ETS-MM: A Multi-Modal Social Bot Detection Model Based on Enhanced Textual Semantic Representation

Anonymous Author(s)
Submission Id: 1033

Abstract

Social bots are becoming increasingly common in social networks, and their activities affect the security and authenticity of social media platforms. Current state-of-the-art social bot detection methods leverage multimodal approaches that analyze various modalities, such as user metadata, text, and social network relationships. However, these methods may not always extract additional dimensions of semantic feature information that could offer a deeper understanding of users' social patterns. To address this issue, we propose ETS-MM, a multimodal detection framework designed to augment multidimensional information from text and extract the semantic feature representation of user text information. We first analyze the user's tweeting behavior based on topic preference and emotion tendency, integrating them into the textual data. Then, we try to extract enhanced semantic representations that reveal the latent relationship between tweeting behavior and tweet content while identifying potential contextual associations and emotional changes. Additionally, to capture the complex interaction between users, we integrate the user's multimodal information, including metadata, textual features, enhanced semantic features, and social network relationships to propagate and aggregate information across various modalities. Experimental results demonstrate that ETS-MM significantly outperforms existing methods across two widely used social bot detection benchmark datasets, validating its effectiveness and superiority.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; **Neural networks**; • **Information systems** → **Social networks**.

Keywords

Social Bot Detection, Language Model, Large Language Model, Graph Neural Network

ACM Reference Format:

Anonymous Author(s). 2018. ETS-MM: A Multi-Modal Social Bot Detection Model Based on Enhanced Textual Semantic Representation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

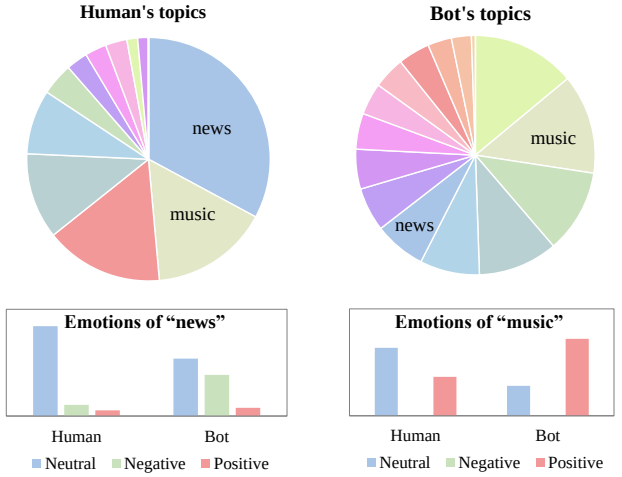


Figure 1: Topic distribution of humans and bots in Cresci15 [9] and emotion distribution within the "news" and "music" topics.

1 Introduction

Social bots often carry specific intentions, such as spreading disinformation [1, 10, 40, 44, 47], manipulating public opinion[7, 20, 21], promoting certain political or commercial agendas[3, 8, 11, 19, 38]. These bots pose a threat to the security of social media platforms. This threat affects users' trust and disrupts the overall information ecosystem[5, 35, 37]. Therefore, effectively detecting social bots has become a critical issue in social media. Researchers have shifted from single-modal to multimodal detection, integrating multiple data sources, such as user metadata, text, and social network relationships. By combining these diverse data sources, multimodal models can more comprehensively and accurately detect social bots[17]. Among these modalities, text, as the main information carrier, is not only the main form of interaction between users but also a key tool for them to influence their behavior and public opinion.

Users' text shows a huge difference between humans and bots regarding emotional expression and social interaction. In terms of social interactions, humans can understand and respond to the emotional states of others and establish deep social relationships. While bots lack an understanding of deep emotions and find it difficult to develop deep social relationships [4]. As shown in Figure 1, humans tend to have in-depth discussions around one or two topics, while bots cover a much wider range. Regarding emotional expression, bots can mimic human emotions in approximate situations. Still, humans have an advantage in the subtle differences in emotional

expression, while bots have difficulty replicating these subtle emotional differences[18]. This leads to differences in the distribution of emotions between bots and humans on specific topics. As shown in Figure 1, the emotion distributions for the "news" topic differ significantly between humans and bots; humans are more likely to express neutral or positive emotions, whereas bots tend to exhibit neutral or negative emotions.

Analyzing users' topics and emotions gives us a more comprehensive and accurate understanding of their behavioral patterns. However, current multimodal detection models tend to ignore the multidimensional semantic information in the text and the connection between this information when processing the text. Two main challenges are included: (1) Some works[6, 28] merely concatenate various users' text information into sequences without delving into the wealth of more dimension information contained within the tweets, such as topics and emotions. (2) Other works[14, 17, 32] directly use pre-trained LMs to obtain the encoding representations of tweets. This method lacks sophisticated mechanisms for extracting enriched semantic features, particularly those related to multidimensional information, which is crucial for distinguishing between bots and humans.

To address the above challenges, we propose a multimodal social bot detection framework incorporating topic preference and emotion tendency to enhance semantic feature representation. Specifically, we use a large language model (LLM) to identify user's most and least frequent topics and emotions to analyze the user's tweeting behavior, incorporating these insights into the tweet data. We then merge each user's description and tweet information (including tweets, topics, and emotions) into textual sequences. Subsequently, we combine each user's metadata with these textual sequences to create hybrid sequences. These hybrid sequences train a language model (LM), capturing enhanced semantic representations that reveal the relationships between topics and emotions within the textual data. Finally, we integrate the user's metadata, description, tweet features, enhanced semantic representation, and social network relationships using a graph neural network (GNN) to accurately determine whether a user is a bot or a human. This work makes the following contributions:

- We analyze the topic preference and emotion tendency in textual data to explore the differences between social bots' and humans' latent social behavior patterns.
- We capture enhanced semantic features that reveal relationships between tweeting behavior and content, identifying contextual associations and emotional changes.
- We integrate enhanced semantic features with multimodal features, enabling effective information propagation and aggregation across modalities, achieving SOTA in social bot detection.

2 Background

2.1 Multi-model Social bot Detection

Early methods for detecting social bots can be broadly categorized into feature-based, text-based, and graph-based. Feature-based approaches identify bots by extracting key attributes from user metadata and applying machine learning classifiers[27, 43, 45]. Text-based methods apply natural language processing techniques (NLP) to analyze and encode users' textual content, primarily focusing on their tweets and descriptions, to identify social bots[15, 27, 42]. Graph-based methods detect bots by constructing social network relationships as graphs and applying GNNs for network analysis[2, 12, 26, 29, 34].

As social bots become increasingly sophisticated in their disguise capabilities, more researchers are employing multimodal approaches for social bot detection. Multimodal methods integrate multiple sources of information, such as user metadata, textual data, and social network information, for more comprehensive detection[24, 36, 46]. Feng et al.[17] enhanced the detection of bots with diverse camouflage behaviors by constructing a heterogeneous graph and exploiting multimodal user semantic and attribute information. Liu et al.[32] synthesized user representations from different prespectives by leveraging multimodal information (metadata, text, network structure). They introduced a modality-specific encoder and a community-aware expert hybrid layer to improve the accuracy and generalization of detection. Feng et al.[14] used a heterogeneous graph information network to learn node representations for graph-based detection of heterogeneity-aware bots, applying relational graph Transformers and semantic attention networks to account for user heterogeneity. Lei et al.[28] combined text and social network information, utilizing a text-graph interaction and semantic coherence approach to assess and counteract bots' evolving behaviors comprehensively. Cai et al.[6] represented each user as a text sequence and performed domain adaptation through an LM, using the model's output as the input features for the GNN. They distilled knowledge from the GNN back into the LM, enhancing the robustness of social bot detection.

Compared to methods that rely on a single source of information, such as user metadata, text, or social network relations, multimodal information provides a more comprehensive analysis.

2.2 Text Processing in Social Bot Detection

The processing of textual modalities is a key aspect of social bot detection. Early methods typically use NLP techniques to extract features from text to identify bots. Kudugunta et al.[27] proposed a long short-term memory(LSTM)-based deep network model to process the tweet and account levels. Wei et al.[42] used a BiLSTM modal to detect Twitter bots for featureless engineering.

Current research has concatenated user data into text sequences and used LMs to obtain the encoded representations[6, 28]. However, these methods fail to extract more dimension information within the tweets. Other approaches to text processing in multimodal methods encoded users' textual information using pretrained LM. Feng et al.[17] obtained the user descriptions and embedded representation of tweets through an LM, averaging the encoded representation of each tweet to get an overall representation of the user's tweets. However, this approach widely used in social bot detection[14, 17, 32] does not include advanced techniques for extracting semantic features, especially those involving multidimensional information.

Although multimodal social bot detection models effectively recognize bots, they often fail to capture textual information. Therefore, we propose extracting the topics and emotions from user tweets through LLM. By leveraging the knowledge of LLM, we can obtain

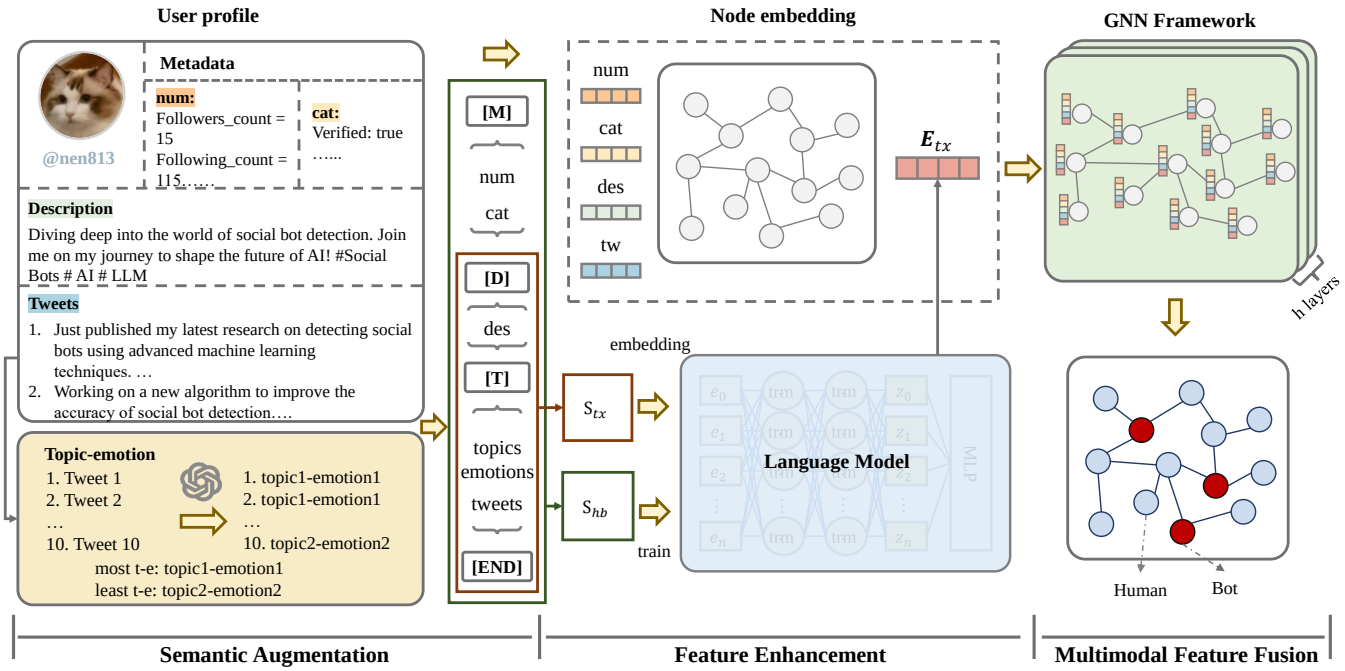


Figure 2: The framework of ETS-MM.

a more comprehensive representation of the tweets, capturing both expressive and contextual features. These topics and emotions are integrated into the user's information, which is then used to train the LM. This process represents the user's textual information (including descriptions, topics, emotions, and tweets) more holistically, ensuring a deeper and more comprehensive understanding of the user's textual behavior.

3 Methodology

We present the ETS-MM framework that consists of three modules: the semantic augmentation module, the feature enhancement module, and the multimodal feature fusion module, as shown in Figure 2. The semantic augmentation module uses an LLM to extract topics and emotions to augment the text data. The feature enhancement module extracts the metadata, textual, and enhanced semantic features. The multimodal feature fusion module employs a GNN to integrate multimodal features, including user metadata, textual information, enhanced semantic features, and social network relationships.

3.1 Semantic Augmentation

To extract more dimensional semantic information, we construct two types of sequences: textual sequences and hybrid sequences.

Textual sequences. The text sequence S_{tx} includes the user's description, tweets:

$$S_{tx} = [Description]\{description\}[Tweet]\{tweet\}[END] \quad (1)$$

where tweets include topics and emotions extracted from LLM, following Chang et al. [8]. We extract their ten longest tweets by GPT-3.5 [30] to identify each tweet's corresponding topic preference and emotion tendency. Appendix A.1 shows the prompts for extracting topics and emotions from GPT-3.5.

- **Topic preference**. We identify the subjects that the user frequently engages within their tweets. This involves categorizing tweets based on content and determining the most and least frequent topics $topic_{most/least}$. The topics are drawn from a list of 16 Twitter categories[1].
- **Emotion tendency**. We further analyze the emotion tendency corresponding to each tweet topic by identifying the most and least frequent emotions $emotion_{most/least}$. The emotions are classified based on Plutchik's three sentiment categories[19].

To further enhance the tweets representation, we incorporate a sequence of topics and emotions into the tweet sequence representation: $\{topic - emotion\}$ = *Most of these tweets are about* $topic_{most}$ *and* $emotion_{most}$, *a few of these are about* $topic_{least}$ *and* $emotion_{least}$. We add this sequence into the sequence of tweets:

$$\{tweet\} = [Tweet]\{topic - emotion\}[SEP]\{tweet_1\}[SEP] \{tweet_2\}[SEP]...[SEP]\{tweet_n\} \quad (2)$$

where $tweet_i$ represents the i-th tweet, and n represents the overall count of user tweets.

Hybrid sequences. The hybrid sequence S_{hb} includes user metadata and text sequence S_{tx}:

$$S_{hb} = [Metadata]\{metadata\}[Description]\{description\} [Tweet]\{tweet\}[END] \quad (3)$$

[1]https://inboundfound.com/twitter-topics-list/

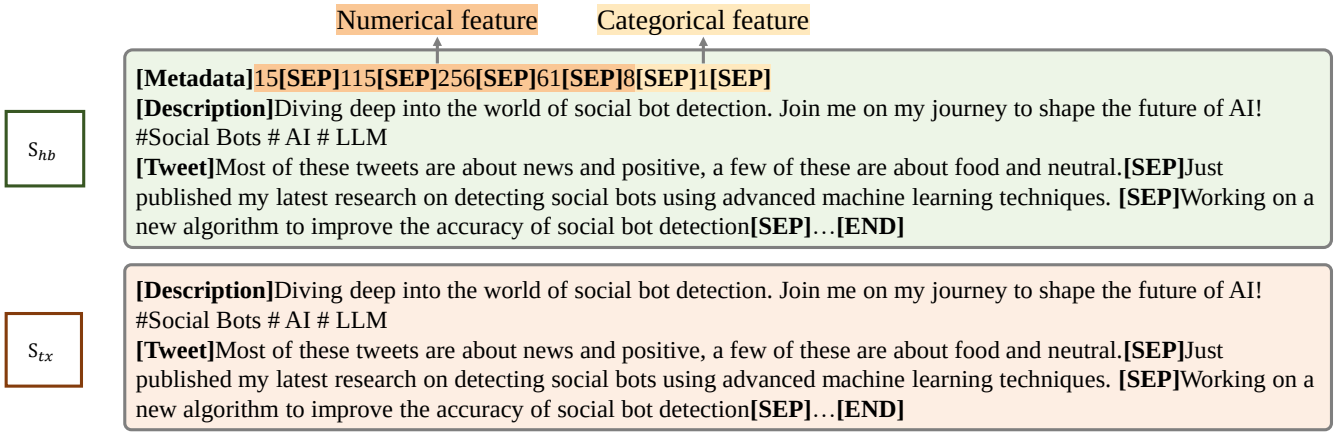


Figure 3: The example of a hybrid sequence and a textual sequence.

where $\{metadata\}$ is the sequence of user metadata:

$$\{metadata\} = [Metadata]\{metadata_1\}[SEP]\{metadata_2\} \\ [SEP]...[SEP]\{metadata_m\} \tag{4}$$

where $metadata_i$ represents numerical and categorical features of users, and m represents the overall count of metadata features. Figure 3 shows the example of a hybrid and textual sequence.

3.2 Feature Enhancement

We extract metadata, textual, and enhanced semantic features for node embedding.

Metadata Feature. Following the feature extraction approach by Feng et al.[17], we process metadata, description, and tweet features. Metadata refers to the user's numerical and categorical features, $num^{(0)}$ and $cat^{(0)}$.

Textual Feature. The description representation $des^{(0)}$ is obtained by a pre-trained LM model. For the tweet representation $twe^{(0)}$, we first use the LM model to obtain the embedding for each tweet and then compute the average embedding of all tweets.

Enhanced Semantic Feature. We train LM with hybrid sequences S_{hb} to capture the enhanced semantic features from the user's augmented semantic data. We design our training framework from the LMs proposed by Cai et al.[6]. Firstly, we use a pre-trained LM model to obtain the user representation z_i and apply mean pooling to the LM output:

$$z_i = \frac{1}{M_i}\sum_{j=1}^{M_i} LM(S_{hbi})_j \tag{5}$$

where S_{hbi} is i - th user's hybrid sequence. $LM(\cdot)$ is a text encoder to obtain the representation of S_{hbi}, and M_i is the number of tokens in S_{hbi}. We obtain a 768-dimensional embedding for the user representation.

We utilize an L layer MLP to perform feature dimensionality reduction and project it into a binary classification space to determine whether the user is human or bot. The final predictions are computed using the $Softmax$ function:

$$\hat{y}_i = Softmax(LeakyReLU(W_{(l)} \cdot z_i^{(l-1)} + b_{(l)})) \tag{6}$$

where $W_{(l)}$ and $b_{(l)}$ are learnable parameters. The model is then optimized using the cross-entropy loss function.

Finally, the enhanced semantic representation E_{tx} within the relationships between topics and emotions is derived by embedding the textual sequences S_{tx} using the trained LM:

$$E_{ts} = LM(X;\theta^*), X \in S_{tx} \tag{7}$$

where θ^* denotes the trained LM parameters.

3.3 Multimodal Feature Fusion

We integrate the user's metadata feature, textual feature, social relationships, and enhanced semantic features using a GNN to classify the user. We construct the social relationships between users as a graph and apply the GNN to learn user representations. The structure of the GNN framework is illustrated in Figure 4. Specifically, we first transform the user's features into node embeddings by passing them through a linear layer followed by ReLU activation (linear_relu layer):

$$\begin{aligned} num_i^{(h)\prime} &= ReLu(Linear(W_{(h)}^0 \cdot num_i^{(h)} + b_{(h)}^0)) \\ cat_i^{(h)\prime} &= ReLu(Linear(W_{(h)}^1 \cdot cat_i^{(h)} + b_{(h)}^1)) \\ des_i^{(h)\prime} &= ReLu(Linear(W_{(h)}^2 \cdot des_i^{(h)} + b_{(h)}^2)) \\ twe_i^{(h)\prime} &= ReLu(Linear(W_{(h)}^3 \cdot twe_i^{(h)} + b_{(h)}^3)) \\ E_{tx_i}^{(h)\prime} &= ReLu(Linear(W_{(h)}^4 \cdot E_{tx_i}^{(h)} + b_{(h)}^4)) \end{aligned} \tag{8}$$

where $W_{(h)}^j$ and $b_{(h)}^j$ are the learnable parameters of the h-layer of the GNN convolution (Gconv), where j is an integer between 0 and 4. i represents the linear layer that processes different user features. Afterward, we concatenate all the five user features to form the final representation $x_i^{(h)}$:

$$x_i^{(h)} = num_i^{(h)\prime} \oplus cat_i^{(h)\prime} \oplus des_i^{(h)\prime} \oplus twe_i^{(h)\prime} \oplus E_{tx_i}^{(h)\prime} \tag{9}$$

The fused features are passed through a linear_relu layer, then combined with the graph-based user relationship features to create

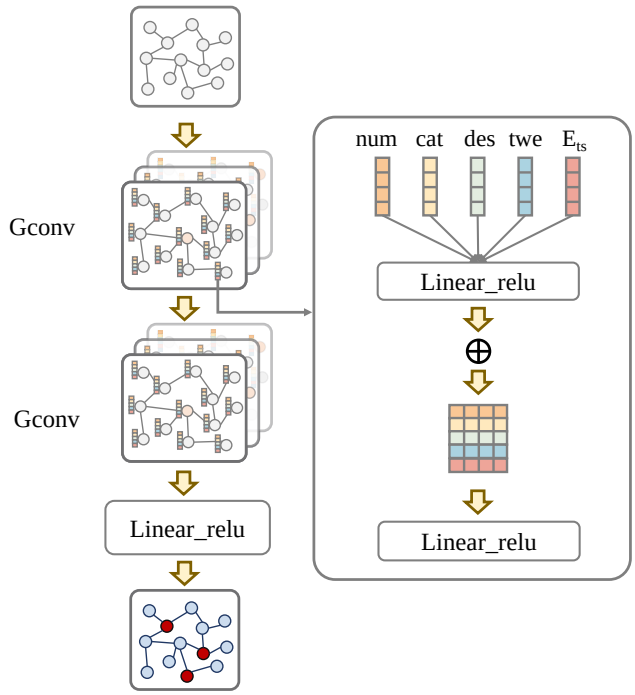


Figure 4: GNN architecture for multimodal Feature Fusion.

a unified representation:

$$a_i^{(h+1)} = \underset{j \in \mathcal{N}(i)}{AGGER}(x_i^h, x_j^h, e_{ij}) \tag{10}$$

$$x_i^{(h+1)} = ReLu(Linear(UPDATE(x_i^h, a_i^h)) \tag{11}$$

where $AGGER(\cdot)$ represents the information aggregated from the neighboring users of user i. $\mathcal{N}(i)$ indicates the set of neighbors of user i. e_{ij} denotes the edge between user i and user j. $UPDATE(\cdot)$ denotes updating the user's representation based on the aggregated information. This updated information is passed through a linear_relu layer to obtain the user representation in layer $h+1$. Finally, the GNN is optimized using a cross-entropy loss function.

4 Experiment

In this section, we aim to answer the following question in this experiment:

- **Q1**: How does the ETS-MM method perform compared to other multimodal-based social bot detection methods?
- **Q2**: Which combinations of LMs and GNNs can significantly enhance the accuracy of social bot detection?
- **Q3**: Can topics and emotions extracted from LLMs improve the effectiveness of social bot detection frameworks?
- **Q4**: Can enhanced semantic features improve the performance of social bot detection tasks?
- **Q5**: How do different modalities influence the model's performance?

4.1 Experiment settings

4.1.1 Datasets. In our experiments, we use two datasets: Cresci15[9] and Twibot20[16]. Cresci15 includes metadatas, descriptions, tweets, and social network information of humans and bots, making it a classic dataset in social bot detection. Twibot20 is larger in scale, providing a broader range of scenarios and diverse user samples. More dataset statistics are outlined in Appendix A.2.

4.1.2 Baselines. Since we utilizes data from various sources, the focus is on comparing it with multimodal-based social bot detection methods. HGT[25] and RGT[14] concentrate on processing heterogeneous graph. SimpleHGN[33] demonstrates that simple isomorphic graph GNNs can achieve high effectiveness when configured appropriately. BotRGCN[17] integrates relational graph convolutional networks with multimodal information to improve detection performance. BIC[28] highlights the deep interaction between text and graphs, along with semantic consistency modeling. BotMoE[32] focuses on multimodal information fusion and incorporates a community-aware mixture of experts. Lastly, LMBot[6] realizes graph-independent bot detection through knowledge distillation of LMs.

4.1.3 Model Parameters. In our model, we use several parameters to achieve optimal performance, as shown in Table 1.

Table 1: Parameter settings of ETS-MM.

Parameter	Value
Batch size in training LM task	32
Epochs of training LM	3
Max length of LM	512
Hidden dimensions of LM	128
Dropout of LM	0.1
Warmup of LM	0.6
Learning rate of LM	1e-5
Weight decay of LM	0.01
Epochs of training GNN	300
Number of GNN convolution layers	2
Dropout of GNN	0.5
Learning rate of GNN	1e-3
Weight decay of GNN	1e-4
Activation of GNN	Leakyrelu
Optimizer of LM and GNN	Adamw

4.2 Performance of ETS-MM(Q1)

Our experiments demonstrate that the ETS-MM model consistently outperforms other state-of-the-art social bot detection methods on Cresci15 and Twibot20, as detailed in Table 2. Specifically, on the Cresci15 dataset, ETS-MM achieves an accuracy of 99.10%, exceeding the second-ranked LMBot. Similarly, on Twibot20, ETS-MM has an accuracy of 90.05%, ahead of BotMoE's. Notably, while LMBot performs well on the smaller Cresci15 dataset, its accuracy drops significantly on the larger Twibot20 dataset, likely due to incomplete semantic feature extraction on the larger dataset. ETS-MM demonstrates consistent and stable performance across multiple

Table 2: Average accuracy and F1-socre of five runs on Cresci15 and Tweibot 20 datasets, with standard deviations in parentheses, best performance bold and second-best underlined.

Method	pre-train LM	train LM	train GNN	Cresci15		Twibot20	
				Accuracy	F1-score	Accuracy	F1-score
HGT[25]			✓	97.45(±0.23)	96.87(±0.43)	86.56(±0.43)	88.75(±0.67)
SimpleHGN[33]			✓	96.84(±0.13)	97.44(±0.74)	87.89(±0.23)	89.56(±0.31)
BotRGCN[17]			✓	96.52(±0.71)	97.30(±0.53)	83.27(±0.57)	85.26(±0.38)
RGT[14]			✓	97.15(±0.32)	97.78(±0.24)	86.57(±0.41)	88.01(±0.41)
BIC[28]		✓	✓	98.35(±0.24)	98.71(±0.18)	87.61(±0.21)	89.13(±0.15)
BotMoE[32]			✓	98.50(±0.00)	98.82(±0.00)	87.76(±0.2)	89.22(±0.3)
LMBot[6]	✓	✓	✓	99.06(±0.30)	99.26(±0.20)	85.25(±0.20)	87.38(±0.20)
ETS-MM		✓	✓	**99.10(±0.08)**	**99.29(±0.06)**	**90.05(±0.19)**	**90.82(±0.18)**

runs, with a standard deviation of only ±0.19% and ±0.18% on Twibot20, highlighting the robustness of the model's improvements. Interestingly, training GNNs using only extracted user metadata and textual features, without additional LM training, led to inferior results. This indicates that LM training effectively captures crucial semantic features in textual information. Enhancing semantic features proves crucial. Almost all models trained with LMs and enriched semantic features perform better, particularly ETS-MM, which reached SOTA on both datasets after semantic enhancement. This confirms the importance of extracting topic preference and emotion tendency in social bot detection.

4.3 LM and GNN Impact on Detection(Q2)

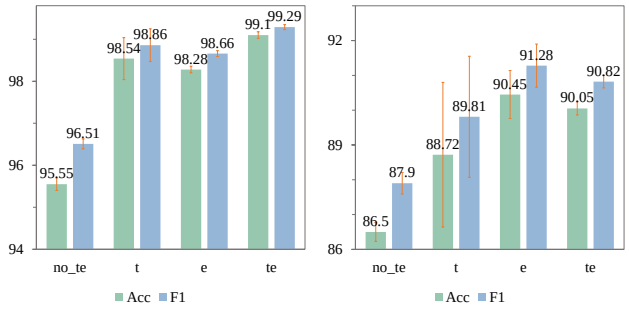
This experiment examines the effect of different combinations of LMs and GNNs on social bot detection performance. We select four LMs–BERT[13], DeBERTa[23], RoBERTa-f[2] and RoBERTa[31]–and four GNNs–GAT[41], HGT[25], SAGE[22], and RGCN[39]. The results indicate that various combinations of LMs and GNNs significantly influence the model's performance, as shown in Table 3.

The Cresci15 dataset shows the most consistent BERT performance and the best RoBERTa results. Since Cresci15 is simpler compared to Twibot20, it may benefit more from the lighter model combinations. In the Twibot20 dataset, RoBERTa and RoBERTa-f outperform BERT and DeBERTa, possibly because RoBERTa generated the original node embeddings, allowing it to analyze complex data structures better. Specifically, in the Cresci15 dataset, the combination of RoBERTa and GAT achieves the best performance, followed by RoBERTa combined with RGCN. This suggests that GAT, with its simplified graph structure aggregation, is more efficient at capturing key features. In contrast, overly complex GNN structures like HGT and RGCN may introduce noise in this scenario. For the Twibot20 dataset, the combination of RoBERTa-f and SAGE performs the best, significantly surpassing other combinations. This implies that SAGE is particularly effective at leveraging the enhanced semantic features extracted by RoBERTa-f to capture the complex relational networks of social bots. Similarly, the combination of RoBERTa and GNN is the most stable, with the model successfully utilizing relational features between users.

[2]https://huggingface.co/yzxjb/roberta-finetuned-20

4.4 Effectiveness of Topics and Emotions(Q3)

To assess the impact of topics and emotions on social bot detection, we design four sets of experiments: without topics and emotions(*no_te*), with only topics(*t*), with only emotions(*e*), and with both topics and emotions(*te*). We analyze the accuracy and F1-score of the four sets.



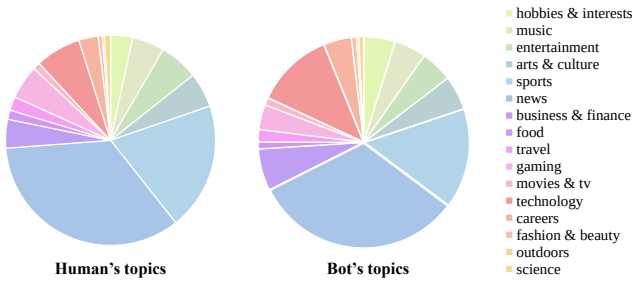
(a) Cresci15 (b) Twibot20

Figure 5: Accuracy, F1-score and standard deviation of ETS-MM across four different scenarios: without topics and emotions(*no_te*), with only topics(*t*), with only emotions(*e*), and with both topics and emotions(*te*).

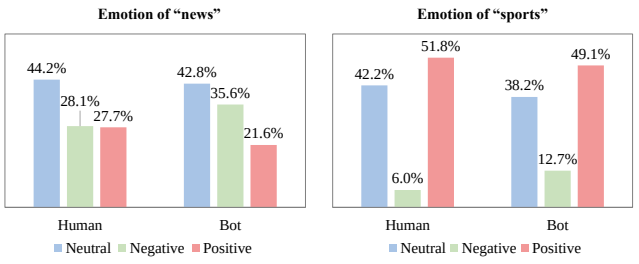
The results of the experiments are shown in Figure 5. In both datasets, adding only topics and emotions improves detection performance. This is because topic information helps capture the potential behavioral patterns of humans and bots, and emotion information allows for the analysis of users' emotional changes and diversity, thus improving the model's performance. By examining the topic distribution in the Cresci15 dataset, as shown in Figure 1, we can clearly distinguish between humans and bots: humans prefer to focus on one or two topics. In contrast, bots have a much broader topic distribution. In Cresci15, the distribution of emotions within topics also varies. For example, in the "news" and "music" topics, there is a significant difference in the emotional distribution between humans and bots. By combining the topics and emotions, richer semantic features are extracted. However, in Twibot20, the model's performance slightly declined. This may be because not all topics and emotions can provide sufficient information for the

Table 3: Accuracy, F1-score and the standard deviations for combinations of four LMs(BERT, DeBERTa, RoBERTa-f and RoBERTa) and four GNNs(GAT, HGT, SAGE and RGCN) on Cresci15 and Twibot20.

Dataset	Cresci15							
Model	BERT		DeBERTa		RoBERTa-f		RoBERTa	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
GAT	98.47(±0.08)	98.80(±0.06)	97.72(±0.16)	98.22(±0.11)	98.06(±0.36)	98.48(±0.28)	**99.14(±0.36)**	**99.32(±0.28)**
HGT	98.47(±0.08)	98.80(±0.06)	97.68(±0.17)	98.19(±0.13)	97.42(±0.16)	98.00(±0.11)	98.5(±0.00)	98.83(±0.00)
SAGE	98.50(±0.00)	98.83(±0.00)	97.57(±0.00)	98.11(±0.00)	97.91(±0.24)	98.37(±0.18)	98.95(±0.31)	99.18(±0.24)
RGCN	98.50(±0.00)	98.83(±0.00)	97.91(±0.08)	98.35(±0.07)	97.50(±0.28)	98.05(±0.21)	99.10(±0.08)	99.29(±0.07)
Dataset	Twibot20							
Model	BERT		DeBERTa		RoBERTa-f		RoBERTa	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
GAT	83.99(±1.81)	85.71(±1.99)	85.21(±0.49)	86.79(±0.43)	87.86(±2.42)	88.97(±2.26)	88.76(±0.34)	89.77(±0.29)
HGT	86.64(±0.34)	87.94(±0.34)	86.97(±0.20)	88.29(±0.11)	90.96(±0.20)	91.63(±0.19)	90.08(±0.19)	90.85(±0.18)
SAGE	87.17(±0.33)	88.58(±0.24)	87.12(±0.30)	88.40(±0.23)	**91.21(±0.28)**	**91.93(±0.27)**	90.16(±0.18)	91.00(±0.13)
RGCN	86.48(±0.32)	87.91(±0.32)	87.13(±0.22)	88.31(±0.23)	91.17(±0.22)	91.85(±0.21)	90.05(±0.19)	90.82(±0.19)



(a) topics in Twibot20



(b) emotions in Twibot20

Figure 6: Topic distribution of humans and bots in Twibot20 and emotion within the "news" and "sports" topics.

detection model. For example, as shown in Figure 6, some emotions (such as "neutral") or highly general topics (such as "news") may not be distinctive enough, thus contributing finitely to the model's detection capability. Twibot20's topic distribution difference between humans and bots is smaller than Cresci15. This phenomenon further demonstrates that bots' ability to disguise themselves continuously improves. Appendix A.3 further enumerates the confusion matrices of the model for four cases: without topics and emotions, only with topics, only with emotions, and with topics and emotions.

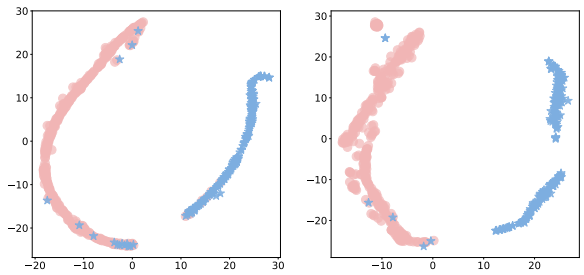
4.5 Contribution of Enhanced Semantic Features(Q4)

This subsection investigates the contribution of enhanced semantic features to the social bot detection framework. We design two different scenarios: without enhanced semantic features (E_{tx}) and with enhanced semantic features (E_{tx}). In our experiments, we compared the feature distributions of the final layer of the ETS-MM model under two conditions across both datasets using t-SNE visualization, as shown in Figure 7.

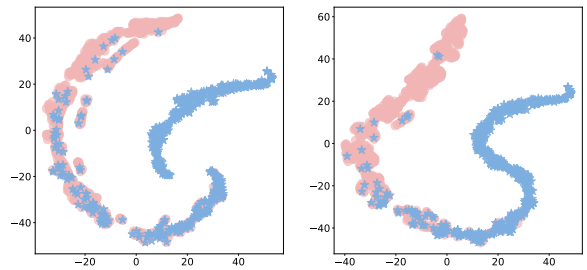
The model displays a much clearer clustering structure when enhanced semantic features were included. This effect is particularly evident in the Cresci15 dataset, where the separation between bots (pink dots) and humans (blue stars) becomes more distinct. Several factors contribute to this improvement: First, enhanced semantic features, such as topics and emotions, allow for a more comprehensive capture of contextual associations and subtle emotional changes within the text. This gives the model a richer semantic background, enabling it to detect behavioral patterns and characteristics specific to social bots, thereby improving detection accuracy. In contrast, without the enhanced feature, the model may rely solely on basic text representations like word vectors or simple sentence embeddings, which are often insufficient for capturing the deeper patterns of user behavior. The model can combine these multidimensional elements by integrating topics and emotions, creating a more comprehensive representation of each user. Appendix A.4 shows the model's performance in two cases: without enhanced semantic features and with enhanced semantic features.

4.6 The Influence of Different Modalities(Q5)

To evaluate the impact of different modalities on the performance of our model, we design four experimental settings: (1) using metadata



(a) ETS-MM without E_{tx} in Cresci15 (b) ETS-MM with E_{tx} in Cresci15



(c) ETS-MM without E_{tx} in Twibot20 (d) ETS-MM with E_{tx} in Twibot20

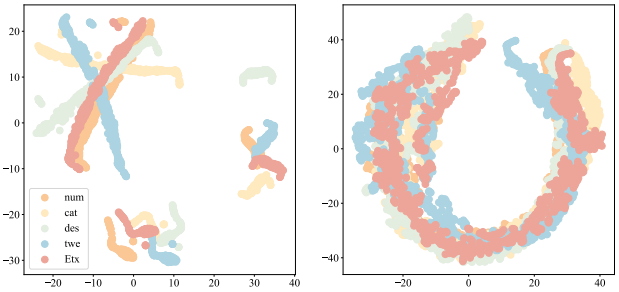
Figure 7: The feature distributions were obtained by applying t-SNE to reduce the dimensionality of output features of the model's final layer across two different scenarios. The first two figures show the results from the Cresci15 dataset, while the last two represent the Twibot20 dataset. Pink dots denote bots, and blue stars indicate humans.

and textual features, (2) using metadata and social network features, (3) using textual and social network features, and (4) using all features (metadata, text, social network), where metadata features include numerical and categorical features, textual features include description representation, tweet representation, and enhanced semantic features.

As shown in Table 4, removing text features led to the most significant drop in performance, indicating the critical role of textual information in detecting social bots. Removing metadata also resulted in a performance decline, but the impact was less severe. Interestingly, removing metadata features had a smaller impact on the Cresci15 dataset, likely due to its simpler data structure. Removing each modality in Twibot20 has a significant effect on the results. In addition, we extracted five types of features for generating user embeddings across two GNN convolution layers and visualized them using t-SNE, as shown in Figure 8. The t-SNE visualization clearly shows that different features contribute to distinct clustering patterns. In both datasets, these modalities integrate with one another. Particularly in the Cresci15 dataset, feature points from different modalities gradually merge into the same category cluster, further validating the effectiveness of our proposed method.

Table 4: Impact of different modalities (metadata, text, and social network) on model performance across Cresci15 and Twibot20 datasets.

Meta	Text	Network	Cresci15	
			Acc	F1
✓	✓		95.96(±0.49)	96.86(±0.36)
✓		✓	92.15(±0.78)	93.99(±0.58)
	✓	✓	98.28(±0.24)	98.65(±0.19)
✓	✓	✓	**99.10(±0.08)**	**99.29(±0.06)**
Meta	Text	Network	Twibot20	
			Acc	F1
✓	✓		83.63(±0.39)	85.34(±0.47)
✓		✓	81.52(±0.22)	85.39(±0.16)
	✓	✓	82.38(±0.56)	83.75(±0.63)
✓	✓	✓	**90.05(±0.19)**	**90.82(±0.18)**



(a) Cresci15 (b) Twibot20

Figure 8: The feature distributions of five features(*num*, *cat*, *des*, *tw* and E_{tx}) across two GNN convolutional layers were visualized using t-SNE.

5 Conclusion

In this study, we propose ETS-MM, a multimodal bot detection model that enhances the semantic representation of user text and addresses the issue of textual modality lacking more dimensional information. Using LLM, we extract topic preference and emotion tendency from tweets, integrating these insights into two custom sequences: textual sequences and hybrid sequences, which combine user metadata and augmented text. These sequences help train the LM and encode enhanced semantic representations. Additionally, GNN is employed to integrate various features, ultimately identifying bots. Our experiments show that the ETS-MM model outperforms existing methods in social bot detection. In the future, we will explore a closer integration of large language models for social bot detection and extract more effective information from user text.

References

[1] Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake News, Disinformation and Misinformation in Social Media: A Review. *Soc. Netw. Anal. Min.* 13, 1 (2023), 30.

[2] Seyed Ali Alhosseini, Raad Bin Tareaf, Pejman Najafi, and Christoph Meinel. 2019. Detect Me If You Can: Spam Bot Detection Using Inductive Representation Learning. In *WWW Companion*. 148–153.

[3] Siva K Balasubramanian, Mustafa Bilgic, Aron Culotta, Libby Hemphill, Anita Nikolich, and Matthew A Shapiro. 2022. Leaders Or Followers? A Temporal Analysis of Tweets from IRA Trolls. In *AAAI*, Vol. 16. 2–11.

[4] Cynthia Breazeal. 2003. Emotion and Sociable Humanoid Robots. *Int. J. Hum. Comput. Stud.* 59, 1-2 (2003), 119–155.

[5] Meng Cai, Han Luo, Xiao Meng, Ying Cui, and Wei Wang. 2023. Network Distribution and Sentiment Interaction: Information Diffusion Mechanisms between Social Bots and Human Users on Social Media. *Inf. Process. Manag.* 60, 2 (2023), 103197.

[6] Zijian Cai, Zhaoxuan Tan, Zhenyu Lei, Zifeng Zhu, Hongrui Wang, Qinghua Zheng, and Minnan Luo. 2024. LMBot: Distilling Graph Knowledge into Language Model for Graph-less Deployment in Twitter Bot Detection. In *WSDM*. 57–66.

[7] Guido Caldarelli, Rocco De Nicola, Fabio Del Vigna, Marinella Petrocchi, and Fabio Saracco. 2020. The Role of Bot Squads in the Political Propaganda on Twitter. *Commun. Phys.* 3, 1 (2020), 81–96.

[8] Ying-Ying Chang, Wei-Yao Wang, and Wen-Chih Peng. 2024. SeGA: Preference-Aware Self-Contrastive Learning with Prompts for Anomalous User Detection on Twitter. In *AAAI*, Vol. 38. 30–37.

[9] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2015. Fame for Sale: Efficient Detection of Fake Twitter Followers. *Decis. Support Syst.* 80 (2015), 56–71.

[10] Stefano Cresci, Roberto Di Pietro, Angelo Spognardi, Maurizio Tesconi, and Marinella Petrocchi. 2023. Demystifying Misconceptions in Social Bots Research. *arXiv preprint arXiv:2303.17251* (2023).

[11] Ashok Deb, Luca Luceri, Adam Badaway, and Emilio Ferrara. 2019. Perils and Challenges of Social Media and Election Manipulation Analysis: The 2018 US Midterms. In *WWW Companion*. 237–247.

[12] Ashkan Dehghan, Kinga Siuta, Agata Skorupka, Akshat Dubey, Andrei Betlen, David Miller, Wei Xu, Bogumił Kamiński, and Paweł Prałat. 2023. Detecting Bots in Social-networks Using Node and Structural Embeddings. *J. Big Data* 10, 1 (2023), 119.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*. 4171–4186.

[14] Shangbin Feng, Zhaoxuan Tan, Rui Li, and Minnan Luo. 2022. Heterogeneity-aware Twitter Bot Detection with Relational Graph Transformers. In *AAAI*, Vol. 36. 3977–3985.

[15] Shangbin Feng, Herun Wan, Ningnan Wang, Jundong Li, and Minnan Luo. 2021. SATAR: A Self-supervised Approach to Twitter Account Representation Learning and its Application in Bot Detection. In *CIKM*. 3808–3817.

[16] Shangbin Feng, Herun Wan, Ningnan Wang, Jundong Li, and Minnan Luo. 2021. Twibot-20: A Comprehensive Twitter Bot Detection Benchmark. In *CIKM*. 4485–4494.

[17] Shangbin Feng, Herun Wan, Ningnan Wang, and Minnan Luo. 2021. BotRGCN: Twitter Bot Detection with Relational Graph Convolutional Networks. In *ASONAM*. 236–239.

[18] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The Rise of Social Bots. *Commun. ACM* 59, 7 (2016), 96–104.

[19] Bilal Ghanem, Davide Buscaldi, and Paolo Rosso. 2019. TexTrolls: Identifying Russian Trolls on Twitter from a Textual Perspective. *arXiv preprint arXiv:1910.01340* (2019).

[20] Henrich R Greve, Hayagreeva Rao, Paul Vicinanza, and Echo Yan Zhou. 2022. Online Conspiracy Groups: Micro-bloggers, Bots, and Coronavirus Conspiracy Talk on Twitter. *Am. Sociol. Rev.* 87, 6 (2022), 919–949.

[21] Anatoliy Gruzd and Philip Mai. 2020. Going Viral: How a Single Tweet Spawned a COVID-19 Conspiracy Theory on Twitter. *Big Data Soc.* 7, 2 (2020), 2053951720938405.

[22] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. *Adv. Neural Inf. Process. Syst.* 30 (2017).

[23] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv preprint arXiv:2006.03654* (2020).

[24] Zheng Hu, Shi-Min Cai, Jun Wang, and Tao Zhou. 2023. Collaborative Recommendation Model Based on Multi-modal Multi-view Attention Network: Movie and Literature Cases. *Appl. Soft Comput.* 144 (2023), 110518.

[25] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous Graph Transformer. In *WWW*. 2704–2710.

[26] Zheng Hu, Satoshi Nakagawa, Liang Luo, Yu Gu, and Fuji Ren. 2023. Celebrity-aware Graph Contrastive Learning Framework for Social Recommendation. In *CIKM*. 793–802.

[27] Sneha Kudugunta and Emilio Ferrara. 2018. Deep Neural Networks for Bot Detection. *Inf. Sci.* 467 (2018), 312–322.

[28] Zhenyu Lei, Herun Wan, Wenqian Zhang, Shangbin Feng, Zilong Chen, Jundong Li, Qinghua Zheng, and Minnan Luo. 2022. BIC: Twitter Bot Detection with Text-graph Interaction and Semantic Consistency. *arXiv preprint arXiv:2208.08320* (2022).

[29] Shudong Li, Chuanyu Zhao, Qing Li, Jiuming Huang, Dawei Zhao, and Peican Zhu. 2023. BotFinder: A Novel Framework for Social Bots Detection in Online Social Networks based on Graph Embedding and Community Detection. *World Wide Web* 26, 4 (2023), 1793–1809.

[30] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Lin Zhao, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. 2023. Summary of ChatGPT-Related research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852* (2023).

[31] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* (2019).

[32] Yuhan Liu, Zhaoxuan Tan, Heng Wang, Shangbin Feng, Qinghua Zheng, and Minnan Luo. 2023. BotMOE: Twitter Bot Detection with Community-aware Mixtures of Modal-specific Experts. In *SIGIR*. 485–495.

[33] Qingsong Lv, Ming Ding, Qiang Liu, Yuxiang Chen, Wenzheng Feng, Siming He, Chang Zhou, Jianguo Jiang, Yuxiao Dong, and Jie Tang. 2021. Are We Really Making Much Progress? Revisiting, Benchmarking and Refining Heterogeneous Graph Neural Networks. In *SIGKDD*. 1150–1160.

[34] Thomas Magelinski, David Beskow, and Kathleen M Carley. 2020. Graph-HIST: Graph Classification from Latent Feature Histograms with Application to Bot Detection. In *AAAI*, Vol. 34. 5134–5141.

[35] Spencer McKay and Chris Tenove. 2021. Disinformation as a Threat to Deliberative Democracy. *Polit. Res. Q.* 74, 3 (2021), 703–717.

[36] Lynnette Hui Xian Ng and Kathleen M Carley. 2023. Botbuster: Multi-platform Bot Detection using a Mixture of Experts. In *AAAI*, Vol. 17. 686–697.

[37] Javier Pastor-Galindo, Félix Gómez Mármol, and Gregorio Martínez Pérez. 2022. Profiling Users and Bots in Twitter through Social Media Analysis. *Inf. Sci.* 613 (2022), 161–183.

[38] Sippo Rossi, Matti Rossi, Bikesh Raj Upreti, and Yong Liu. 2020. Detecting Political Bots on Twitter During the 2019 Finnish Parliamentary Election. In *HICSS*. 2430–2439.

[39] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling Relational Data with Graph Convolutional Networks. In *Semant. Web*. 593–607.

[40] Kate Starbird. 2019. Disinformation's Spread: Bots, Trolls and all of us. *Nature* 571, 7766 (2019), 449–450.

[41] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph Attention Networks. *arXiv preprint arXiv:1710.10903* (2017).

[42] Feng Wei and Uyen Trang Nguyen. 2019. Twitter Bot Detection Using Bidirectional Long Short-term Memory Neural Networks and Word Embeddings. In *TPS-ISA*. 101–109.

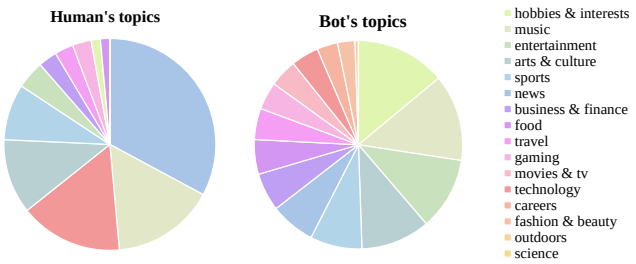
[43] Jun Wu, Xuesong Ye, and Chengjie Mou. 2023. BotShape: A Novel Social Bots Detection Approach via Behavioral Patterns. *arXiv preprint arXiv:2303.10214* (2023).

[44] Kai-Cheng Yang, Francesco Pierri, Pik-Mai Hui, David Axelrod, Christopher Torres-Lugo, John Bryden, and Filippo Menczer. 2021. The COVID-19 Infodemic: Twitter versus Facebook. *Big Data Soc.* 8, 1 (2021), 20539517211013861.

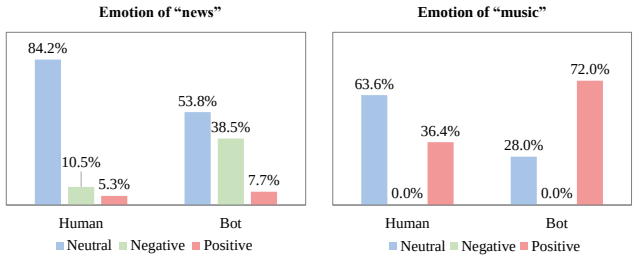
[45] Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. 2020. Scalable and Generalizable Social Bot Detection through Data Selection. In *AAAI*, Vol. 34. 1096–1103.

[46] Yingguang Yang, Renyu Yang, Hao Peng, Yangyang Li, Tong Li, Yong Liao, and Pengyuan Zhou. 2023. FedACK: Federated Adversarial Contrastive Knowledge Distillation for Cross-lingual and Cross-model Social Bot Detection. In *WWW*. 1314–1323.

[47] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2019. Disinformation Warfare: Understanding State-sponsored Trolls on Twitter and Their Influence on the Web. In *WWW Companion*. 218–226.



(a) topics in Cresci15



(b) emotions in Cresci15

Figure 9: Topic distribution of humans and bots in Cresci15 and emotion distribution within "news" and "music" topics.

A APPENDIX

A.1 Prompts for Extracting Topics and Emotions from ChatGPT

We reference [8] to design the instruction prompts we extract for topics and emotions for the construction of the textual dataset in Section 3.1, as shown in Table 5.

Table 5: The instruction prompt and an example of the output of GPT-3.5-Turbo

Instruction prompt	Example output
Please classify each tweet into the topics and corresponding emotions. The available topics are arts & culture, business & finance, careers, entertainment, fashion & beauty, food, gaming, hobbies & interests, movies & TV, music, news, outdoors, science, sports, technology, and travel. The emotions to consider are positive, negative, and neutral. Please provide the classification for each post in the format 'topic - emotion'. If you are not sure about the 'topic' corresponding to this tweet, classify the 'topic' as none. Limit the response to less than 100 words and use lowercase.	1: news - positive 2: news - negative 3: news - positive 4: news - neutral 5: news - positive 6: news - positive 7: news - positive 8: news - negative 9: news - neutral 10: news - neutral

A.2 Statistics of Datasets

Table 6 shows the number of training sets, validation sets, test sets, and overall in the Cresci15 and Twibot20 datasets.

Table 6: Statistics of Datasets.

Datasets	train	dev	test	total
Cresci15	3708	1058	535	5301
Twibot20	8278	2365	1183	11826

A.3 Confusion Matrices of ETS-MM with Topics and Emotions

Figure 10 shows the confusion matrices of the model for the four cases: without topics and emotions, with only topics, with only emotions, and with both topics and emotions in Section 4.4. On the Cresci15 dataset, adding topics and emotions makes the model able to distinguish humans more clearly, especially topics. This also shows that topics are important in distinguishing bots from humans. On the Twibot20 dataset, the model's accuracy for predicting humans and bots increases. This demonstrates that emotions are also beneficial for improving social bot detection.

A.4 Results of Enhanced Semantic Features

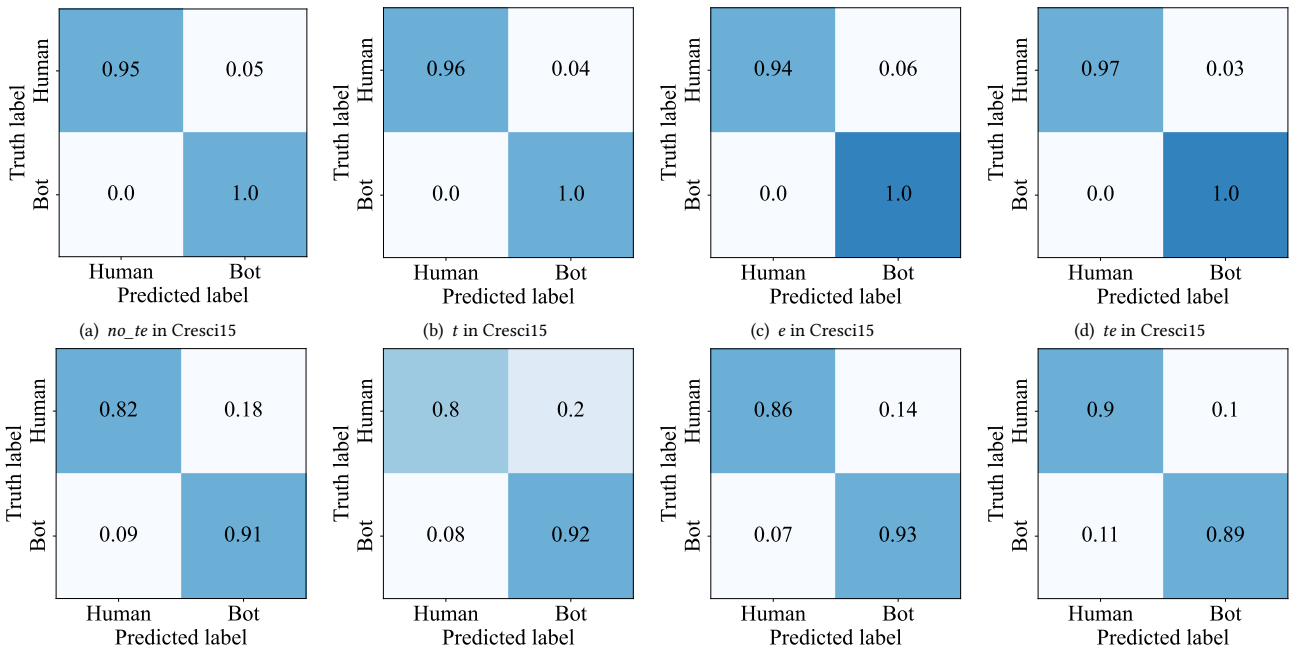
Table 7 illustrates the performance changes in two different scenarios: without enhanced semantic features E_{tx} and with enhanced semantic features E_{tx}, in Section 4.5. The performance changes are illustrated in Table 7. In the Cresci15 and Twibot20 datasets, after removing enhanced semantic features, the model's accuracy decreased by 3.55% and 2.78%, while the F1-score dropped by 3.55% and 2.92%.

Table 7: Accuracy, F1-score, and standard deviation of ETS-MM across two different scenarios: without and with enhanced semantic features.

Methods	Cresci15	
	Accuracy	F1-score
without E_{tx}	95.55(±0.15)	96.51(±0.12)
with E_{tx}	**99.10(±0.08)**	**99.29(±0.06)**
Methods	**Twibot20**	
	Accuracy	F1-score
without E_{tx}	86.50(±0.27)	87.90(±0.31)
with E_{tx}	**90.05(±0.19)**	**90.82(±0.18)**

A.5 Topics and Emotions Distribution of Cresci15

Figure 9 shows the distribution of topics and emotions with data labels in Cresci15,



(a) *no_te* in Cresci15 (b) *t* in Cresci15 (c) *e* in Cresci15 (d) *te* in Cresci15

(e) *no_te* in Twibot20 (f) *t* in Twibot20 (g) *e* in Twibot20 (h) *te* in Twibot20

Figure 10: Confusion matrices for ETS-MM across four scenarios(*no_te*, *t*, *e* and *te*). The first four matrices correspond to the Cresci15 dataset, while the four matrices correspond to the Twibot20 dataset.