Do We Trust What They Say or What They Do? A Multimodal User Embedding Provides Personalized Explanations

Zhicheng Ren[1,*], Zhiping Xiao[2,*] and Yizhou Sun[3]

[1]*Aurora Innovation, 280 N Bernardo Ave, Mountain View, CA 94043*
[2]*University of Washington, 1410 NE Campus Pkwy, Seattle, WA 98195*
[3]*University of California, Los Angeles, Los Angeles, CA 90095*

Abstract

With the rapid development of social media, the importance of analyzing social network user data has also been put on the agenda. User representation learning in social media is a critical area of research, based on which we can conduct personalized content delivery, or detect malicious actors. Being more complicated than many other types of data, social network user data has inherent multimodal nature. Various multimodal approaches have been proposed to harness both text (i.e. post content) and relation (i.e. inter-user interaction) information to learn user embeddings of higher quality. The advent of Graph Neural Network models enables more end-to-end integration of user text embeddings and user interaction graphs in social networks. However, most of those approaches do not adequately elucidate which aspects of the data – text or graph structure information – are more helpful for predicting each specific user under a particular task, putting some burden on personalized downstream analysis and untrustworthy information filtering. We propose a simple yet effective framework called **C**ontribution-**A**ware **M**ultimodal **U**ser **E**mbedding (**CAMUE**) for social networks. We have demonstrated with empirical evidence, that our approach can provide personalized explainable predictions, automatically mitigating the impact of unreliable information. We also conducted case studies to show how reasonable our results are. We observe that for most users, graph structure information is more trustworthy than text information, but there are some reasonable cases where text helps more. Our work paves the way for more explainable, reliable, and effective social media user embedding which allows for better personalized content delivery.

Keywords

Multi-modal representation learning, Social network analysis, User embeddings

1. Introduction

The advancement of social networks has placed the analysis and study of social network data at the forefront of priorities. User-representation learning is a powerful tool to solve many critical problems in social media studies. Reasonable user representations in vector space could help build a recommendation system [1, 2], conduct social analysis [3, 4, 5], detect bot accounts [6, 7, 8], and so on. To obtain user-embeddings of higher quality, many multimodal methods are proposed to fully utilize all types of available information from the social networks, including interactive graphs, user profiles, images, and texts from their posts [9, 10, 11, 12]. Compared with models using single modality data, multimodal methods utilize more information from the social-media platforms, and hence usually achieve better results in downstream tasks.

Among all modalities in social networks, user-interactive graphs (i.e., what they do) and text content (i.e., what they say) are the two most frequently-used ones, due to their good availability across different datasets and large amount of observations. The graph-neural-network (GNN) models [13, 14, 15] makes it more convenient to fuse both the text information and graph-structure information of social-network users, where text-embeddings from language-models such as GloVe [16] or BERT [17] are usually directly incorporated into GNNs as node attributes. Although those approaches have achieved great
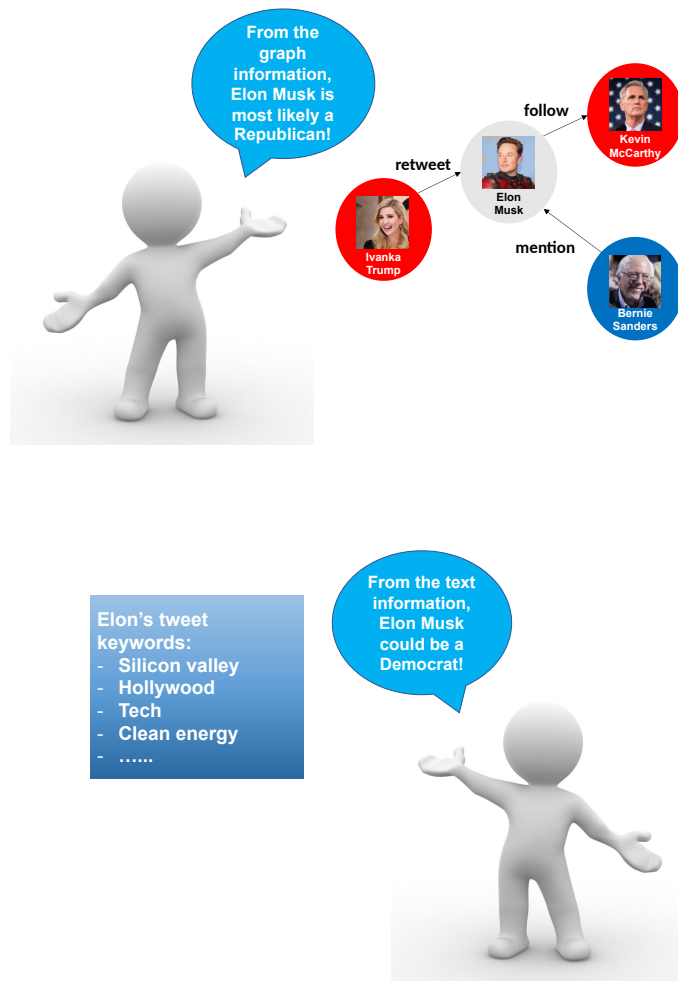

Figure 1: When predicting which political party Elon Musk voted in 2020, graph structure-based methods and text-based methods might reach opposite conclusions. Top: A small subset of Elon's activity with other Twitter users. Bottom: Some keywords extracted from Elon's tweets. All of which are extracted before year 2020.

performance in a bunch of downstream tasks [18], the text information and graph-structure information are fully-entangled with each other, which makes it hard to illustrate the two modalities' respective contributions to learning each user's representation.

It is already found by researchers that different groups of users can behave very differently on social media [19]. If such differences are not correctly captured, it might cause significant bias in the user attribute prediction (e.g., political stance prediction) [20]. Hence, when learning multi-modal user-representation, it is not only important to ask what the prediction results are, but also important to ask why we are making such predictions for different users (e.g. Are those predictions due to the same reason?). Only in that way, we could provide more insights into the user modelings, and potentially enable unbiased and personalized downstream analysis for different user groups.

On the other hand, under a multi-modality setting, if one aspect of a user's data is not trustworthy and misleading, it might still be fused into the model and make the performance lower than single-modality models [21]. Consider the case when we want to make a political ideology prediction for Elon Musk based on his Twitter content before 2020 U.S presidential election (Figure 1), when he has not revealed his clear Republican political-stance yet. If we trust the follower-followee graph structure information, we can see that he is likely to be a Republican since he follows more Republicans than Democrats, and has more frequent interactions with the verified Republicans accounts. However, in his tweet content, his word choice also shows some Democratic traits. Due to the existence of such conflicting information,

being able to automatically identify which modality is more trustworthy for each individual becomes essential in building an accurate social media user embedding for different groups of users.

To address the above two shortcomings of text-graph fusion in social networks, we propose a simple yet effective framework called **C**ontribution-**A**ware **M**ultimodal **U**ser **E**mbedding (CAMUE), which can identify and remove misleading modality from specific social network users during text-graph fusion, in an explainable way. CAMUE uses a learnable attention module to decide whether we should trust the text information or the graph structure information when predicting individual user attributes, such as political stance. Then, the framework outputs a clear contribution map for each modality on each user, allowing personalized explanations for downstream analysis and recommendations. For ambiguous users whose text and graph structure information disagree, our framework could successfully mitigate the unreliable information among different modalities by automatically adjusting the weight of different information accordingly.

We conduct experiments on the TIMME dataset [21] used for a Twitter political ideology prediction task. We observed that our contribution map can give us some interesting new insights. A quantitative analysis of different Twitter user sub-groups shows that link information (i.e., interaction graph) contributes more than text information for most users. This provides insights that political advertising agencies should gather more interaction graph information of Twitter users in the future when creating personalized advertisement content, instead of relying too much on their text data. We also observe that when the graph and text backbone are set to R-GCN and GloVe respectively, our approach successfully ignores the unreliable GloVe embedding and achieves better prediction results. When the text modality is switched to a more accurate BERT embedding, our framework can assign graph/text weights for different users accordingly and achieve comparable performance to existing R-GCN-based fusion methods. We pick 9 celebrities among the 50 most-followed Twitter accounts [1], such as Elon Musk. A detailed qualitative analysis of their specific Twitter behaviors shows that our contribution map models their online behaviors well. Finally, we run experiments on the TwiBot-20-Sub dataset [22] used for a Twitter human/bot classification task, showing that our framework could be generalized to other user attribute prediction tasks. By creating social media user embeddings that are more explainable, reliable, and effective, our framework enables improved customized content delivery.

2. Preliminaries and Related Work

2.1. Multimodal Social Network User Embedding

Social network user embedding is a popular research field that aims to build accurate user representations. A desirable user embedding model should accurately map sparse user-related features in high-dimensional spaces to dense representations in low-dimensional spaces. Multimodal social network user embedding models utilize user different types of user data to boost their performance. Commonly-seen modality combinations include graph-structure (i.e. link) data and text data [23], graph-structure data and tabular data [24, 25, 9], and graph-structure data, text data and image data altogether [26] [27], etc.

Among those multi-modality methods, the fusion of graph-structure data and text data has always been one of the mainstream approaches for user embedding. At an earlier stage, without much help from the GNN models, most works trained the network-embedding and text-embedding separately and fused them using a joint loss [28, 29, 30, 31]. With help of the GNN models, a new type of fusion method gained popularity, where the users' text-embeddings are directly incorporated into GNNs as node attributes [23, 32, 33].

Despite their good performance, all existing models did not explain how much the graph structure or text information of a particular user contributed to its final prediction result, make it difficult to give customized modality weight for downstream analysis or recommendations. Also, if one modality is very poorly learned, it can be counter-effective to the user embedding quality, make it even worse

---

than their single-modality counterparts [21]. How to address this problem in a universally-learned way instead of heuristic-based information filtering, has largely gone under-explored. Hence, we propose a framework that would not only utilize both text and graph-structure information together but also reveal their relative importance along with our prediction result.

2.2. Graph Neural Network

Graph Neural Network (GNN) refers to a collection of deep learning models that learn node embedding through iterative aggregation of information from neighboring nodes, using a convolutional operator. The majority of the GNN architectures include a graph convolution layer in a form that can be characterized as message-passing and aggregation. A general formula for such convolution layers is:

$$\mathbf{H}^{(l)} = \sigma(\tilde{\mathbf{A}}\mathbf{H}^{(l-1)}\mathbf{W}^{(l)}), \tag{1}$$

where $\mathbf{H}^{(l)}$ represents the hidden node representation of all nodes at layer l, operator σ is a non-linear activation function, and the graph-convolutional filter $\tilde{\mathbf{A}}$ is a matrix that usually takes the form of a transformed (e.g., normalized) adjacency matrix $\mathbf{A}$, and the layer-l's weight $\mathbf{W}^{(l)}$ is learnable.

In the past few years, GNN models have reached SOTA performances in various graph-related tasks, and are widely regarded as a very promising technique to generate node embedding for users in social-network graphs. [13, 34, 14, 15].

2.3. Neural Network-based Language Models

The field of natural language processing has undergone a significant transformation with the advent of neural-network-based language models. Word2Vec [35] introduced two architectures: Continuous Bag-of-Words (CBOW) and Skip-Gram. CBOW predicts a target word given its context, while Skip-Gram predicts context words given a target word. GloVe [16] model went beyond by incorporating global corpus statistics into the learning process. ELMo [36] was another significant step forward, as it introduced context-dependent word representations, making it possible for the same word to have different embeddings if the context is different. BERT [17] is a highly influential model that is built on the transformer architecture [37], pre-trained on large text corpora using, for example, masked language modeling and next-sentence prediction tasks. Recently, large language models like GPT-3 [38], InstructGPT [39], and ChatGPT have achieved significant breakthroughs in natural-language-generation tasks. All of those large language models (LLMs) are frequently used to generate text embedding for social network users.

Our framework does not rely on any specific language model, and we do not have to use LLMs. Instead, we use language-models as a replaceable component, making it possible for either simpler ones like GloVe or more complicated ones like BERT to fit in. We will explore some different options in the experimental section.

2.4. Multimodal Explanation Methods

In the past, several methods have been proposed to improve the interpretability and explainability of multimodal fusions [40]. Commonly used strategies include attention-based methods [41, 42, 43], counterfactual-based methods [44, 45], scene graph-based methods [46] and knowledge graph-based methods [47]. Unfortunately, most of them focus on the fusion of image modality and text modality, primarily the VQA task, while to the best of our knowledge, no work focuses on improving the explainability between the network structure data and text data in social-network user embedding.

3. Problem Definition

Our general goal is to propose a social network user embedding fusion framework that could answer: 1. which modality (i.e. text or graph structure, saying or doing) contributes more to our user attribute

prediction, hence allowing more customized downstream user behavior analysis and 2. which modality should be given more trust for each user, and automatically filter out the untrustworthy information when necessary, in order to achieve higher-quality multi-modal user-embedding.

3.1. Problem Formulation

A general framework of our problem could be formulated as follows: given a social media interaction graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with node set $\mathcal{V}$ representing users and edge set $\mathcal{E}$ representing links between users. Let $\mathbf{X} = [x_1, x_2, x_3, \cdots, x_n]$ be the text content of $n = |\mathcal{V}|$ users, $\mathbf{Y} = [y_1, y_2, y_3, \cdots, y_n]$ be the labels of those users, $\mathbf{A} = [\mathbf{A}^1, \mathbf{A}^2, \cdots, \mathbf{A}^m]$ be the adjacency matrices of $\mathcal{G}$, m be the number of link types and $\mathbf{A}^i \in \mathbb{R}^{n \times n}$, our training objective is:

$$\min \mathbb{E}\left[\mathcal{L}\left(f\left(\mathcal{G}, \mathbf{X}\right), \mathbf{Y}\right)\right] \tag{2}$$

Here, $\mathcal{L}$ is the loss of our specific downstream task, and f is some function that combines the graph structure information and text information, producing a joint user embedding.

3.2. Preliminary Experiment

To investigate the effectiveness of the existing GNN-based multimodal fusion methods in filtering the unreliable modality when the graph structure and text contradict, we run experiments using a common fusion method that feeds the fine-tuned BERT features into the R-GCN backbone, similar to the approaches in [23] and [10]. We observe that this conventional fusion method fails to filter the unreliable information for some of those ambiguous users. Table 1 show two politicians whose Twitter data contains misleading information, either in the graph structure or text data. While the single-modality backbones which are trained without misleading information give the correct predictions, the multi-modality fusion method is fooled by the misleading information and is not able to make correct predictions.

These insights revealed the importance of having a more flexible and explainable framework for learning multimodal user embedding.

4. Methodology

We propose a framework of **C**ontribution-**A**ware **M**ultimodal **U**ser **E**mbedding (CAMUE), a fusion method for text data and graph structure data when learning user embedding in social networks. The key ingredient of this framework is an attention gate-based selection module which is learned together with the link and text data, and decide which information we want to trust more for each particular user.

Our framework has three main parts: a text encoder, a graph encoder, and an attention-gate learner. The text content of each user passes through the text encoder and generates a text embedding for that user. The embedding is then passed through a three-layer MLP for fine-tuning. The adjacency matrix of the users passes through the graph encoder and generates a node embedding for that user. At the same time, both the text embedding and the graph adjacency matrix pass through our attention gate learner. The output of this module is two attention weights, α and β, which control the proportion of our graph structure information and text information. Without loss of generality, if we make R-GCN our graph encoder and BERT our text encoder, our model will be trained in the following way (Equation 3-6, also illustrated in Figure 2):

$$\mathbf{H}^{(1)} = \sigma(\text{concat}(\mathbf{A}^1 + \mathbf{A}^2 + \cdots + \mathbf{A}^m, \text{BERT}emb\left(\mathbf{X}\right)\mathbf{W}^{(1)}) \tag{3}$$

$$\mathbf{H}^{(2)} = \sigma(\mathbf{H}^{(1)}\mathbf{W}^{(2)}) \tag{4}$$

Table 1

Left: A snippet for former U.S. Rep. Ryan Costello, his tweet text content is informative for the political Party prediction, but his Twitter interaction graph data could be misleading.

Right: A snippet for U.S. Senator Sheldon Whitehouse, his Twitter interaction graph data is informative for the political Party prediction, but his tweet text content could be misleading.

Name	Ryan Costello
Ground Truth Party	Republican
Sample Graph Data	Liked Ben Rhodes (Democrat) 20 times. Liked Donald Trump 0 time. Following Mike Quigley (Democrat).
Sample Text Data	Despite Trump, Iran's elections & chaotic ME, some Democrats want to race ahead with ill-conceived Iran sanctions RT @SaeedKD: Iran's people care about elections. The so-called democratic fringe doesn't - by me
Graph-backbone Prediction	Democrat (Wrong)
Text-backbone Prediction	Republican (Right)
Fused Model Prediction	Democrat (Wrong)

Name	Sheldon Whitehouse
Ground Truth Party	Democrat
Sample Graph Data	Liked Senate Democrats Official Account 26 times. Not following Donald Trump. Following Barack Obama.
Sample Text Data	My Republican partner on the CARA bill, @SenRobPortman, writes a powerful editorial on the success of CARA and CURES (which provided a needed boost of funding to match CARA). Good move by Trump Administration. Cong. @JimLangevin &
Graph-backbone Prediction	Democrat (Right)
Text-backbone Prediction	Republican (Wrong)
Fused Model Prediction	Republican (Wrong)

$$[e_\alpha, e_\beta] = \mathbf{H}^{(2)}\mathbf{W}^{(3)} \tag{5}$$

$$\alpha = \mathrm{softmax}(e_\alpha), \ \beta = \mathrm{softmax}(e_\beta) \tag{6}$$

Where $\mathbf{H}$ and $\mathbf{W}$ are hidden layers and weights of our attention gate learner, $\mathbf{X} = [x_1, x_2, x_3, \cdots, x_n]$ is the text content, BERTemb is the BERT encoding module, $\mathbf{A} = [\mathbf{A}^1, \mathbf{A}^2, \cdots, \mathbf{A}^m]$ is the adjacency matrices of $\mathcal{G}$ and m is the number of link types.

Then, our overall training objective becomes:

$$\min \mathbb{E}[\mathcal{L}((\alpha + \lambda) \ \mathrm{R\text{-}GCN}emb(\mathcal{G})$$
$$+ (\beta + \lambda) \ \mathrm{BERT}emb(\mathbf{X}), \mathbf{Y})]$$

Here, λ acts as a regularizer to ensure our model is not overly dependent on a single modality.

Our methods offer two levels of separation. First, we separate the text encoder and graph encoder to allow better disentanglement on which data contributes more to our final prediction results. Second, we separate the learning of the downstream tasks and the learning of which data modality (i.e. text or graph structure) we can rely on more. This makes our framework adaptable to different downstream social media user prediction tasks. The learned trustworthiness of different modalities allows for auto-adjustment of the weight between graph structure and text modalities, hence filtering any unreliable information once they are discovered.

Figure 2 shows the overall architecture of our framework, note that the graph structure encoder and text encoder could be replaced by any other models which serve the same purposes.
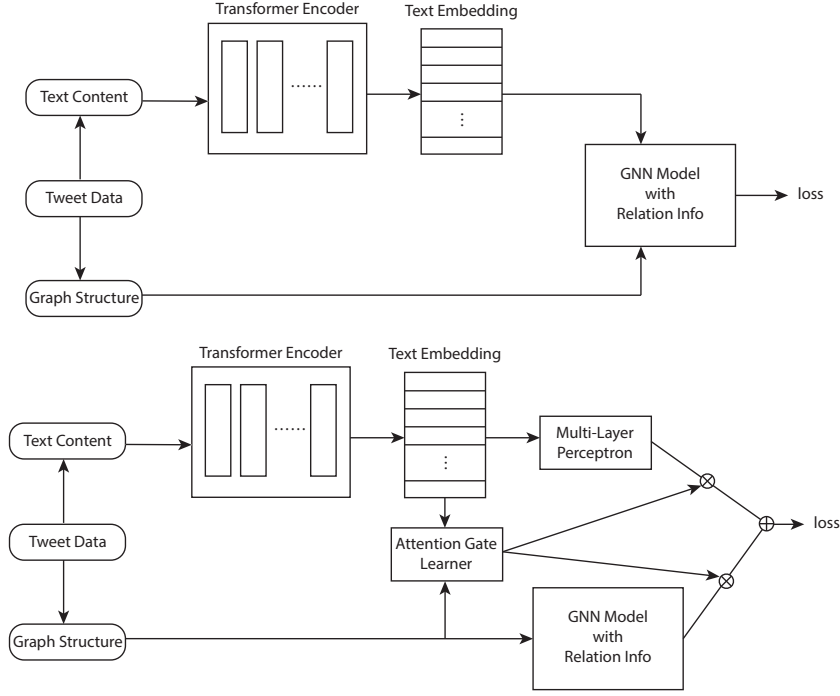
Figure 2: The architectures of our framework. Top: Simple Fusion method for GNN (baseline), bottom: CAMUE

We give a short complexity analysis of our architecture for the case of R-GCN + BERT: Since we are using sparse adjacency matrix for R-GCN, the graph encoder part has a complexity of $\mathcal{O}(L_{\text{graph}}EF_{\text{graph}} + L_{\text{graph}}NF_{\text{graph}}^2)$ (according to [48]), where L is the number of layers, E is the number of edges, N is the number of nodes, and F is the feature dimension. Since we fixed the maximum text length to be a constant for the text encoder, it has a complexity of $\mathcal{O}(F_{\text{text}}^2)$ (based on [49]). Since F_{text} and F_{graph} are about comparable size, our fusion module has the complexity of $\mathcal{O}(F_{\text{graph}}^2 + F_{\text{text}}^2)$, so the overall complexity is $\mathcal{O}(L_{\text{graph}}EF_{\text{graph}} + L_{\text{graph}}NF_{\text{graph}}^2 + F_{\text{text}}^2)$, hence we are not adding extra time complexity.

5. Experiments

5.1. Tasks and Datasets

We run experiments on two Twitter user prediction tasks: 1. Predicting the political ideology of Twitter users (Democrat vs Republican) and 2. Predicting whether a Twitter user account is a human or a bot.

5.1.1. TIMME

TIMME [21] introduced a multi-modality Twitter user dataset as a benchmark of political ideology prediction task for Twitter users. TIMME contains $21,015$ Twitter users and $6,496,112$ Twitter interaction links. Those links include follows, retweets, replies, mentions, and likes, together they form a large heterogeneous social network graph. TIMME also contains $6,996,310$ raw Twitter content from those users. Hence, it will be a good dataset to study different fusion methods of text features and graph structure features. In TIMME, there are 586 labeled politicians and $2,976$ randomly sampled users with a known political affiliation. Some of them are ambiguous users we investigated before. Labeled nodes belong to either Democrats or Republicans. Note that the dataset cut-off time is 2020, so the political polarity of many public figures (e.g. Elon Musk) have not been reviewed at that time.

5.1.2. TwiBot-20-Sub

TwiBot-20 [22] is an extensive benchmark for Twitter bot detection, comprising $229,573$ Twitter accounts, of which $11,826$ are labeled as human users or bots. The dataset also contains $33,716,171$ Twitter interaction links and $33,488,192$ raw Twitter content. The links in TwiBot-20 include follows, retweets, and mentions. To further examine the generalizability of our method, we run experiments for Twitter bot account detection on the TwiBot-20 dataset. To reduce the computation cost of generating node features and text features, we randomly subsample $3,000$ labeled users and $27,000$ unlabeled users from the TwiBot-20 dataset, and form a new dataset called TwiBot-20-Sub. In this way, the size and label sparsity of the TwiBot-20-Sub dataset becomes comparable with the TIMME dataset.

5.1.3. Train-test Split

We split the users of both datasets into an $80\%:10\%:10\%$ ratio for the training set, validation set, and test set respectively.

5.2. Implementation Detail

To test the effectiveness of our framework across different models, we choose two single-modality text encoders, GloVe and BERT, and two single-modality graph encoders, MLP and R-GCN.

The GloVe embedding refers to the Wikipedia 2014 + Gigaword 5 (300d) pre-trained version. [2] The BERT embedding refers to the sentence level ($[CLS]$ token) embedding of BERT-base model [50] after fine-tuning the pre-trained model's parameters on the tweets from our training set consisting of 80% of the users. We chose a max sequence length of 32. After the encoding, we have a 300-dimension text embedding for GloVe and a 768-dimension text embedding for BERT.

We choose a modified version of R-GCN from TIMME [21] as an R-GCN graph encoder. R-GCN [34] is a GNN model specifically designed for heterogeneous graphs with multiple relations. In the TIMME paper, it is discovered that assigning different attention weights to the relation heads of the R-GCN model could improve its performance. Hence, we adopt their idea and use the modified version of R-GCN. We did not use the complete TIMME model since it is designed for multiple tasks which is outside our research scope, and will overly complicate our model.

We also choose a 3-layer MLP as another graph encoder for comparison, the adjacency list for each user is passed to the MLP.

Large language models (LLMs) like ChatGPT are powerful in understanding texts, but they usually have a great number of parameters, making traditional supervised fine-tuning a hard and costly task [38]. Instead, less resource-intensive methods like few-shot learning, prompt tuning, instruction tuning, and chain-of-thought are more frequently used to adapt LLMs on specific tasks [51]. We do not use large language models as one of the options for the text encoder since those methods are not compatible with our framework – they do not provide a well-defined gradient to train our attention gate learner.

We run experiments on a single NVIDIA Tesla A100 GPU. We used the same set of hyper-parameters as in the TIMME paper, with the learning rate being 0.01, the number of GCN hidden units being 100, and the dropout rate being 0.1, on a PyTorch platform. For a fair comparison, we run over 10 random seeds for each algorithm on each task.

6. Results and Analysis

6.1. Contribution Map

To show that our framework could essentially provide personalized explanations during the fuse of modalities, we draw the contribution map based on α (graph weight) and β (text weight) attention for users from each dataset. The darker the color is, the weight of the corresponding modality is closer to 1.

---

[2] glove.6B.zip from https://nlp.stanford.edu/projects/glove/

In the contribution map, pure white indicates a zero contribution (0) from a modality, while pure dark blue indicates a full contribution (1).

The top figure of Figure 3 shows the contribution map output when the text encoder is BERT and the graph encoder is R-GCN, on a subgroup of the TIMME dataset consisting of some politicians and some random Twitter users. To avoid any misuse of personal data information, we hide the names of random Twitter users and only include politicians whose Twitter accounts are publically available at [3]. As we can see, there is a clear cut between the percentage of contributions from different modalities to the final prediction. It is notable that for the two ambiguous politician users we mentioned earlier (Ryan Costello and Sheldon Whitehouse), CAMUE could give correct attention, where we should trust more text data from Mr. Costello while trusting more graph structure data from Mr. Whitehouse.

The bottom figure of Figure 3 shows the contribution map output when the text encoder is GloVe and the graph encoder is R-GCN, on the same subgroup of the TIMME dataset. Note that for all shown users text information does not contribute to the final prediction. This could be attributed to the fact that GloVe is not very powerful for sentence embedding, especially when the text is long. This contribution map shows us our framework filters out the text modality almost completely when it is not helpful for our user embedding learning. As we can see from table 2, the traditional fusion method for GloVe+R-GCN only yields an accuracy of 0.840, which is much lower than the single graph structure modality prediction (0.953) using R-GCN, due to unreliable GloVe embedding. In contrast, our CAMUE method obtains a higher accuracy (0.954) than the single modality models by disregarding the unreliable information.

Figure 4 shows the contribution map output for the same set of encoders on a subgroup of the Twibot-20-Sub dataset. There is also a clear cut between the percentage of contributions from different modalities, for both the human Twitter accounts and bot accounts.

Hence, we verify that our framework could both provide personalized modality contribution and drop low-quality information during the fuse of modalities. Some quantitative analysis of how this low-quality information filtering could benefit the general model performance could be found in the next section, and some qualitative analysis about what new insights we could gain from the output of our framework could be found in the case study section.

6.2. General Performance

Table 2 shows the performance of CAMUE on different combinations of encoders. The traditional fusion method in Figure 2 is denoted as "simple fusion". For MLP, we do not have such a natural fusion method. We also add "CAMUE, fixed params" as an ablation experiment to prove the effectiveness of our attention gate-based selection module.

We observe that within those combinations, sometimes simple fusion methods are significantly worse than single-modality methods (e.g. GloVe+R-GCN vs R-GCN only) due to some untrustworthiness in one of the modalities. However, any fusion under our CAMUE framework always performs better than their respective single modality methods. That suggests that our algorithm can benefit from attending to the more reliable modality between text and graph structure, if one particular modality is not trustworthy (e.g. GloVe embedding), and learning not to consider it when making predictions (as we can see in Figure 3, bottom).

It is also notable that our CAMUE method outperforms "CAMUE, fixed params". These results suggest that adjusting the weight of different modalities dynamically yields better performance than fixed weights of modalities. Finally, when the text modality is switched to a more accurate BERT embedding, our framework still gives comparable performance to its corresponding simple fusion methods.

6.3. Case Studies

User Sub-groups Table 3 gives a quantitative analysis when the text encoder is BERT and the graph encoder is R-GCN, for different sub-groups of Twitter users we are interested in. In general, graph

---

Table 2
The overall performance (format: accuracy ; f1-score) of CAMUE on different social media data sets.

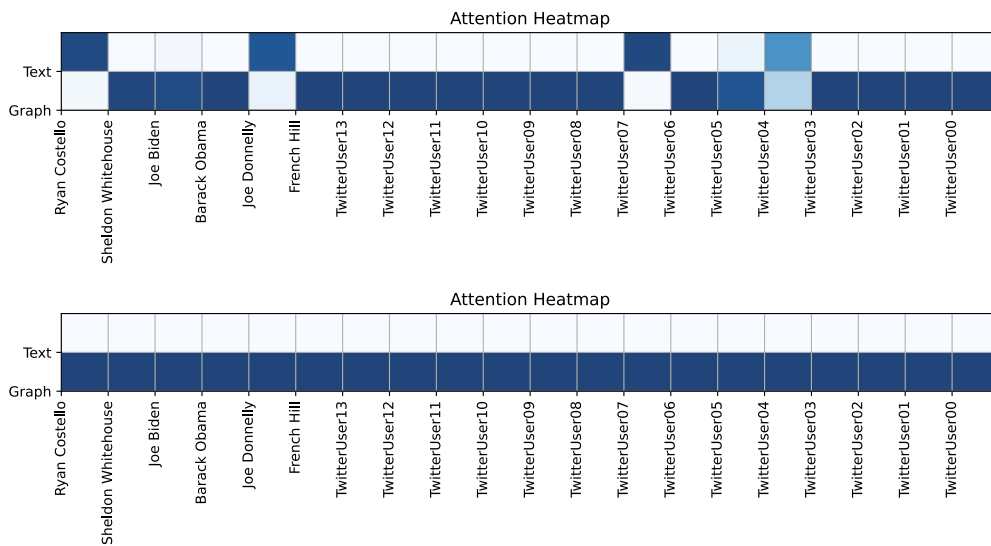| Algorithm	Encoder Variant		Data Set	
	Text	Graph	TIMME	TwiBot-20-Sub
text-only	GloVe	N/A	0.688 ; 0.681	0.565 ; 0.511
	BERT		0.862 ; 0.859	0.731 ; 0.722
link-only	N/A	MLP	0.932 ; 0.930	0.707 ; 0.697
		R-GCN	0.953 ; 0.953	0.735 ; 0.728
simple fusion	GloVe	R-GCN	0.840 ; 0.837	0.683 ; 0.675
	BERT	R-GCN	0.959 ; 0.959	0.791 ; 0.787
CAMUE w. fixed params	GloVe	MLP	0.938 ; 0.937	0.700 ; 0.691
		R-GCN	0.952 ; 0.951	0.734 ; 0.727
	BERT	MLP	0.940 ; 0.938	0.732 ; 0.722
		R-GCN	0.952 ; 0.951	0.779 ; 0.771
CAMUE	GloVe	MLP	0.945 ; 0.944	0.707 ; 0.697
		R-GCN	0.954 ; 0.953	0.738 ; 0.731
	BERT	MLP	0.935 ; 0.933	0.744 ; 0.738
		R-GCN	0.961 ; 0.960	0.782 ; 0.776

Figure 3: Contribution map for TIMME dataset, top: CAMUE(BERT, R-GCN), bottom: CAMUE(GloVe, R-GCN)
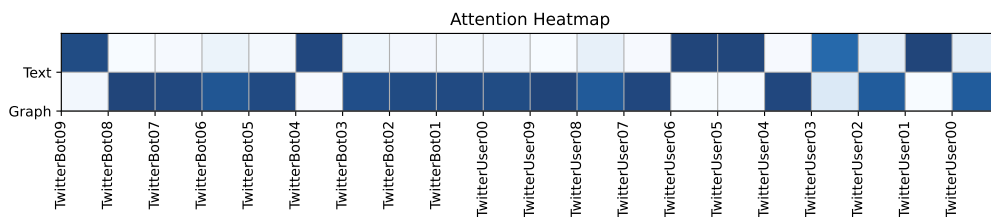

Figure 4: Contribution map for TwiBot-20-Sub dataset, CAMUE(BERT, R-GCN)

structure information contributes the most when it comes to bot accounts. One possible explanation for this is the variety of bot accounts on Twitter, such as those for business advertising, political outreach, and sports marketing [22]. Bots with different usage might talk very differently, however, they may share some common rule-based policies when trying to interact with humans on Twitter [52, 53].

Graph structure information contributes the second highest when it comes to politicians. This is also not surprising since politicians are generally more inclined to retweet or mention events related to their political parties [21]. It is also notable that the weight of text information for Republicans is slightly

less than it is for Democrats. This aligns with the findings in [54] that Democrats have a slightly more politically polarized word choice than Republicans.

For random users, the weight of text information is the largest, although still not as large as the weight of graph structure information. This could be attributed to the pattern that many random users interact frequently with their non-celebrity families and friends on Twitter, who are more likely to be politically neutral.

Table 3
% Users whose $\alpha > \beta$ for different subgroups

Subgroup	% Users
Democrats	70.9
Republicans	76.1
Politicians	<u>76.2</u>
Non-politicians with Party affiliations	72.4
Non-bot random users	61.2
Bot accounts	**77.3**
TIMME, aggregated	73.5
TwiBot-20-Sub, aggregated	70.1

Table 4
Predicted Political Stance of Some News Agencies

News Agency	Prediction	Text or Graph
New York Times	D	Graph
Washington Post	D	Graph
Wall Street Journal	R	Text
USA Today	D	Graph
CNN	D	Graph
Fox News	R	Text
Guardian	D	Text
Associated Press	R	Graph
US News	D	Graph
MSNBC	D	Graph
BBC	R	Graph
National Review	R	Graph
Bloomberg	D	graph

Table 4 shows some predicted political stances and the main contributing modalities of a group of news agencies. We could see that the majority of them have more reliable graph structure information than text information. This is not surprising since most news agency tends to use neutral words to increase their credibility, hence it is hard to gather strong political stances from their text embedding, except for some of them like Fox News and Guardian which are known to use political polarized terms more often [54, 55]. Our framework is able to capture this unique behavior pattern for Fox News and Guardian, meanwhile giving mostly accurate political polarity predictions aligning with results in [54] and [4].

To conclude, we are able to obtain customized user behavior patterns through our multi-modality fusion. Those patterns could provide insights on which modality we should focus on more for different type of users, for downstream tasks such as personalized recommendations, social science analysis, or malicious user detection.

Selected Celebrities from TIMME Dataset Since we are not allowed to disclose regular Twitter users' information, we instead selected 9 celebrities among the top 50 most followed Twitter

---

[4]https://www.allsides.com/media-bias/ratings

Name	Elon Musk	Lebron James	Lady Gaga	Bill Gates	Oprah Winfrey	Jimmy Fallon	Katy Perry	Justin Timberlake	Taylor Swift
Photo									
Field	Business	Sports	Music	Business	Television	Television	Music	Music	Music
Text or Graph?	Graph	Text	Text	Graph	Graph	Graph	Text	Graph	Graph
% Text	1	93	100	33	15	0	96	58	1
% Graph	99	7	0	67	85	100	4	42	99
Political Polarity Prediction	R	D	D	R	D	D	D	D	R

Figure 5: A subset of the top 50 followed celebrities, note that the political polarity prediction is coming from our model and may not reflect their actual stances.

accounts from [5], whose Twitter accounts appear in the TIMME dataset, as a case study to show our frameworks' capability to give personalized explanations. We run the political polarity prediction task on those people and obtained some predictions. We also record the percentage of text information and percentage of graph structure information that contributes to their political polarity prediction (See Figure 5).

- Elon Musk: Before 2020 (dataset cut-off), Elon Musk's political views in his tweet text content are often complex. He has claimed several times not to take the viewpoints in his tweets too seriously [6]. This aligns with the low contribution weight of his texts on his political stand prediction. However, on the graph level, 66.67% of the politicians Elon Musk liked have liked Trump at least once, which is significantly larger than the average number in the TIMME dataset (23.67%). This could be a strong reason why our graph structure weight is so high and why we predict Elon Musk to be Republican-leaning. Our prediction is proved correct when in 2022 (which is beyond our dataset cut-off time, 2020 [54]), Elon Musk claimed that he would vote for Republicans in his tweet [7]. This is a strong indicator that our framework is using correct information.
- LeBron James: In his tweets, LeBron James frequently shows his love and respect to Democratic President Obama [8]. Our prediction for him to be Democrat-leaning with a strong text contribution aligns with this observation.
- Lady Gaga: Similarly to James, Lady Gaga also expresses explicitly in her tweets about her support of Democratic candidates [9]. Our graph weight becomes 0 for her case, meaning that using the text alone we could make sure she is Democrat-learning.
- Bill Gates: He usually avoids making explicit statements about whether he supports Democrats or Republicans in his tweets. Although our model predicts him as the Republican, the probability edge is very marginal (11%).

- Oprah Winfrey: During the 2016 presidential campaign, she retweeted and mentioned her support for Democratic candidate Hillary Clinton frequently [10], making the graph structure information a strong indicator of her Democratic stance.
- Jimmy Fallon: Jimmy Fallon has managed to maintain a sense of political neutrality in his tweets. His text contribution to the final prediction is 0. Even though the Twitter graph structure information indicates that he is Democrat-leaning, we still do not know for sure in real life whether he is a Democrat or Republican.
- Katy Perry: Just like Oprah Winfrey, Katy Perry also interacted with and supported Hillary Clinton during the 2016 election, a reason why we predict her as Democrat-leaning from the graph structure. Although she supports some republican politicians in 2022 [11], that is beyond the dataset cutoff.
- Justin Timberlake: Justin Timberlake has frequent positive interactions with President Obama [12] and firmly supports Hillary Clinton in his tweets [13], both suggesting that he is Democrat-learning. Our model assigns a similar weight to text and graph structure, suggesting that both of them are effective in making that prediction.
- Taylor Swift: In the case of Taylor Swift, the model fails to give the correct prediction. Her tweets show that she voted for Biden during 2020 [14], but the prediction is Republican. One reason is that at the graph structure level, the majority of Taylor Swift's followers are classified as Republican (67.09 %) in the dataset, which could mislead the graph encoder.

Overall, we find out that for those celebrities, graph structure information is usually more useful when making political polarity predictions. That aligns with the quantitative results in table 3. As we see, different celebrities could have very different behavior patterns, and those patterns could be correctly captured and explained by our contribution weight. That confirms the effectiveness of our framework.

7. Conclusion

In this paper, we investigate some potential limitations of existing fusion methods for text information and graph structure information in user representation learning from social networks. We then propose a contribution aware multimodal social-media user-embedding with a learnable attention module. Our framework can automatically determine the reliability of text and graph-structure information when learning user-embeddings. It filters out unreliable modalities for specific users across various downstream tasks. Since our framework is not bound to any specific model, it has great potential to be adapted to any graph-structure-embedding component and text-embedding component, if affordable. More importantly, our models can give a score on the reliability of different information modalities for each user. That gives our framework great capability for personalized downstream analysis and recommendation. Our work can bring research attention to identifying and removing misleading information modality due to differences in social network user behavior, and paves the way for more explainable, reliable, and effective social media user representation learning.

Some possible future extensions include adding more modalities other than text and graphs (e.g., image and video data from user's posts). Also, we consider the user identities to be static throughout our analysis, which might not be the case in many scenarios. We can bring time as a factor to produce a multi-modality dynamic social media user embedding. For example, it is possible to observe that a user's text content is more trustworthy in the first few months, and then that user's interactive graph structure information or interaction becomes more reliable in longer-terms.

---

[10]https://twitter.com/Oprah/status/780588770726993920

[11]https://twitter.com/katyperry/status/1533246681910628352

[12]https://twitter.com/jtimberlake/status/1025867320407846912

[13]https://twitter.com/jtimberlake/status/768191007036891136

[14]https://twitter.com/taylorswift13/status/1266392274549776387

References

[1] W. X. Zhao, S. Li, Y. He, E. Y. Chang, J.-R. Wen, X. Li, Connecting social media to e-commerce: Cold-start product recommendation using microblogging information, IEEE Transactions on Knowledge and Data Engineering 28 (2016) 1147–1159. doi:10.1109/TKDE.2015.2508816.

[2] W. Fan, Y. Ma, Q. Li, Y. He, E. Zhao, J. Tang, D. Yin, Graph neural networks for social recommendation, in: The World Wide Web Conference, WWW '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 417–426. URL: https://doi.org/10.1145/3308558.3313488. doi:10.1145/3308558.3313488.

[3] D. Preoţiuc-Pietro, Y. Liu, D. Hopkins, L. Ungar, Beyond binary labels: Political ideology prediction of Twitter users, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 729–740. URL: https://aclanthology.org/P17-1068. doi:10.18653/v1/P17-1068.

[4] T. Islam, D. Goldwasser, Twitter user representation using weakly supervised graph embedding, Proceedings of the International AAAI Conference on Web and Social Media 16 (2022) 358–369. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/19298. doi:10.1609/icwsm.v16i1.19298.

[5] J. Jiang, X. Ren, E. Ferrara, Retweet-bert: Political leaning detection using language features and information diffusion on social networks, Proceedings of the International AAAI Conference on Web and Social Media 17 (2023) 459–469. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/22160. doi:10.1609/icwsm.v17i1.22160.

[6] O. Varol, E. Ferrara, C. Davis, F. Menczer, A. Flammini, Online human-bot interactions: Detection, estimation, and characterization, Proceedings of the International AAAI Conference on Web and Social Media 11 (2017) 280–289. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14871. doi:10.1609/icwsm.v11i1.14871.

[7] S. Kudugunta, E. Ferrara, Deep neural networks for bot detection, Information Sciences 467 (2018) 312–322. URL: https://www.sciencedirect.com/science/article/pii/S0020025518306248. doi:https://doi.org/10.1016/j.ins.2018.08.019.

[8] L. H. X. Ng, K. M. Carley, Botbuster: Multi-platform bot detection using a mixture of experts, Proceedings of the International AAAI Conference on Web and Social Media 17 (2023) 686–697. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/22179. doi:10.1609/icwsm.v17i1.22179.

[9] Z. Jin, X. Zhao, Y. Liu, Heterogeneous graph network embedding for sentiment analysis on social media, Cognitive Computation 13 (2021) 81–95.

[10] Q. Guo, H. Xie, Y. Li, W. Ma, C. Zhang, Social bots detection via fusing bert and graph convolutional networks, Symmetry 14 (2021) 30.

[11] L. Wu, P. Sun, R. Hong, Y. Fu, X. Wang, M. Wang, Socialgcn: An efficient graph convolutional network based model for social recommendation, CoRR abs/1811.02815 (2018). URL: http://arxiv.org/abs/1811.02815. arXiv:1811.02815.

[12] F. Huang, X. Zhang, J. Xu, C. Li, Z. Li, Network embedding by fusing multimodal contents and links, Knowledge-Based Systems 171 (2019) 44–55.

[13] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: International Conference on Learning Representations, 2017. URL: https://openreview.net/forum?id=SJU4ayYgl.

[14] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, Advances in neural information processing systems 30 (2017).

[15] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, et al., Graph attention networks, stat 1050 (2017) 10–48550.

[16] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[18] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, M. Sun, Graph neural networks: A review of methods and applications, AI open 1 (2020) 57–81.

[19] E. Mustafaraj, S. Finn, C. Whitlock, P. T. Metaxas, Vocal minority versus silent majority: Discovering the opionions of the long tail, in: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, 2011, pp. 103–110. doi:10.1109/PASSAT/SocialCom.2011.188.

[20] G. Blank, The digital divide among twitter users and its implications for social research, Social Science Computer Review 35 (2017) 679–697.

[21] Z. Xiao, W. Song, H. Xu, Z. Ren, Y. Sun, Timme: Twitter ideology-detection via multi-task multi-relational embedding, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 2258–2268. URL: https://doi.org/10.1145/3394486.3403275. doi:10.1145/3394486.3403275.

[22] S. Feng, H. Wan, N. Wang, J. Li, M. Luo, Twibot-20: A comprehensive twitter bot detection benchmark, Proceedings of the 30th ACM International Conference on Information & Knowledge Management (2021).

[23] M. Ribeiro, P. Calais, Y. Santos, V. Almeida, W. Meira Jr, Characterizing and detecting hateful users on twitter, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 12, 2018.

[24] Z. Zhang, H. Yang, J. Bu, S. Zhou, P. Yu, J. Zhang, M. Ester, C. Wang, Anrl: attributed network representation learning via deep neural networks., in: Ijcai, volume 18, 2018, pp. 3155–3161.

[25] L. Liao, X. He, H. Zhang, T.-S. Chua, Attributed social network embedding, IEEE Transactions on Knowledge and Data Engineering 30 (2018) 2257–2270. doi:10.1109/TKDE.2018.2819980.

[26] W. Zhang, W. Wang, J. Wang, H. Zha, User-guided hierarchical attention network for multi-modal social image popularity prediction, in: Proceedings of the 2018 world wide web conference, 2018, pp. 1277–1286.

[27] J. Ni, Z. Huang, Y. Hu, C. Lin, A two-stage embedding model for recommendation with multimodal auxiliary information, Information Sciences 582 (2022) 22–37.

[28] S. Pan, T. Ding, Social media-based user embedding: A literature review, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 6318–6324. URL: https://doi.org/10.24963/ijcai.2019/881. doi:10.24963/ijcai.2019/881.

[29] T. H. Do, D. M. Nguyen, E. Tsiligianni, B. Cornelis, N. Deligiannis, Twitter user geolocation using deep multiview learning, CoRR abs/1805.04612 (2018). URL: http://arxiv.org/abs/1805.04612. arXiv:1805.04612.

[30] X. Song, Z.-Y. Ming, L. Nie, Y.-L. Zhao, T.-S. Chua, Volunteerism tendency prediction via harvesting multiple social networks, ACM Trans. Inf. Syst. 34 (2016). URL: https://doi.org/10.1145/2832907. doi:10.1145/2832907.

[31] A. Benton, R. Arora, M. Dredze, Learning multiview embeddings of twitter users, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016, pp. 14–19.

[32] L. Yao, C. Mao, Y. Luo, Graph convolutional networks for text classification, in: Proceedings of the AAAI conference on artificial intelligence, volume 33, 2019, pp. 7370–7377.

[33] Y. Dou, K. Shu, C. Xia, P. S. Yu, L. Sun, User preference-aware fake news detection, CoRR abs/2104.12259 (2021). URL: https://arxiv.org/abs/2104.12259. arXiv:2104.12259.

[34] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, M. Welling, Modeling relational data with graph convolutional networks, in: The semantic web: 15th international conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, proceedings 15, Springer, 2018, pp. 593–607.

[35] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, volume 26, Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

[36] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2227–2237. URL: https://aclanthology.org/N18-1202. doi:10.18653/v1/N18-1202.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[38] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[39] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2024.

[40] G. Joshi, R. Walambe, K. Kotecha, A review on explainability in multimodal deep neural nets, IEEE Access 9 (2021) 59800–59821. doi:10.1109/ACCESS.2021.3070212.

[41] A. Mani, W. Hinthorn, N. Yoo, O. Russakovsky, Point and ask: Incorporating pointing into visual question answering, CoRR abs/2011.13681 (2020). URL: https://arxiv.org/abs/2011.13681. arXiv:2011.13681.

[42] Y. Goyal, A. Mohapatra, D. Parikh, D. Batra, Interpreting visual question answering models, CoRR abs/1608.08974 (2016). URL: http://arxiv.org/abs/1608.08974. arXiv:1608.08974.

[43] H. Lee, S. T. Kim, Y. M. Ro, Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis, in: K. Suzuki, M. Reyes, T. Syeda-Mahmood, E. Konukoglu, B. Glocker, R. Wiest, Y. Gur, H. Greenspan, A. Madabhushi (Eds.), Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support, Springer International Publishing, Cham, 2019, pp. 21–29.

[44] A.-H. Karimi, B. Schölkopf, I. Valera, Algorithmic recourse: from counterfactual explanations to interventions, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 353–362. URL: https://doi.org/10.1145/3442188.3445899. doi:10.1145/3442188.3445899.

[45] L. A. Hendricks, R. Hu, T. Darrell, Z. Akata, Generating counterfactual explanations with natural language, CoRR abs/1806.09809 (2018). URL: http://arxiv.org/abs/1806.09809. arXiv:1806.09809.

[46] K. Alipour, A. Ray, X. Lin, J. P. Schulze, Y. Yao, G. T. Burachas, The impact of explanations on ai competency prediction in vqa, in: 2020 IEEE International Conference on Humanized Computing and Communication with Artificial Intelligence (HCCAI), IEEE, 2020, pp. 25–32.

[47] M. Gaur, K. Faldu, A. Sheth, Semantics of the black-box: Can knowledge graphs help make deep learning systems more interpretable and explainable?, IEEE Internet Computing 25 (2021) 51–59. doi:10.1109/MIC.2020.3031769.

[48] D. Blakely, J. Lanchantin, Y. Qi, Time and space complexity of graph convolutional networks, Accessed on: Dec 31 (2021).

[49] A. Vaswani, S. Bengio, E. Brevdo, F. Chollet, A. Gomez, S. Gouws, L. Jones, Ł. Kaiser, et al., Tensor2tensor for neural machine translation, in: Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track), 2018, pp. 193–199.

[50] J. Devlin, M.-W. Chang, K. Lee, K. N. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL: https://arxiv.org/abs/1810.04805.

[51] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei, et al., The flan collection: Designing data and methods for effective instruction tuning, arXiv preprint arXiv:2301.13688 (2023).

[52] E. Alothali, N. Zaki, E. A. Mohamed, H. Alashwal, Detecting social bots on twitter: a literature review, in: 2018 International conference on innovations in information technology (IIT), IEEE, 2018, pp. 175–180.

[53] M. Mazza, S. Cresci, M. Avvenuti, W. Quattrociocchi, M. Tesconi, Rtbust: Exploiting temporal patterns for botnet detection on twitter, in: Proceedings of the 10th ACM Conference on Web Science, WebSci '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 183–192. URL: https://doi.org/10.1145/3292522.3326015. doi:10.1145/3292522.3326015.

[54] Z. Xiao, J. Zhu, Y. Wang, P. Zhou, W. H. Lam, M. A. Porter, Y. Sun, Detecting political biases of named entities and hashtags on twitter, EPJ Data Science 12 (2023) 20.

[55] K. Brown, A. Mondon, Populism, the media, and the mainstreaming of the far right: The guardian coverage of populism as a case study, Politics 41 (2021) 279–295. URL: https://doi.org/10.1177/0263395720955036. doi:10.1177/0263395720955036. arXiv:https://doi.org/10.1177/0263395720955036.