A user perspective to quantify controversy on Twitter using Graph Neural Networks

Anonymous ACL submission

Abstract

This paper investigates the quantification of controversy in online discussions, focusing on social media platforms, notably Twitter. Emphasizing the prevalence of echo chambers, where users are exposed to opinions aligned with their own, we propose a novel approach leveraging Large Language Models (LLM) and Graph Neural Networks (GNN). Our methodology integrates both structural and textual information in social networks to provide a nuanced understanding of controversy. Contributions include a theoretical model for quantifying controversy based on the expected probability of user participation in controversial topics. We introduce an empirical estimation method using a GNN-based model. Unlike existing approaches focused on structural polarity, our model captures the rich textual content. Empirical evaluations on Twitter topics demonstrate the effectiveness of our methodology, outperforming variant methods using only textual or structural information, as well as state-of-the-arts methods. In conclusion, we introduce an innovative approach to controversy quantification, emphasizing user participation within social networks.

1 Introduction

Internet and social media have today replaced in many ways discussion and debates in society. People interacting on local or international topics come from different communities, countries, ethnicities, etc. By discussing their point-of-views, sharing content, or embracing some other user arguments, they participate in public debates and potentially controversial topics. Controversy represents a prolonged public disagreement on a topic or event (Hessel and Lee, 2019). With the increased use of social media, controversial topics are widely discussed, especially on social media like Twitter, where it is easy to share about specific events (hashtags) or to endorse someone (retweet). The presence of the phenomenon of echo-chambers [1] indirectly helps to build strong communities around a single topic/event, such as the pros and cons of mandatory vaccination for the COVID-19 vaccine, only taking into account arguments from their perspective.

The rise of deep neural networks enables us to investigate more deeply large amounts of data, such as large graphs and texts. Large Language Models (LLM) based on the attention mechanism and the transformer's architecture, such as BERT, enable us to represent texts regarding the context. Graph Neural Networks (GNN) are commonly used to work with deep neural networks on unstructured data, for different tasks like node or graph classification.

In this paper, we quantify controversy on topics from the perspective of user polarization around communities with diverse viewpoints and opinions within social networks.

Contributions. Most approaches focus on polarity, ignoring the information contained in the texts published by users. We propose a user-based approach to quantify controversy, using both structural and textual information through GNN layers, resulting from the three contributions:

- **Controversy quantification**. We propose a theoretical model to quantify the controversy according to the users. This score is based on the expected probability of a user participating in a controversial topic.
- **Empirical estimation of the quantification score**. We propose a method for estimating this score empirically. We estimate the conditional probabilities of users participating in a controversial topic. We show that minimizing the loss function of this estimator is equivalent to optimizing our score. To estimate this

[1] Environment where a person only encounters information or opinions that reinforce their own opinions.

probability for each user, we propose a model based on GNNs, exploiting both textual and structural information of the graph.

- **Results and analysis**. We establish an evaluation protocol to compare our different models. We compare our model with some of their variants using only textual or structural information and show that it achieves the best performance. Then, by comparing the quantification scores with the literature on topics from Twitter, we show that our approach performs better.

The paper is organized as follows: after a state-of-the-art in Section 2, Section 3 presents the theoretical modeling of our Twitter controversy quantification score and a method for empirically estimating this score. Section 4 presents the experimental setup and Section 5 sets out our experiments and the results, before closing in section 6.

2 Related Work

Controversy work is mainly concerned with detection and quantification. Controversy detection on social media aims to guess if a topic is controversial or not. Most of the proposed approaches exploit a user interactions graph (i.e. retweet graph) (Garimella et al., 2018; Mendoza et al., 2020; Hessel and Lee, 2019). Such a graph is partitioned into two disjoint classes C and $\bar{C}$, each one representing users supporting the same opinion on a given topic. Well-separated classes characterize controversial topics. Some recent approaches consider controversy detection as a graph classification problem (Zhong et al., 2020). They exploit Graph Neural Networks and Natural Language Processing techniques to classify the whole graph as controversial or not.

Controversy quantification, which is our purpose in this paper, aims to measure to what extent a topic is controversial. Different methods have been suggested. Random Walk Controversy (RWC) is introduced in (Garimella et al., 2018) and works on classes C and $\bar{C}$. It aims to capture how likely a random user of a class is to be exposed to the content of the most connected users of the opposite class. Some adaptations of the RWC metric were proposed in (Emamgholizadeh et al., 2020) and (Darwish, 2019) to take into account useful information that could be present in user nodes (i.e. influencer or not, used hashtags, etc.).

Considering that a force-directed embedding technique (Jacomy et al., 2014) fosters a clear separation of partitions of a graph (modularity), the two-dimensional embedding of user nodes of classes C and $\bar{C}$ are exploited to define the Embedding Controversy score (EC) (Garimella et al., 2018). EC is based on the average embedded distance among pairs of user nodes in C (respectively $\bar{C}$), and the average embedded distance among pairs of nodes across C and $\bar{C}$. Controversial (respectively non-controversial) topics tend to have an EC score close to 1 (respectively 0). (Guerra et al., 2013) consider that the modularity polarization metric is not necessarily sufficient, since non-polarized graphs may also be divided into two disjoints classes. A community boundary-based polarization metric is proposed, which characterizes polarized communities by a low concentration of high-degree nodes along the boundary.

Inspired by the electric dipole moment, (Morales et al., 2015) consider that perfect controversy can be characterized by the fact that the classes C and $\bar{C}$ are of the same size and with opposite opinions. A dipole controversy measure is then proposed, it defines the controversy level as a function of the difference in size between C and $\bar{C}$, and the distance between the opinions of the two classes (i.e. the gravity centers). A model is defined to estimate the opinion distributions of users of both classes. (Zarate and Feuerstein, 2020) proposed a vocabulary-based controversy measure that adapts the dipole measure by replacing the opinions of both classes with their respective vocabularies used by users.

These controversy quantification methods work on graph partitions to define metrics. In this paper, we quantify controversy by focusing on users instead of communities. We base our quantification method on the Probabilistic Theory of Pattern Recognition (Xu et al., 2019). We consider that a perfect controversial (respectively non-controversial) topic corresponds to a graph for which we can predict without error that any of its subgraphs is controversial (respectively non-controversial). The controversy level is then quantified as the error when predicting the controversy label of a selected subgraph centered on a random user. To the best of our knowledge, our work is the first work that exploits conditional probability of subgraphs belonging to controversial topics for the need for controversy quantification.

3 Method

3.1 Controversy quantification

We propose a theoretical score for quantifying controversy based on user subgraphs. Given G a random graph and L a random label, we denote their realizations by g and l. The graph and the labels will be indexed as g_i and l_i when necessary. Let $\mathbb{P}$ be the true unknown joint distribution of G, L ($(G, L) \sim \mathbb{P}$) which can be decomposed into μ, the marginal law of G, and η the conditional probability of L given the observation of a graph (Devroye et al., 2013):

$$\eta(g) = \mathbb{P}(L = 1|G = g) \tag{1}$$

The equation 1 corresponds to the true probability that a graph is controversial ($L = 1$) conditionally on the observation of the graph ($G = g$). We sometimes call η the *posterior*, as opposed to the *prior* which represents the frequency of labels. Again, η is an unknown quantity that we generally aim at estimating when minimizing a cross-entropy in deep learning.

Our controversy quantification score is related to the features used in the user graph. Thus, if any part of the graph necessarily implies controversy, then the quantification score should be high. On the other hand, if only certain parts of the graph indicate that the subject is controversial, then the score should be low. Finally, if the graph is not inherently linked to a controversial subject, the corresponding score is expected to approximate zero.

Let $g_u^{(k)} \subseteq g$ be a subgraph of g centered on user u and including up to k levels of neighbors. The true conditional probability that the subgraph $g_u^{(k)}$ is associated with a controversial subject is given by equation 2.

$$\eta\left(g_u^{(k)}\right) = \mathbb{P}\left(L = 1|G = g_u^{(k)}\right) \tag{2}$$

Hence, $\eta\left(g_u^{(0)}\right)$ represents the probability that the content published by user u is associated with a controversial subject, independently of any interaction. In contrast, $\eta\left(g_u^{(\infty)}\right)$ represents the probability that the subject is controversial when the entire graph is analyzed. We expect the latter to be close to 1 if controversial, or 0 if not controversial, even if the absence of contextual elements can sometimes limit the certainty of the prediction. Note that the latter remains a quantity that depends on η, the unknown *posterior*.

We now define in equation 3 the quantification score CQS ("Controversy Quantification Score"):

$$CQS(g, k) = \mathbb{E}_{u\sim\mathbb{U}(g)}\left[\eta\left(g_u^{(k)}\right)\right] \tag{3}$$

This score corresponds to the expectation that a user chosen uniformly in the graph will be associated with a controversial topic, looking only at k neighborhood levels. The only unknown quantity in equation 3 is the true conditional probability η, which must be estimated.

3.2 Consistent loss functions with CQS

In this section, we show that loss functions such as *cross-entropy* are consistent with our CQS score estimation. Minimizing a consistent loss function minimizes the estimation error of CQS.

We define $\ell : [0, 1] \to \mathbb{R}^+$ as a binary loss function. The argument is the probability estimation of the true label. η is the true conditional probability and $\hat{\eta}$ its estimation. The risk associated with ℓ is given in 4.

$$\mathcal{L}_\ell(\hat{\eta}, \eta) = \eta\ell\left(\hat{\eta}\right) + (1 - \eta)\ell\left(1 - \hat{\eta}\right) \tag{4}$$

If $\ell(s) = -\log(s)$, then we obtain the *binary cross entropy*.

Definition 1. ***(Strictly) Proper loss*** **(Lorieul, 2020)** *A loss $\ell : [0, 1] \to \mathbb{R}^+$ is considered proper if its infimum (largest minorant) is reached by η:*

$$\mathcal{L}_\ell(\eta, \eta) = \inf_{s\in[0,1]} \mathcal{L}_\ell(s, \eta)$$

And strictly proper if η is the unique minimizer.

Definition 2. ***μ-strongly proper loss*** **(Lorieul, 2020)** *a loss $\ell : \{0, 1\} \times [0, 1] \to \mathbb{R}^+$ is μ-strongly proper if:*

$$\mathcal{L}_\ell(\hat{\eta}, \eta) - \mathcal{L}_\ell(\eta, \eta) \geq \frac{\mu}{2}|\hat{\eta} - \eta|_1^2$$

The regret is the gap between our risk and its optimal value:

$$\mathrm{Reg}_\ell(\hat{\eta}; x) = \mathcal{L}_\ell(\hat{\eta}(x), \eta(x)) - \mathcal{L}_\ell(\eta(x), \eta(x))$$

The following proposition shows that a *strongly proper* loss function is consistent with our score if the graphs are generated according to a specific procedure shown in 5:

$$\begin{aligned} G &\sim \mu \\ u &\sim \mathbb{U}(G) \end{aligned} \tag{5}$$

Graphs $g_u^{(k)}$ are built from this procedure, μ being the marginal law of graphs.

Proposition 1. *Any strongly proper loss function ℓ is consistent with CQS if it is minimized for graph generation.*

$$\mathbb{E}_G\left[\mathbb{E}_{u\sim\mathbb{U}(G)}\left[\mathrm{Reg}_\ell\left(\hat{\eta}; g_u^{(k)}\right)\right]\right] \to 0 \Rightarrow$$

$$\mathbb{E}_G\left[\left|\widehat{CQS}(G,k) - CQS(G,k)\right|\right] \to 0$$

If the former term converges to 0, then so does the error in estimating our score. We can generalize the previous theorem by taking a random k according to a probability distribution. If the regret tends towards 0 for a random k, then so does the estimation error. We define μ_1 as the marginal probability distribution of k and μ_2 as the marginal probability distribution of the graphs:

$$k \sim \mu_1$$
$$g \sim \mu_2$$
$$u \sim U(g)$$

We present in section 3.3 a method to empirically estimate this quantification score.

3.3 Empirical estimation of the conditional probabilities

After defining the score $CQS(g,k)$ theoretically in equation 3, we establish a method for estimating this score empirically. To this end, we compute the conditional probabilities $\eta(g_u^{(k)})$ from the k-level subgraphs of each user u. Figure 1 shows the various stages in the process of computing this probability η and estimating our controversy quantification score.

Firstly, the user retweet graph is created from the tweets retrieved related to the topics. The graph is then fed into our GNN-based model, to predict the user's participation in a controversial topic. Secondly, this model, combining both the structural properties of the retweet graph and the textual information published by users in their original tweets, is presented.

3.3.1 Graph building

We consider retweets as user endorsement. From all tweets and retweets belonging to a topic, a user retweet graph G is created, representing the topic discussion on Twitter. Nodes represent users, and two users u_i and u_j are related by an edge if one has retweeted the other at least once. The graph is undirected. Each user is represented by his tweets on the subject concerned. As social networks are known for their low density, after creating the graph, we only keep users and edges from the biggest connected component, as we want to propagate information through the graph.

More formally, a topic t is represented as a graph $G = (\mathcal{U}, \mathcal{E}, X)$ where $\mathcal{U} = \{u_1, u_2, ..., u_n\}$ denotes the user nodes and $\mathcal{E} = \{(u_i, u_j)\}_{1\leq i,j\leq n}$ denotes the edges of the graph. A node represents a user, and an edge between two nodes exists if there is at least one interaction between the corresponding users. The set X represents node features, represented by tweets for each user.

3.3.2 Predicting user participation in a controversial topic

Figure 1 step a) Each tweet is represented by a vector using the BERT language model (Devlin et al., 2019). The output of the last layer of the BERT model is used as the representation vector for each tweet. Tweets representations are then aggregated by user, as indicated in figure 1 by the AGG block. Tweets representations are refined as the model is trained, implying the refining of the textual representation of users after the aggregation block. Users who have not posted any tweets are assigned an empty tweet by default, with its corresponding vector representation. User input representations are gathered in the matrix $X \in \mathbb{R}^{n\times d}$, with n the number of nodes and d the dimension of the vectors.

Figure 1 step b) From the user-embedded representation of its tweets, the model learns new node representations from the structural representation of the graph, using graph convolutional networks with multiple layers. These convolutional layers enable the integration of node features and local neighborhood information, and effectively learn expressive node representations that capture both local and global graph structure, enabling downstream tasks such as node classification. Two different GNN approaches are tested, based on the spatial theory, with different characteristics.

1. **Inductive representation learning on large graphs.** GRAPHSAGE (Hamilton et al., 2017) uses neighborhood sampling and aggregation to generate informative node embeddings from local neighbors. Compared to the classic spatial theory of convolutional layers (Xu et al., 2019), GRAPHSAGE uses node

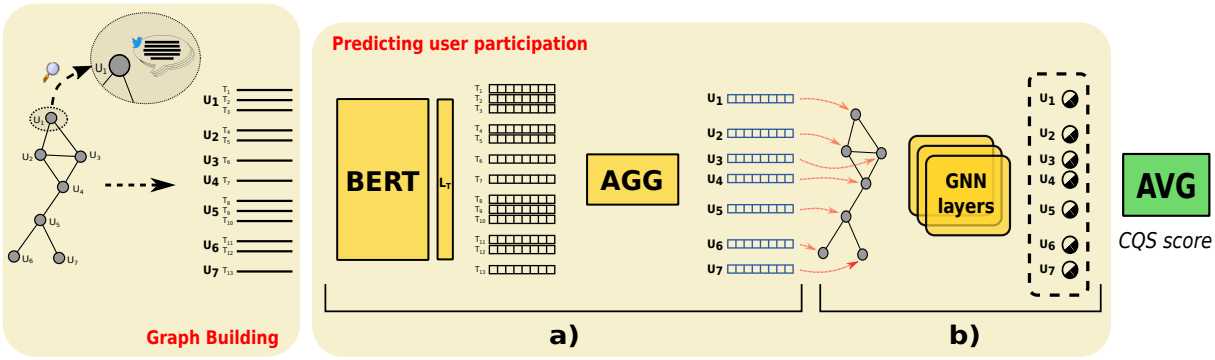

Figure 1: Overview of the different steps in our approach to quantifying controversy. The textual representation of users is refined by learning new representations in the model. The output of the last GNN layer is an estimation of the probability of a user participating in a controversial topic.

sampling to keep the computational footprint of each batch fixed. Each layer l follows the equation 6 to create new node representations.

$$h_u^{(l)} = \sigma\Big(W^{(l)} \cdot CONC\Big(h_u^{(l-1)}, AGGR(\{h_v^{(l-1)}, \forall v \in \mathcal{N}_{(u)}\})\Big)\Big) \quad (6)$$

GRAPHSAGE first performs neighborhood sampling. Given a target node u, the method randomly samples a fixed-size set of its neighbors, gathered in set $\mathcal{N}_{(u)}$. This sampling process is repeated for each node in the graph, allowing efficient computation on large-scale graphs. $W^{(l)}$ is the weight matrix optimized at layer l. Embeddings of the sampled neighbor nodes are then aggregated ($AGGR$) using the mean-aggregator to have an average embedded representation of neighbors. Finally, this average embedded representation of neighbors is concatenated with the target node v representation at the layer $l-1$, and fed to a classic perceptron combined with an activation function σ. The output represents the embedding of the target user u at the layer l.

2. **Attention for node representation.** GAT (Velickovic et al., 2018) creates new node representations in graph-structured data by leveraging attention mechanisms. Self-attention mechanism is introduced to assign different attention weights between users in the graph. By learning the importance of each user's neighbors, GAT focuses on the most relevant users during representation learning. weights are computed using a shared attention mechanism across all nodes. For each user u, the model learns attention coefficients $a_{u,v} = ATT(W_{att}^{(l)} h_u^{(l-1)}, W_{att}^{(l)} h_v^{(l-1)})$ between the target input user u representation ($h_u^{(l-1)}$) and every neighbor (including himself) $v \in \tilde{N}_{(v)}$ representation (($h_v^{(l-1)}$) from the previous layer $l-1$. The function ATT represents a classical single-layer neural network and $W_{att}^{(l)}$ the attention weight matrix at layer l. These weights are then normalized using a $softmax$ function, shown by equation 7.

$$\alpha_{uv} = softmax(a_{uv}) = \frac{\exp(a_{uv})}{\sum_{w \in \tilde{\mathcal{N}}_{(u)}} \exp(e_{uw})} \quad (7)$$

These attention coefficients are then used in the propagation function of the convolution layer, as shown by equation 8. GAT employs multiple attention heads to capture different aspects of the graph structure and interactions. Each attention head k independently computes attention coefficients and generates a weighted sum of neighbor features. The outputs of attention heads are concatenated to generate the embedding of the target user u at the layer l, as shown in 8.

$$h_u^l = \|_{k=1}^{K} \sigma\left(\sum_{v \in \tilde{\mathcal{N}}_{(u)}} \alpha_{uv}^k \mathbf{W}^k h_v^{(l-1)}\right) \quad (8)$$

These two methods present the advantage of being inductive, with predicting capabilities of nodes from unseen graphs. As spatial methods, only the local neighborhoods of nodes are needed to com-

pute new representations and not the full graph, reducing the computational and memory costs. The last graph convolutional layer will be used in both models to classify the user node as belonging to a controversial topic or not, using $softmax$ as the activation function. Note that for the first layer $l = 0$, the input features of users correspond to the textual representation X of users.

With our $\hat{\eta}$ estimator defined, we estimate for each user u the probability of their participation in a controversial topic and empirically estimate the quantification score by averaging probabilities.

4 Experiments

In this section, we present the experiments carried out to evaluate our approach and the dataset used.

4.1 Dataset

We use the Twitter dataset provided by Zarate and Feuerstein (2020), retrieved from the Twitter API [2], composed of 30 topics from 2019 to 2020.

Fifteen of these topics are controversial and fifteen are not non-controversial. Each topic has been manually labeled using multiple sources such as mainstream media. Non-controversial topics are represented by soft news such as entertainment or dramatic events with no controversy, whereas controversial topics are focused on political events (especially election and justice cases). To retrieve multiple controversial datasets, some of them represent the same event, but at different times. Each topic contains tweets being retrieved from hashtags or keywords from the corresponding event [3]. We only keep original tweets at least retweeted once, and users who have been tweeting or retweeting at least once (involved in the debate).

4.2 Evaluation protocol

To train and test our model, the dataset is divided into two balanced sets (train and test). The training set $\mathcal{G}_{\text{train}}$ contains 20 subjects (10 from each label), and the test set $\mathcal{G}_{\text{test}}$ contains 10 subjects (5 from each label). To avoid biasing our analysis and overfitting the model, we ensure that the controversial subjects are separated by time period are part of the same (training or test) set.

As presented in section 3.1, we define a metric to test and compare our approaches. We create several test subsets according to the value of k. For one subset, we randomly take 5000 users, from a random topic selected from $\mathcal{G}_{\text{test}}$. From those users, we create their corresponding subgraph $g_u^{(k)}$ centered on them, containing all neighbors at k level. In theory, we define a true level k, at which a user participates in a controversial subject. However, this k value is difficult to choose without sociological and philosophical knowledge of the controversy. Therefore, we compare model performances at different levels of k, using the *cross-entropy* function as our metric.

4.3 Baseline

We define two baseline models using different types of features as input.

- "$\text{GRAPH}_{\text{DEGREE}}$" uses the same methodology as our approach, with 2 GAT layers, but it uses only structural information. Instead of textual features as input, we use the degree of the node as user features.
- "$\text{TEXT}_{\text{BERT}}$" is based on a BERT model and uses only the user's tweets to predict participation in a controversial topic. The BERT model is fine-tuned using the original tweets, labeled controversial or not, according to the topic label. Each user has a collection of tweets and retweets. We treat tweets and retweets equivalently. The final user's prediction corresponds to the average predicted probability of each of his tweets (and retweets) belonging to a controversial topic.

5 Results

We evaluate our approach by varying various characteristics and parameters. To obtain the tweet representations, we add a layer of dimension 768 to the BERT output. To reduce computing and time costs, the weights of this additional layer only are updated during the training phase. The MEAN aggregator is used to represent users based on their tweets. Finally, concerning the GNN layers, we test the two methods presented in section 3.3.2 (GAT and GRAPHSAGE) respectively with 1, 2 and 3 convolution layers of dimensions 192, to compare local and global representation of users. These models are recalled in table 1 as follows : "GNNMODEL_AGGREGATOR_NBSLAYERS"). The models in this study were trained using a learning rate of 1×10^{-3}, a weight decay of 0.05, and a batch size of 64. The models underwent training for a maximum of 300 epochs, with

[2] https://developer.twitter.com/en/docs/twitter-api

[3] Statistics of the dataset are available in the appendices

	Average cross-entropy	
	Sampling training	full training
TEXT$_{\text{BERT}}$	23.176	
GRAPH$_{\text{DEGREE}}$	0.748	0.693
GAT_MEAN_1	0.686	**0.529**
GAT_MEAN_2	0.621	0.751
GAT_MEAN_3	0.523	1.616
GRAPHSAGE_MEAN_1	0.504	2.170
GRAPHSAGE_MEAN_2	**0.478**	1.630
GRAPHSAGE_MEAN_3	0.485	7.447

Table 1: Comparison of performance for user prediction, evaluated by averaging the *cross-entropy* loss on subgraphs $g_u^{(k)}$ for k between 0 and 5. The Sampling training process takes a fixed number of random users at each epoch, while the full training process takes every user when optimizing the loss function.

early termination implemented if the loss function fails to exhibit improvement within the initial 100 epochs. We used *ADAM* as the optimizer of the *cross-entropy* loss function.

5.1 Predicting user participation in a controversial topic

Table 1 summarizes the average loss scores for subgraphs $g_u^{(k)}$ with k ranging from 0 to 5. These subgraphs collectively account for a substantial dataset of 30 000 samples. Our objective is to assess the performance of different models in estimating a user's probability of participating in a controversial topic. To this end, we compare our models against a baseline. The results demonstrate the effectiveness of our proposed models. Across all values of k within the specified range, our models consistently outperform the baseline. These findings are consistent with our hypothesis that a combination of both structural and textual information is crucial for capturing meaningful features in the context of user participation in controversial topics. These results show that training models with a few samples selected, by randomly picking a fixed number of users from which to build our subgraphs, at each epoch, regularizes and enables most of our models to better performs. This enables methods based on GRAPHSAGE to generalize more effectively from training data. Notably, among our models, GRAPHSAGE_MEAN_2 emerges as the top-performing model, achieving an average loss value of 0.478. Our results suggest that this model excels in capturing the intricate interplay of user behavior and content characteristics in such scenarios. From this point, all future analyses will consider models trained using the sampling training process.

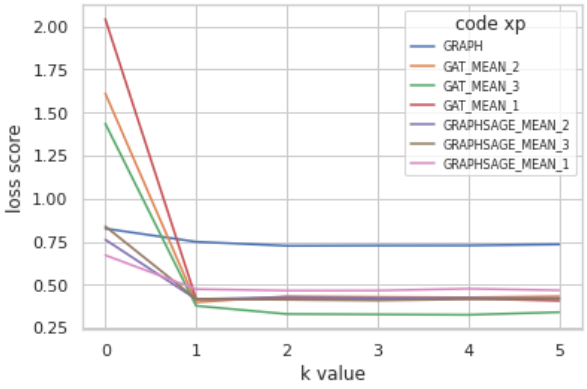

Figure 2: Loss variation across models for different k values. Note that the TEXT$_{\text{BERT}}$ baseline model is not included in the report, as its performances are significantly elevated in comparison to other values.

In figure 2, we explore how model performance evolves concerning the parameter k, which represents the number of neighbor levels considered. Notably, we observe that the estimation task is challenging when $k = 0$, highlighting the inherent complexity of accurately predicting user participation in controversial topics with only local user information. As we move to higher values of k, particularly when $k = 1$, we witness a substantial improvement in performance across all models. This suggests that considering immediate neighbors in the graph significantly enhances the accuracy of predictions. Moreover, it is interesting to note that the loss values appear to stabilize from $k = 1$ onward, indicating that the added benefit of expanding the graph to include further steps is limited. The model GAT_MEAN_3 consistently outperforms other models for all values of k beyond $k = 1$. However, it faces specific challenges when $k = 0$. The reason behind this performance gap may lie in the model's reliance on attention weights computed over neighbors. In cases where the user u is isolated within the graph $g_u^{(0)}$, this reliance becomes challenging and may explain the observed difficulties in estimation accuracy.

5.2 Controversy Quantification

As demonstrated in Section 3.2, the model that best estimates the probability of a user's participation in a controversial topic is also the one obtaining the best quantification score. Using equation 3, we compute our quantification score CQS for the ten topics included in $\mathcal{G}_{\text{test}}$, with $\hat{\eta}$ represented by our top-performing model, GAT_MEAN_2.

Quantification scores		ROC-AUC
Baseline	*rwc_score*	0.76
	dipole_score	0.8
CQS	GRAPH$_{\text{DEGREE}}$	0.84
	TEXT$_{\text{BERT}}$	0.92
	GRAPHSAGE_MEAN_2	**1.0**

Table 2: Comparison of the area under the ROC curve for scores computed from estimators, using different features. The ROC-AUC are computed based on scores estimated for the topics included in the test set $\mathcal{G}_{\text{test}}$.

Next, we investigate whether our score provides a good separation between controversial and non-controversial topics. To do so, we evaluate, using the ROC-AUC score, the ability of the scores to distinguish between classes by measuring the area under the ROC curve. We also visually analyze the kernel density estimation of the distribution's density based on the topic labels for each score. Furthermore, we compare our score with two controversial polarization scores (Garimella et al., 2018): the *rwc_score* based on random walk sampling and the *dipole_score* based on the distribution and alignment of electrical charges (nodes) in a molecule (graph). Additionally, we compare our score CQS with two other estimators $\hat{\eta}$, which only use structural or textual characteristics: GRAPH$_{\text{DEGREE}}$ and TEXT$_{\text{BERT}}$.

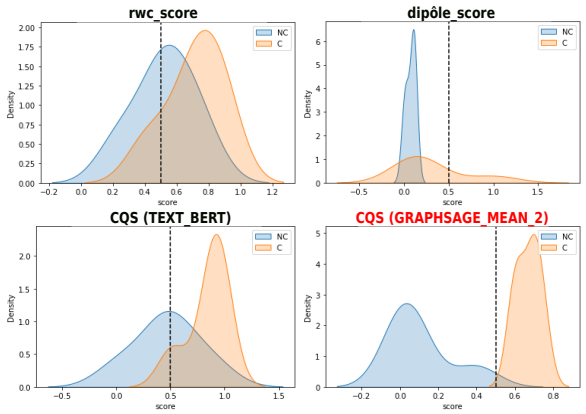

Figure 3: Diagram of kernel density estimations based on topic labels for each of the quantification scores. The blue curve depicts the distribution of non-controversial topics (NC), whereas the orange curve illustrates controversial topics (C). The more distinct the distributions, the better the quantification quality.

Table 2 compiles the ROC-AUC for all scores. CQS, with $\hat{\eta}$ represented by GRAPHSAGE_MEAN_2, achieves better results than scores based solely on structure or tweets, or than literature scores. CQS provides additional insights on how to distinguish between topics. In figure 3, it can be observed that our approach shows fewer overlaps between the curves of controversial and non-controversial topics, compared to other estimators. The distribution curves based on topic labels are centered and spread around the average score [4] ($x = 0.5$) for CQS with GRAPHSAGE_MEAN_2 as estimator. These results demonstrate the proper distribution of subjects during the quantification phase.

6 Conclusion

This scientific paper introduces a theoretical method for quantifying controversy based on user participation, presenting an innovative approach to estimate controversy scores through a graph neural network (GNN) that incorporates both structural and textual information. The results of our study demonstrate the efficacy of our proposed model, particularly when employing a sampling training process, which consistently outperforms our baseline in predicting the probability of user participation in controversial topics. Moreover, our approach surpasses existing state-of-the-art quantification scores, which predominantly rely on the structural polarity of controversial Retweet graphs (Garimella et al., 2018). This suggests the robustness and versatility of our GNN-based methodology in capturing the nuanced dynamics of controversy within online topics.

In considering avenues for improvement, one perspective involves the calibration of our model (Ghoshal and Tucker, 2022). Neural networks are prone to challenges in outputting accurate probabilities during label predictions. Addressing this aspect could enhance the precision of our controversy quantification model (Kull et al., 2019). Another promising perspective for future research lies in the augmentation and training of our data on a more extensive set of graphs spanning various fields. This broader dataset would ensure the coverage of a diverse array of topics during training, potentially enhancing the generalizability and applicability of our model across different domains. By incorporating these perspectives, our proposed methodology could be further refined and adapted to better address the evolving landscape of online controversy detection.

[4] CQS ranges from 0 to 1. 0 indicates no controversy, whereas 1 shows high controversy.

Limitations

The main limitation of our work concerns the used dataset. Indeed, only 20 subjects are employed in the training dataset, rendering the sample size rather small. Several controversial topics are related to the same main topic and only separated by different timeframes. It precludes our model from learning from more different patterns and therefore reduces its generalization capability. Similarly, the test dataset comprises only 10 subjects. To mitigate this issue and enhance our model's generalization, a fixed number of subgraphs are selected at each epoch during the training phase, facilitating better regularization of our models. The same protocol is followed during the model evaluation phase to improve the quality of our metric. Expanding the number of subjects (controversial or not) in our database, as well as the underlying domains, would contribute to a more robust generalization of our model.

Furthermore, the controversy quantification performed is static and corresponds to an image at a given moment. Unfortunately, if a user changes their opinion, this evolution would not be accounted for during the analyzed time frame. A temporal study of the evolution of controversy (Wang and Aste, 2022) would be necessary to address this limitation.

Acknowledgement

This work was supported by grants from the Janssen Horizon endowment fund. It was granted access to the HPC resources of IDRIS under the allocation AD011012604 made by GENCI.

References

Kareem Darwish. 2019. Quantifying polarization on twitter: the kavanaugh nomination. In *Social Informatics: 11th International Conference, SocInfo 2019, Doha, Qatar, November 18–21, 2019, Proceedings 11*, pages 188–201. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT Conference: Human Language Technologies, Volume 1*, pages 4171–4186.

Luc Devroye, László Györfi, and Gábor Lugosi. 2013. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media.

Hanif Emamgholizadeh, Milad Nourizade, Mir Saman Tajbakhsh, Mahdieh Hashminezhad, and Farzaneh Nasr Esfahani. 2020. A framework for quantifying controversy of social network debates using attributed networks: biased random walk (BRW). *Soc. Netw. Anal. Min.*, 10(1):90.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *ACM Trans. Soc. Comput.*, 1(1):3:1–3:27.

Biraja Ghoshal and Allan Tucker. 2022. On calibrated model uncertainty in deep learning. *arXiv preprint arXiv:2206.07795*.

Pedro Henrique Calais Guerra, Wagner Meira Jr., Claire Cardie, and Robert Kleinberg. 2013. A measure of polarization on social media networks based on community boundaries. In *Seventh International Conference on Weblogs and Social Media, ICWSM*. The AAAI Press.

William L. Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, pages 1024–1034.

Jack Hessel and Lillian Lee. 2019. Something's brewing! early prediction of controversy-causing posts from discussion features. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 1648–1659.

Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PloS one*, 9:e98679.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32.

Titouan Lorieul. 2020. *Uncertainty in predictions of Deep Learning models for fine-grained classification*. Ph.D. thesis.

Marcelo Mendoza, Denis Parra, and Álvaro Soto. 2020. GENE: graph generation conditioned on named entities for polarity and controversy detection in social media. *Inf. Process. Manag.*, 57(6):102366.

Alfredo Jose Morales, Javier Borondo, Juan Carlos Losada, and Rosa M. Benito. 2015. Measuring political polarization: Twitter shows the two sides of venezuela. *CoRR*.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations, ICLR*. OpenReview.net.

Yuanrong Wang and Tomaso Aste. 2022. Sparsification and filtering for spatial-temporal gnn in multivariate time-series. *arXiv preprint arXiv:2203.03991*.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How Powerful are Graph Neural Networks? *arXiv:1810.00826 [cs, stat]*. ArXiv: 1810.00826.

Juan Manuel Ortiz De Zarate and Esteban Feuerstein. 2020. Vocabulary-based method for quantifying controversy in social media. In *Ontologies and Concepts in Mind and Machine - 25th International Conference on Conceptual Structures, ICCS*, volume 12277 of *Lecture Notes in Computer Science*, pages 161–176. Springer.

Lei Zhong, Juan Cao, Qiang Sheng, Junbo Guo, and Ziang Wang. 2020. Integrating semantic and structural information with graph convolutional network for controversy detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 515–526. Association for Computational Linguistics.

Proof.

$$\begin{aligned}
&\mathbb{E}_G\left[\left|\widehat{CQS}(G,k) - CQS(G,k)\right|\right] \\
&\quad= \mathbb{E}_G\left[\left|\mathbb{E}_{u\sim\mathbb{U}(G)}\left[\hat{\eta}(g_u^{(k)}) - \eta(g_u^{(k)})\right]\right|\right] \\
&\quad\leq \mathbb{E}_G\left[\mathbb{E}_{u\sim\mathbb{U}(G)}\left[\left|\hat{\eta}(g_u^{(k)}) - \eta(g_u^{(k)})\right|\right]\right] \\
&\quad\leq \sqrt{\frac{2}{\mu}\mathbb{E}_G\left[\mathbb{E}_{u\sim\mathbb{U}(G)}\left[\mathrm{Reg}\ell(\hat{\eta}; g_u^{(k)}\right]\right]}
\end{aligned} \tag{9}$$

□

Topic	Timeframe	# Tweets	# Users (nodes)	# Retweets (edges)	Description
IMPEACHMENT-5-10	31oct–10nov, 2015	123 697	20 878	51 404	Roussef impeachment
MENCIONES-1-10ENERO	1–11jan, 2018	81 209	25 591	49 034	Macri's mentions
MENCIONES-11-18MARZO	11–18mar, 2018	406 869	31 659	58 797	Macri's mentions
MENCIONES-20-27MARZO	24–26mar, 2018	97 950	34 975	68 990	Macri's mentions
MENCIONES-05-11ABRIL	5–10apr, 2018	220 460	63 358	144 600	Macri's mentions
MENCIONES05-11MAYO	5–10may, 2018	267 283	63 030	146 217	Macri's mentions
BOLSONARO27	27oct, 2018	120 162	45 629	88 160	Brazilian elections
BOLSONARO28	28oct, 2018	151 952	84 986	104 955	Brazilian elections
BOLSONARO30	30oct, 2018	174 565	73 399	130 599	Brazilian elections
KAVANAUGH06-08	8oct, 2018	157 721	71 933	123 055	Kavanaugh's nomination
KAVANAUGH16	3oct, 2018	168 571	66 765	131 270	Kavanaugh's nomination
KAVANAUGH02-05	5oct, 2018	181 202	74 834	145 476	Kavanaugh's nomination
LULA_MORO_CHATS	10–11jun, 2019	199 423	66 462	143 318	Lula's mentions during Moro chats news
LEADERSDEBATE	11–21nov, 2019	250 000	76 863	174 466	Candidates debate
PELOSI	6dec, 2019	252 000	95 558	209 044	Trump Impeachment
AREA51	3–13jul, 2019	178 220	107 460	156 481	Jokes about Area51
OTDIRECTO20E	13–20jan, 2020	148 061	25 436	95 321	Event of a Music TV program in Spain
VANDUMURUGANAJITH	23jun, 2019	167 434	8401	113 208	Ajith's fans
NINTENDO	19–28may, 2019	166 145	94 255	105 793	Nintendo's release
MESSICUMPLE	23–24jun, 2019	177 770	98 448	128 099	Messi's birthday
WRESTLEMANIA	8apr, 2019	213 355	61 051	106 347	Wrestlemania event
KINGJACKSONDAY	24–27mar, 2019	142 240	39 838	107 298	popstar's birthday
NOTREDAM	16apr, 2019	171 306	99 346	146 280	Notredam fire
THANKSGIVING	28nov, 2019	250 000	155 358	164 174	Thanksgiving day
HALSEY	7–8jun, 2019	237 501	98 008	204 149	Halsey's concert
FELIZNATAL	25–26dec, 2019	305 879	193 989	212 893	Happy Christmas wishes
EXODEUX	7nov, 2019	179 908	37 384	135 579	EXO's new album
BIGIL	21–22jun, 2019	205 557	25 830	171 322	Vijay's birthday
CHAMPIONSASIA	24nov–1dec, 2019	221 925	68 754	145 829	Al-Hilal champion
SEUNGWOOBIRTHDAY	23dec, 2018	251 974	18 977	193 183	Segun Woo singer birthday

Table 3: Statistics on the dataset and graph for data retrieved from each different topic. The first 15 topics represent controversial topics, whereas the last 15 represent non-controversial topics.