$. f is the activation function, W' is another linear layer $\in \mathcal{R}^{F' \times 1}$. Since we have a fully connected graph, we normalize the attention coefficients across all the other nodes,

$$\alpha_{ij} = \frac{\exp\left(e_{ij}\right)}{\sum_{k=1, k \neq i}^{n} \exp\left(e_{ik}\right)} \tag{2}$$

The final output features are computed as follows,

$$h'_i = \sigma\left(\sum_j \alpha_{ij} W h_j\right) \tag{3}$$

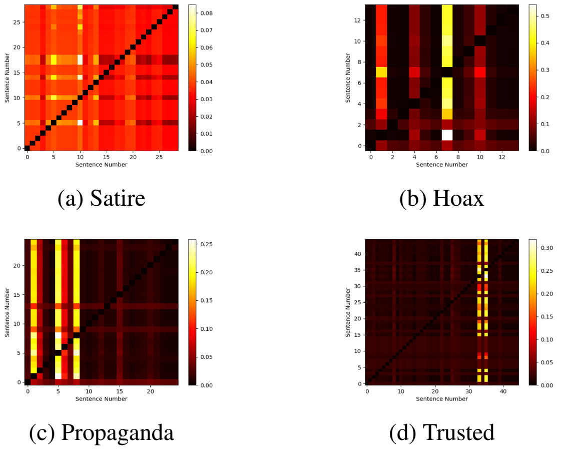
For the *GAT + 2 Attn Heads* model,

$$h'_i = \sigma\left(\sum_j \alpha_{ij} W_1 h_j\right) \| \sigma\left(\sum_j \alpha_{ij} W_2 h_j\right)$$

We can see that in the entire formulation, the adjacency matrix is not used, and hence for our fully connected graph, GAT is essentially equivalent to $GAT + SS$.

Qualitative Results

We closely inspect the attention maps generated by the GAT model for the four way classification, as shown in Figure 1.



(a) Satire

(b) Hoax

(c) Propaganda

(d) Trusted

Figure 1: Attention heatmaps generated by GAT for different types of news articles in 4-way classification.

We see that the model is learning to attend to specific sentences for trusted news articles and is attending to all the sentences for satirical articles. For the satirical articles, since there are no factual jumps, the model can't easily assign higher weights to particular sentence representation. This results in a heatmap with almost equal weights assigned to all the sentences. However, there is no obvious difference in the kind of attention maps generated for the hoax and propaganda articles. From the confusion matrix shown in Figure 2, we see that the model is highly confused between hoax and propaganda.

This confusion explains the similar nature of attention maps generated for hoax and propaganda articles. This behaviour is as per our expectation, because we faced the most confusion in classifying propaganda and hoax, given the similar nature of these kinds of articles.

[1] Task 1 of SemEval-2017

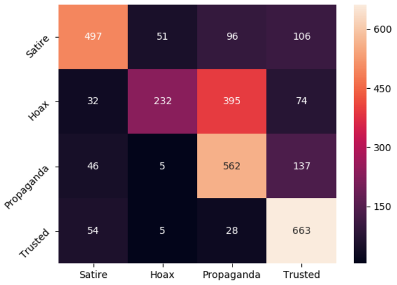


Figure 2: Confusion matrix for the GAT model in a 4-way classification setting on LUN-test.

References

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Petar Velikovi, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Li, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.