HyperTwitter: A Hypergraph-based Approach to Identify Influential Twitter Users and Tweets

Lulwah Alkulaib* ‡, Abdulaziz Alhamadani*, Shailik Sarkar*, and Chang-Tien Lu*
* Department of Computer Science, Virginia Tech, Falls Church, VA 22043 USA
‡ Department of Computer Science, Kuwait University, Kuwait
{lalkulaib, hamdani, shailik, ctlu}@vt.edu

***Abstract*—Social media platforms have become an easy method of communication for many users. Content posted on social media can influence those who are exposed to it, and users who posted that content are referred to as influencers. Identifying influencers has many applications in marketing, politics, and even health awareness. While research identifying influential users across multiple fields has been studied extensively, users' influence varies in different topics. Recent studies in topic-specific influence have shown that identifying influencers on the topic-level is more effective. However, most of the existing influencer detection approaches focus only on influential user identification and do not consider that some content can be influential regardless of who published it. This paper investigates the problem of detecting topic-specific influential users and tweets in Twitter datasets. We introduce HyperTwitter, a framework that uses a Twitter sub-graph consisting of users, tweets, and interactions as input. HyperTwitter generates a hypergraph with hyperedges of two types: networks and topic edges, then measures the topic distribution for both users and tweets. With this distribution and the constructed hypergraph, we create a local, topic-based influence ranking for each user and tweet. We conduct extensive experiments with two Twitter datasets and show that the proposed framework outperforms existing baselines significantly.**

***Index Terms*—hypergraph, hypergraph learning, influential user, topic modeling**

I. Introduction

The growth of social media platforms led millions of users worldwide to connect with each other, share information, and express their feelings or ideas. Users tend to interact with friends, family, and content related to their interests by forming the following relationship on those platforms. Users can perform various actions on Twitter when engaging with others they follow. Tweets, follow, replies, likes, retweets, and quote tweets of users in the same network are behavioral evidence of the relationship between users. The literature on influential users focused on the number of followers one has [1] then evolved to the ability of a user to affect the behavior, attitudes, or feelings of other users in their network [2]. Identifying influential users has a wide range of applications, such as marketing [3], political campaigning [4], and health awareness [5].

Analyzing the influence of users on social media has attracted a lot of attention recently. There has been a significant amount of work studying influence in networks using topological features [6], textual-similarity features [7], and handcrafted-user-related features [6]. While research identifying influential users across multiple fields is important, topic-specific influence has been proven to be more effective [8]. Even though multiple studies utilized hybrid methods where combined features were used to identify influential users, user-topic-specific influence in Twitter networks has been overlooked.

On Twitter, users share a tweet with their local network of users following them. The tweet's Original Poster (OP) profile and their followings interaction with the tweet provide rich information that is useful to infer influence in OP's network. Information such as the number of followers, retweets, likes, replies, and the amount of time it takes other users to interact with OP's tweets are indicative of their influence in their network. The combination of interaction and network information in identifying the influence of a Twitter user on a specific topic in their network is the core idea of our work.

This paper proposes using a hypergraph-based learning approach to measure topic-specific influence in a Twitter network. We address two main challenges in this proposal:

1) **Utilizing the short-text data in Twitter networks to accurately determine user and tweet topic distribution.** Modeling short texts has been a challenge that is neglected by many existing works. The scarcity in datasets to train those models makes it just as difficult.
2) **Measuring influence for both users and tweets based on the composite features inferred from interactions, networks, and topic distributions.** Even though exploiting multiple features to infer user influence on other Twitter users has been implemented, studying features extracted from tweets to determine influence is a complex task. According to our survey, existing works haven't built models that determine influential tweets.

Different from traditional influencer detection models, which relied on textual-topic-specific features that worked well for users in large networks [3], our proposed method utilizes topic, network, and interaction features for various learning objectives. We construct two types of hyperedges: (1)network-hyperedges to rank influence in a Twitter network based on interactions and network features and (2)topic-hyperedges to learn topic distribution using textual-topic features and rank influence in specific topics. Our main contributions are summarized as follows:

- **Develop a hypergraph framework that detects in-**

978-1-6654-8045-1/22/$31.00 ©2022 IEEE

fluential users and tweets. The framework constructs a hypergraph from a Twitter sub-graph and interaction information and calculates topic distribution to rank both users and tweets based on their influence on specific topics. To the best of our knowledge, this is the first hypergraph framework that detects both influential users and tweets.

- **Propose an effective topic modeling method for short texts.** Short texts retrieved from Twitter are a challenge to model. We extend a DMM-based model by integrating a corpus to correctly model the short texts for each category in our dataset. Moreover, we adopt an LDA-based model for our user topic modeling.
- **Perform extensive experiments to demonstrate the efficacy of our proposed framework.** The proposed framework was evaluated on both influential node detection and topic modeling. An ablation study was also performed to confirm the significance of hyperedge types in our proposed model. The results show that the proposed framework outperforms existing baselines on all tasks.

II. Related Work

A. Influential User Identification

Identifying influential users has attracted increasing attention in recent years. Initially, methods focused on identifying users with the largest number of followers [1], then researchers combined network structure and textual content to identify influential users [6]. The drawback of these methods is that they were modeling user influence individually and ignored collective influence, which can be incorrect [3]. More recently, machine learning models have been adopted in identifying influential users. All these mentioned works either identify influencers without considering the related topic or do not utilize the comprehensive information that can be collected from Twitter content and networks.

B. Topic Modeling

A huge collection of documents can be organized using topic modeling by classifying the documents into various subjects. A standard clustering method assigns one topic to each document and takes into account the distance between topics, whereas topic modeling distributes a document to a group of topics with varying weights or probabilities without taking into account the distance between topics. In previous works, LDA has been used to perform topic modeling for a list of user-published texts to get specific user interests [9], [10]. Dirichlet Multinomial Mixture (DMM)-based models assume that each text is sampled from one topic; these models have been successfully adopted to infer latent topics in short texts [11]. Existing works have enhanced the performance of DMM-based models by incorporating word embeddings into the model [12]. In our work, we use LDA to learn the topic distributions for a user node and incorporate a word vector trained on relevant data into our DMM-based model to correctly identify topics in short tweet texts.

C. Hypergraph Learning

Recently, learning with hypergraphs has attracted a lot of attention in tasks like classification [13], link prediction [14], community detection [15], and others. Since a graph generalizes to a hypergraph, hypergraph learning can be thought of as passing information along the hypergraph structure to analyze structured data and solve tasks like the ones mentioned above. Unlike graphs, hypergraph learning models the high order correlation between data, which expands the graph learning models to a high dimensional and more comprehensive nonlinear space, resulting in higher correlation modeling capabilities and subsequently better performance [16]. In our work, the hypergraph is used to effectively model users, tweets, subgraphs, and the various complex interactions on Twitter. Our hypergraph learning method combines the content of the tweets and the interaction information to identify the influence of users and tweet contents for specific topics.

III. Hypergraph Preliminary

In this section, we explain the fundamentals of hypergraphs [16]. A hypergraph is a generalization of a graph in which the edges ,known as hyperedges, are non-empty, arbitrary subsets of the vertex set. As a result, hypergraphs are exceptionally well suited for modeling social media networks because they can be used to represent different entity types and model complex relations. A hypergraph denoted as $\mathcal{G}$, consists of a set of vertices $\mathcal{V}$, a set of hyperedges $\mathcal{E}$, and each hyperedge $e_i \in \mathcal{E}$ is assigned a weight $w(e_i)$.

Let $\mathbf{W}$ denote the diagonal matrix of the hyperedge weights, i.e.,

$$\mathbf{W}(i,j) = \begin{cases} w(e_i) & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Given a hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$, the structure of the hypergraph is usually represented by an incidence matrix $\mathbf{H}$, with each entry $\mathbf{H}(v,e)$ indicating whether the vertex v is in the hyperedge e

$$\mathbf{H}(v,e) = \begin{cases} 1 & \text{if } v \in e \\ 0 & \text{if } v \notin e \end{cases}. \tag{2}$$

Based on the definition of $\mathbf{H}$, we define the degree of a hyperedge $e \in \mathcal{E}$ and the degree of vertex $v \in \mathcal{V}$ by

$$\delta(e) = \sum_{v \in \mathcal{V}} \mathbf{H}(v,e), \tag{3}$$

and

$$d(v) = \sum_{c \in \mathcal{E}} w(e) * \mathbf{H}(v,e), \tag{4}$$

respectively. Let $\mathbf{D}_e \in \mathbb{R}^{|\mathcal{E}|*|\varepsilon|}$ and $\mathbf{D}_v \in \mathbb{R}^{|V|*|V|}$ be the diagonal matrices of the hyperedge weights and vertex degrees, respectively.

The Laplacian matrix plays an important role in graph theory. For example, in spectral analysis like clustering and partitioning a graph, the solution is based on finding eigenvalues and eigenvectors for the graph's Laplacian matrix. In an ordinary graph, the Laplacian matrix is defined as $\Theta = \mathbf{D} - \mathbf{J}$,

where $\mathbf{D}$ is the diagonal matrix of vertex degrees and $\mathbf{J}$ is the adjacency matrix. Whereas in a hypergraph, the Laplacian matrix is more complicated, and using the previous definitions and considering I as the identity matrix, it is defined as:

$$\Theta = \mathbf{D}_v - \mathbf{HWD}_e^{-1}\mathbf{H}^T. \tag{5}$$

This Laplacian matrix can be normalized as

$$\Theta = \mathbf{I} - \mathbf{D}_v^{-1/2}\mathbf{HWD}_e^{-1}\mathbf{H}^T\mathbf{D}_v^{-1/2}. \tag{6}$$

To identify whether the user and tweet nodes are influential, we utilize the objective function of a generic hypergraph learning model. The objective function consists of $\Omega(\mathbf{F})$ as a regularizer which indicates the smoothness of the hypergraph label distribution, λ as the trade-off parameter, and $R(\mathbf{F})$ as a loss of learned labels from topic modeling. It is formulated as:

$$\underset{f}{\operatorname{argmax}}\Psi(f) := \{\Omega(f) + \lambda R(f)\} \tag{7}$$

A generic regularizer on the hypergraph is defined as:

$$\begin{aligned}\Omega(f) = \frac{1}{2}\sum_{e\in\mathcal{E}}\sum_{u,v\in\mathcal{V}}\frac{w(e)h(u,e)h(v,e)}{\delta(e)} \\ \times\left(\frac{f(u)}{\sqrt{d(u)}} - \frac{f(v)}{\sqrt{d(v)}}\right)^2 \\ = f^T\Theta f\end{aligned} \tag{8}$$

IV. Proposed Method

In this section, we formally define the problem statement and then introduce our proposed approach, HyperTwitter, a hypergraph-based learning method to measure topic-specific influence of users and tweets in a Twitter sub-graph.

A. Problem Definition

Suppose we have a set of Twitter users $\mathcal{U}$ where each user $u \in \mathcal{U}$ defines a tuple$(\mathcal{F}_u, \mathcal{T}_u)$. $\mathcal{F}_u$ denotes a set of followers and $t = \{\mathcal{X}_t, \mathcal{I}_t\} \in \mathcal{T}_u$ a set of tweets consisting of the textual content and interaction information. The interaction information is about which users commented, liked, or retweeted that specific tweet. The goal is to first learn the topic distribution for each user and initial tweet, we define these distributions as $\theta_u = P(z|u) \in \mathbb{R}$ and $\theta_t = P(z|t) \in \mathbb{R}$ respectively. With this information, it will be possible to create a Twitter sub-graph-based influence score $\mathcal{I}$ for each user and tweet in its specified topic. We denote this score with $\mathcal{I} = \{I_u \in \mathbb{R}, I_t \in \mathbb{R}\}$ where $u \in \mathcal{U}$ and $t \in \mathcal{T}$.

B. Framework Overview

Here we propose HyperTwitter to detect influential users and tweets in specific topics, as shown in Figure 1. As previously stated, we have a Twitter sub-graph consisting of users, tweets, and interactions as input. Based on this input, a hypergraph is constructed with topic edges between tweet nodes and network edges between multiple user nodes and tweet nodes they have interacted with. We use the short-text topic modeling framework to obtain the topic distribution of the tweet texts and the historical user texts. With this distribution and the constructed hypergraph, we are then able to create a local, topic-based influence ranking for each user and tweet. We describe these steps in detail below.

C. Hypergraph Construction

We define our Hypergraph as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, w)$. $\mathcal{V}$ is the set of vertices present in our graph. For each user and tweet, we generate one vertex in our graph denotes as $\mathcal{V}_u, \mathcal{V}_t$ respectively. $\mathcal{E}$ has edges of two different types. The **network hyperedges** to represent user interaction relationships with tweets, and **topic hyperedges** for tweets textual content relations. The set w represents the weight of each edge.

1) Network hyperedges: We denote this group of hyperedges $\mathcal{E}_{network}$. They are used to represent the interaction relations between the user and tweet nodes. To construct $\mathcal{E}_{network}$ we consider any interaction between multiple users on the same tweets. For example, if two users retweeted the same content, those user nodes and tweet nodes are connected by a hyperedge. $\mathcal{E}_{network}$ hyperedges are constructed for all interactions: likes, and retweets. For each edge $e \in \mathcal{E}_{network}$, we have that $w_e = 1$ for $w_e \in w$. Meaning that the weight for all network edges is equal to 1. Network hyperedges are later used to rank influence based on these interactions.

2) Topic hyperedges: This group of hyperedges $\mathcal{E}_{topic}$ are used to represent the textual features relations between tweet nodes. To construct $\mathcal{E}_{topic}$, we first create a list of keywords based on the text of the tweets. These keywords are then compared between all tweet pairs and create an edge based on the matching information. For example, if two tweets' text content contains the same keyword, they are connected by a hyperedge, and their hyperedge weight is based on the textual similarity, which is set to the number of matches. Formally, we define the keywords of a tweet $t \in \mathcal{T}$ as $k_t = \{k_{t_i}\}|k_{t_i} \in t$. The weight of the edge between tweet t_i and t_j is then defined as: $w(t_i, t_j) = |k_i^t = k_j^t|$. Topic hyperedges are later used to learn topic distribution using textual-topic similarity features.

D. Short-Text Topic Modeling

In order to successfully apply topic modeling to short texts such as tweets, we have to combine tweets into one document to overcome data sparsity [11]. We merge tweets connected by the same user into one document to learn the user's topic distribution by adopting an LDA-based model [9]. Suppose that we have a collection of S short-text-tweets $\mathcal{T} = \{t_1, .., t_s\}$ that share the same set of N topics of interest $\mathcal{Z} = \{z_1, .., z_n\}$. And we have a corpus of short texts $\mathcal{T}_{corpus}$ where each short text is connected to a specific user $u \in \mathcal{U}$. More specifically, following the problem definition,

$$\mathcal{T}_{corpus} = \{\mathcal{X}_t^u : \mathcal{X}_t^u \in \mathcal{T}_u | (\mathcal{F}_u, \mathcal{T}_u) = u \in U\} \tag{9}$$

To get the topic distribution for users, we combine the last 3000 tweets $\mathcal{X}_t^u$ of each user u into one document and adapt an LDA algorithm [9] to identify topics in each document by 1) learning the word representation of each topic Z, and 2)

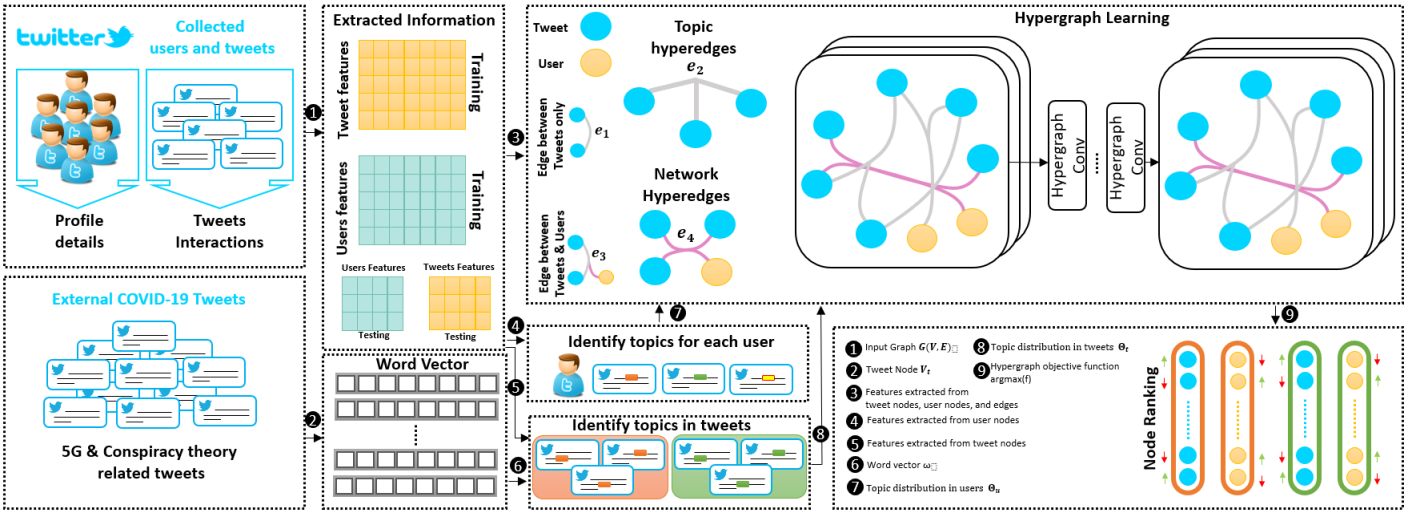

Fig. 1. Framework Overview

learning the sparse topic representation of the documents. This gives us θ_u the user topic distribution.

To enrich our text modeling, we train our own word vector using a corpus described in section V-A, this allows us to capture the semantic relations between all words in the corpus. Aiming to calculate the topic distribution for tweets, we combine word vector features learned from a large corpus of Twitter datasets with our task's adapted DMM model [17], Tweet-DMM. Like other DMM models, Tweet-DMM assumes that each short-text document is sampled by a single topic. For each word w in tweet document t, sample a binary indicator variable $\mathcal{D}_{t,w}$ from a Bernoulli distribution to indicate which model is applied to generate w, Dirichlet multinomial model or latent feature model. We describe the generative process in algorithm 1, where λ is a hyperparameter signifying the probability of a word w generated by the latent feature model, ϕ is the posterior distribution of each word belonging to a topic z, σ is the softmax function that generates word w, μ is the latent feature vectors associated with a topic z, and ω is the pre-trained word vector.

The text topic distribution is then obtained as in Tweet-DMM algorithm 1, resulting in θ_t, the tweet text topic distribution. In the next step, these distributions are used with the hypergraph to generate a local, topic-specific influence score for each node which is then used in the node ranking task.

E. Topic-Specific Node Ranking

Using the constructed hypergraph with two types of edges: topic hyperedges and network hyperedges. And the probability distributions obtained in our short-text topic modeling: θ_u for the user topic distribution and θ_t for the tweet node topic distribution. We propose a Topic-Specific Node Ranking algorithm to measure and rank the user and tweet nodes based on their influence. We explain Algorithm 2 in detail below.

The input to our ranking algorithm is given by the constructed hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, w)$, the user and tweet topic distributions θ_u, θ_t, and controlling parameters τ, β. First, we calculate the topic-specific interest of each user. Denote $i^z(f)$ as the influence score from the user u_f on the topic z and for each interaction hyperedge define the sets $\{r_{fj}\}_{z=1}^{Z}$, and $\{a_{fj}\}_{z=1}^{Z}$ as the retweet interaction and like interaction hyperedge sets, respectively. Further, the topical similarity between the tweets published by users u_f and u_j is measured with the variable $b^z(f,j)$. We define the update rules for the variables as follows:

$$r^z(f,j) \leftarrow b^z(f,j) - \max_{t \in T(f)} \{b^z(f,t) + l^z(f,t)\} \tag{10}$$

$$l^z(f,j) \leftarrow \min \left\{ 0, r^z(j,j) + \sum_{t \notin \{f,j\}} \max\{0, r^z(t,j)\} \right\} \tag{11}$$

$$l^z(j,f) \leftarrow \sum_{f' \neq j} \max\{0, r^z(f',j)\} \tag{12}$$

$T(f)$ represents the nodes of users whose tweets interest the user u_f. When aggregating the influence scores of all the tweets that user u_f is interested in, we can obtain the value $b^z(f,j)$.

$$b^z(f,j) = log \frac{g(u_f, u_j, z)}{\sum_{t \in T} g(u_f, u_t, z)} \tag{13}$$

Here, $g(u_f, u_j, z)$ is the influence score for tweets that connect users u_f and u_j:

$$g(u_f, u_j, z) = \sum_{t_p \in T_{u_j}} i^z(t_p) \tag{14}$$

where $i^z(t_p)$ denotes the influence score of tweet t_p w.r.t topic z.

We calculate the social and overall influence scores between users u_f and u_j to be able to learn a user's influence. The social influence score is calculated as:

$$q^z(f,j) = \frac{1}{1 + e^{-(r^z(f,j)+l^z(f,j))}} \tag{15}$$

The overall influence score is then calculated with the PageRank [18] algorithm with topic-based influence for each user. Where $p^z(j|f)$ is the transition probability on topic z from user u_f to u_z. It is defined as

$$p^z(j|f) = \frac{q^z(f,j)}{\sum_{j':f\to j'} q^z(f,j')} \tag{16}$$

For a user on a specific topic z, the ranking process is defined as

$$i_t^z(j) = \tau \sum_{f:f\to j} i_{t-1}^z(f) p^z(j|f) + (1-\tau) v_j^z \tag{17}$$

where v_j^z is the initial probabilistic influence score of user u_j, and t is the random walk process iteration number. v_j^z is initialized by aggregating the topic distribution of tweets of user u_j:

$$v_j^z = \sum_{t_f \in T_{v_j}} P(z_n|t_f) \tag{18}$$

The value is normalized with 1-unit. τ controls the probability of random teleportation in the PageRank algorithm and thus lies in range (0,1).

The updated calculation for the topic-specific user influence scores in Eqn. 17 becomes:

$$i_t^z = \tau Q_z i_{t-1}^z + (1-\tau) v^z \tag{19}$$

which has a unique solution derived as

$$i_\pi^z = (1-\tau)(I - \tau Q_z)^{-1} v^z \tag{20}$$

giving the social influence score $i_u^z = i_\pi^z$ for users on topic z.

We can now learn the influence of tweets after obtaining the influence for users:

Each tweet t_f has a corresponding user u_h in the hypergraph. Additionally, it has interaction relations with other users $\{u_r\}_{r=1}^C$, thus the sum is taken of the social influence of these users to compute the influence of t_p on topic z as:

$$i_t^z(t_p) = P(z_n|t_p)(\beta \sum_{r=1}^{C} i_t^z(r) + (1-\beta) i_t^z(h)) \tag{21}$$

where $P(z_n|t_p)$ is the probability of t_p on topic z, β is a parameter controlling the users contribution, and C is the number of influenced users. Until either a convergence or maximum iteration is reached, the above process is iteratively repeated between user influence and tweet influence learning.

After performing these calculations, we obtain the topic-sensitive influence scores for the sets of users and tweets $\{i_u^z, i_t^z\}_{z=1}^Z$ in the hypergraph.

Algorithm 1 Tweet-DMM

Input: Tweets short-text dataset, Twitter word vector
Output: the tweet text topic distribution θ_t
Sample a Multiple distribution of a topic proportion using Dirichlet distribution $\theta \sim Dirichlet(\alpha)$
for each topic $z \in 1, ..., Z$ **do**
    Sample a topic-word distribution using Dirichlet distribution $\theta_z \sim Dirichlet(\beta)$
end for
for tweet $t \in 1, ..., T$ **do**
    sample a topic $z_t \sim Multinomial(\theta)$
    for each word $w \in \{w_{t,1}, ..., w_{t,n_t}\}$ **do**
        sample a variable weight probability from the Bernoulli distribution $\mathcal{D}_{t,w} \sim Bernoulli(\lambda)$
        Sample a word from the topic multinomial distribution $w \sim (1-d_w) Multinomial(\phi_{z_t}) + d_w(\sigma(\mu_{z_t}\omega^M))$
    end for
end for

Algorithm 2 Topic-Specific Node Influence Ranking

Input: Hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, w)$, θ_u, θ_t and parameters τ, β
Output: Topic-specific influence scores of users and tweets $\{i_u^z, i_t^z\}_{z=1}^Z$
Initialization: Initialize each $i_{u_z}^z = \frac{1}{|V_u|}$, $i_{t_z}^z = \frac{1}{|V_t|}$
Update social influences for tweets: Compute i_t^z according to (21)
Update social influence for users: Compute i_u^z according to the calculation as described in section IV-E

V. Experiments

A. Dataset

We combine two datasets collected from Twitter regarding misinformation in connection to the COVID-19 pandemic, WICO-Text [19], and WICO-Graph [20]. Since our work incorporates topic modeling, we utilize tweets, their interaction information, and their corresponding graphs from two categories: 5G and Other conspiracies. We manually label influential tweets and users resulting in 10 influential users in the 5G category, 14 influential users in the other conspiracies category, and there are 10 influential tweet texts per category.

In order to have ground truth data for topic modeling, each category was investigated for topic areas that exist in tweets. We found that each category had three main topics and manually labeled tweets with the relevant topic. Moreover, to correctly perform topic modeling in our short texts, we collected tweets from other datasets related to our topics generating a large corpus to be used in the word vector training. The tweets were collected from 14 sources [21]–[34] related to COVID-19 conspiracy theories, and we extracted tweets that matched topics found in the 5G and Other Conspiracies datasets. Dataset statistics are summarized in Table I.

TABLE I
DATASETS DESCRIPTION

	5G	Other
Initial tweets and users	406	596
Total # of retweets	18588	32015
Total # of Users	14792	38095
Total # of Likes	26,023	68,504
Topics	Cell towers, Radiation weakens immunity, Effects on oxygen	Bill Gates, Chinese Government, Contaminated ventilators
Tweets for Word Vector	2.8M	2.4M

B. Baselines

We compare our proposed method, HyperTwitter, against five existing influential user detection baselines that are trained on textual and network features:

- **TwitterRank** [10]: An extension of the PageRank algorithm to measure user influence by considering topical similarities between users and link structures.
- **TS-SRW** [35]: Ranks users according to their topic influence using supervised PageRank-like random walks.
- **GCN** [36]: Classifies tweet and user nodes by learning their hidden representations based on their textual and network features and their neighboring nodes' features.
- **CoupledGNN** [37]: Uses two coupled graph neural networks to capture the interplay between node activation states and the spread of user influence.
- **DID** [38]: Uses language attention network and influence convolution network to identify influential users.

We also compare our proposed short-text topic modeling method, Tweet-DMM, against these conventional techniques:

- **LDA** [9]: is a generative probabilistic model for documents where each word's presence can be counted as one of the document's topics.
- **Twitter-BTM** [39]: aggregates user-based terms to learn user-specific topic distribution.
- **GPU-DMM** [17]: a DMM-based model that promotes the semantically related words under the same topic during the sampling process by the generalized Pólya urn (GPU) model.
- **GPU-PDMM** [40]: is a Poisson-based Dirichlet Multinomial Mixture model (PDMM), which was extended as the GPU-PDMM model by incorporating the GPU model during the sampling process.

C. Evaluation Metrics

We evaluate our proposed method, HyperTwitter, against the baselines using four conventional metrics: the F_1 score is used to evaluate the performance of tweets and user classification. Whereas the Area Under the Curve (AUC) is used to measure the ability to distinguish between influential and non-influential nodes. We utilized the Normalized Discounted Cumulative Gain (NDCG) and Mean Average Precision (MAP) to measure the ranking quality of influential nodes. We evaluated the performance for the topic modeling using text classification and chose accuracy and F_1 score as a metrics.

TABLE II
TOPIC-SPECIFIC INFLUENCER DETECTION PERFORMANCE

Methods	**5G**				**Other Conspiracies**			
	F1	**AUC**	**NDCG**	**MAP**	**F1**	**AUC**	**NDCG**	**MAP**
Twitter Rank	0.207	0.573	0.079	0.566	0.366	0.645	0.182	0.727
TS-SRW	0.141	0.667	0.055	0.445	0.244	0.733	0.094	0.543
GCN	0.290	0.480	0.238	0.428	0.374	0.513	0.296	0.512
CoupledGNN	0.443	0.515	0.325	0.570	0.475	0.560	0.362	0.604
DID	0.647	0.783	0.476	0.819	0.577	0.746	0.614	0.860
HyperTwitter	**0.952**	**0.984**	**0.832**	**0.989**	**0.927**	**0.973**	**0.827**	**0.976**

TABLE III
TOPIC-MODELING PERFORMANCE

Dataset	**5G**		**Other**	
Model	**Accuracy**	**F1**	**Accuracy**	**F1**
LDA	0.621	0.546	0.633	0.557
T-BTM	0.657	0.578	0.670	0.589
GPU-DMM	0.710	0.624	0.724	0.637
GPU-PDMM	0.732	0.644	0.746	0.657
Tweet-DMM	0.875	0.831	0.910	0.882

D. Implementation Details

Based on the analysis of HyperTwitter, we elaborate on some implementation details. To learn the topic distribution for tweets and users, we use the initial tweets and user nodes from each dataset category described in Table I. In the 5G dataset, we had 406 initial nodes representing users and tweets; in the other conspiracies dataset, we had 596 initial nodes. In our social influence analysis, we consider the number of hidden topics in the topic model $Z = 6$. To evaluate our topic-sensitive influence ranking, we set the parameters β and τ to 0.6 and 0.8 respectively. Topic modeling is usually affected by the number of topics z and the number of associated topic words T, but since we extend the DMM-based model with a word vector, it is less affected by the z and T values which are both set to 6.

VI. RESULTS

In our approach, the hypergraph was adopted to learn topic distributions for tweet and user nodes and perform topic-based influence ranking on those nodes. Our proposed approach significantly outperforms existing baselines in all of our experiments, demonstrating our model's superiority. We discuss our results in detail with regards to each task in our proposed method in the following:

1) Influential Nodes Detection: Table IV summarizes our model's performance in detecting the top influential nodes in specific topics. The top influential users and tweets for each topic were detected . Overall, HyperTwitter is able to outperform the existing baselines on both datasets across all conventional metrics. Compared to TwitterRank [10], and TS-SRW, [35] which identify topic influence by relying on textual features, HyperTwitter outperforms them by considering the interaction relations between users and tweets. In addition, compared to more complex models like GCN [36], CoupledGNN [37], and DID [38] that combine textual features with network information to identify influential users, HyperTwitter demonstrated superiority over other methods, which indicates that hypergraph learning is more effective in modeling the interactions between tweets and users and that the hypergraph captures these high order relations better than the baseline models. The GCN results show poor performance

TABLE IV
TOP INFLUENTIAL USERS AND TWEETS DETECTED

Dataset	Top 3 User Influencers	Top 3 Influential Tweets
5G	1)@Truth_Rises_	1)"These 5G towers are the real danger it's those damn towers causing this coronavirus I been talking about this since last year and people didn't take it serious...get you some colloidal silver and build up that immune system."
	2)@aerburr	2)" @officialWHO On #coronavirus #CoronavirusPandemic #COVID19 do you get paid to not allow @WHO to research any relation between #5G and the effects on oxygen?"
	3)@Kay_Gee20	3)"The 5G Towers are doing all this. It's not an illness.. its radiation."
Other Conspiracies	1)@ItsTommyDee1	1)"Fauci and Bill Gates have some big ideas about sickness that has fear mongering and dollar signs all over it. They're problem children of the hour for today and in some ways are threatening to hold us hostage without vaccines."
	2)@urbanx_f	2)"Chinese Communist Party funds DC think tanks and is engaged in aggressive influence ops through the United Front Work Department"
	3)@eaglechrisgold	3)"The Chinese are selling the world contaminated ventilators,they spread the virus, why buy ventilators from them?"

TABLE V
ABLATION STUDY

Methods	5G				Other Conspiracies			
	F1	AUC	NDCG	MAP	F1	AUC	NDCG	MAP
HyperTwitter	**0.95**	**0.98**	**0.83**	**0.98**	**0.92**	**0.97**	**0.82**	**0.97**
HyperTwitter -network hyperedges	0.81	0.84	0.75	0.90	0.79	0.88	0.75	0.88
HyperTwitter -topic hyperedges	0.89	0.92	0.78	0.93	0.87	0.91	0.77	0.91

that was not expected to be that low. The probable reason is that the neighborhood node features used in learning could have been more harmful than useful. The results show that the performance on the 5G dataset is slightly better than influencer detection in the Other conspiracies dataset. After inspecting the labeled data, we attribute the results to the distinctly defined topics in the 5G dataset, whereas there's more overlap in topics found in the other conspiracies dataset. Table IV shows the top-3 influential users and tweets that were detected in each topic.

2) Topic Distribution: In each dataset, classifying tweets accurately to their corresponding topic is challenging due to the overlap in topics under one category. Figure 2 shows a word cloud of keywords representing each topic and tweets that were classified to that topic area. Table III shows the classification performance of the topic distribution using our topic modeling component. The results in Table III show that for the small number of topics we have in our datasets, we were able to achieve comparative results when compared to the baselines like LDA [9], Twitter-BTM [39], GPU-DMM [17], and GPU-PDMM [40]. LDA is a standard baseline for text modeling, but it does not perform well when working on short-text data, and even though Twitter-BTM incorporates background topics for each user when modeling, on a small dataset like ours, it did not perform well. Tweet-DMM utilizes word vectors to obtain latent feature representations from external corpora to better model short texts. That proved useful in correctly identifying topics that exist in tweets, even when tweets had keywords that belonged to multiple topics. All three DMM-based methods which utilize word embeddings outperform the other baselines, which verifies the assumption that short text data is sampled from one topic. While GPU-DMM and GPU-PDMM improved on the baseline's performance, they seemed to be dataset dependant where GPU-DMM outperforms GPU-PDMM on the 5G dataset, and GPU-PDMM improves its performance over GPU-DMM on the other conspiracy dataset.

3) Ablation Study: We conduct an ablation study of HyperTwitter by removing one type of hyperedges at a time to investigate their influence. Based on the results shown in Table V, each type of hyperedges is significant in achieving high performance in our model. Although, removing topic hyperedges seems to affect the performance on both datasets more than removing the network hyperedges. The results show that our proposed method, HyperTwitter, benefits from both types of hyperedges in topic-specific influencer detection.

VII. CONCLUSION

This paper investigates the problem of detecting topic-specific influential nodes in Twitter datasets where nodes refer to users and tweets. We introduce HyperTwitter, a framework that uses a Twitter network consisting of users, tweets, and interactions as input, constructing the topic hyperedges and network hyperedges. Then, short-text topic modeling obtains the topic distribution of the tweet texts and user texts. With this distribution and the constructed hypergraph, we create a local, topic-based influence ranking for each user and tweet. Comprehensive experiments were conducted with two Twitter datasets, and the results show that the proposed model provides significant improvement to existing baselines.

VIII. ACKNOWLEDGEMENT

This research is supported in part by National Science Foundation grants CNS-2141095. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF, or the U.S. Government.

REFERENCES

[1] F. Morone, B. Min, L. Bo, R. Mari, and H. A. Makse, "Collective influence algorithm to find influencers via optimal percolation in massively large social media," *Scientific Reports*, vol. 6, 7 2016.

[2] N. Liu, L. Li, G. Xu, and Z. Yang, "Identifying domain-dependent influential microblog users: A post-feature based approach," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 28, 2014.

[3] M. A. Al-Garadi, K. D. Varathan, D. Ravana, E. Ahmed, M. Usman, S. Khan, M. A. Al-Garadi, G. Mujtaba, S. U. Khan, M. A. Al-Garadi, K. D. Varathan, and S. Devi, "Analysis of online social network connections for identification of influential users: Survey and open research issues," *ACM Computing Surveys*, vol. 51, p. 16, 2018.

[4] S. Stieglitz and L. Dang-Xuan, "Social media and political communication: a social media analytics framework," *Social Network Analysis and Mining*, vol. 3, pp. 1277–1291, 1 2013.

[5] K. Zhao, G. E. Greer, J. Yen, P. Mitra, and K. Portier, "Leader identification in an online health community for cancer survivors: a social network-based classification approach," *Information Systems and e-Business Management*, vol. 13, pp. 629–645, 11 2015.

[6] J. Jiang, C. Wilson, X. Wang, W. Sha, P. Huang, Y. Dai, and B. Y. Zhao, "Understanding latent interactions in online social networks," *ACM Transactions on the Web (TWEB)*, vol. 7, no. 4, pp. 1–39, 2013.

[7] J.-V. Cossu, N. Dugué, and V. Labatut, "Detecting real-world influence through twitter," in *2015 Second European Network Intelligence Conference*, pp. 83–90, IEEE, 2015.

[8] R. Panchendrarajan and A. Saxena, "Topic-based influential user detection: a survey," *Applied Intelligence*, 7 2022.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

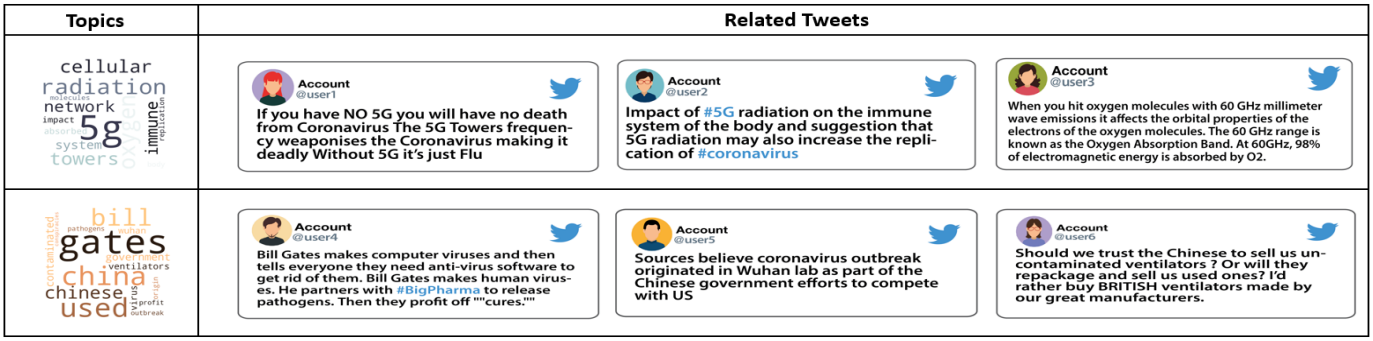

Fig. 2. Topic keywords and tweets associated with them

[10] J. Weng, E. P. Lim, J. Jiang, and Q. He, "Twitterrank: Finding topic-sensitive influential twitterers," *WSDM 2010 - Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, pp. 261–270, 2010.

[11] J. Qiang, Z. Qian, Y. Li, Y. Yuan, and X. Wu, "Short text topic modeling techniques, applications, and performance: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 3, pp. 1427–1445, 2020.

[12] D. Q. Nguyen, R. Billingsley, L. Du, and M. Johnson, "Improving topic models with latent feature word representations," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 299–313, 2015.

[13] D. Zhou, J. Huang, and B. Schölkopf, "Learning with hypergraphs: Clustering, classification, and embedding," *Advances in neural information processing systems*, vol. 19, 2006.

[14] L. Xia, C. Huang, Y. Xu, P. Dai, L. Bo, X. Zhang, and T. Chen, "Link prediction in social networks based on hypergraph," *WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web*, pp. 41–42, 2013.

[15] Z. T. Ke, F. Shi, and D. Xia, "Community detection for hypergraph networks via regularized tensor power iteration," 9 2019.

[16] Y. Gao, Z. Zhang, H. Lin, X. Zhao, S. Du, and C. Zou, "Hypergraph learning: Methods and practices," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[17] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma, "Topic modeling for short texts with auxiliary word embeddings," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 165–174, 2016.

[18] W. Xing and A. Ghorbani, "Weighted pagerank algorithm," in *Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004.*, pp. 305–314, IEEE, 2004.

[19] K. Pogorelov, D. T. Schroeder, P. Filkuková, S. Brenner, J. Langguth, and S. Bren-Ner, "Wico text: a labeled dataset of conspiracy theory and 5g-corona misinformation tweets," *dl.acm.org*, pp. 21–25, 10 2021.

[20] D. T. Schroeder, F. Schaal, P. Filkukova, K. Pogorelov, and J. Langguth, "Wico graph: A labeled dataset of twitter subgraphs based on conspiracy theory and 5g-corona misinformation tweets," vol. 2, pp. 257–266, SciTePress, 2021.

[21] S. C. Phillips, L. H. X. Ng, and K. M. Carley, "Hoaxes and hidden agendas: A twitter conspiracy theory dataset: Data paper," in *Companion Proceedings of the Web Conference 2022*, pp. 876–880, 2022.

[22] E. Ferrara, "What types of covid-19 conspiracies are populated by twitter bots?," *arXiv preprint arXiv:2004.09531*, 2020.

[23] M. Cheng, S. Wang, X. Yan, T. Yang, W. Wang, Z. Huang, X. Xiao, S. Nazarian, and P. Bogdan, "A covid-19 rumor dataset," *Frontiers in Psychology*, vol. 12, p. 644801, 2021.

[24] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, and T. Chakraborty, "Fighting an infodemic: Covid-19 fake news dataset," in *International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation*, pp. 21–29, Springer, 2021.

[25] H.-y. Lu, C. Fan, X. Song, and W. Fang, "A novel few-shot learning based multi-modality fusion model for covid-19 rumor detection from online social media," *PeerJ Computer Science*, vol. 7, p. e688, 2021.

[26] D. Gerts, C. D. Shelley, N. Parikh, T. Pitts, C. W. Ross, G. Fairchild, N. Y. V. Chavez, A. R. Daughton, *et al.*, ""thought i'd share first" and other conspiracy theory tweets from the covid-19 infodemic: exploratory study," *JMIR public health and surveillance*, vol. 7, no. 4, p. e26527, 2021.

[27] K. Hayawi, S. Shahriar, M. A. Serhani, I. Taleb, and S. S. Mathew, "Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection," *Public health*, vol. 203, pp. 23–30, 2022.

[28] M. R. DeVerna, F. Pierri, B. T. Truong, J. Bollenbacher, D. Axelrod, N. Loynes, C. Torres-Lugo, K.-C. Yang, F. Menczer, and J. Bryden, "Covaxxy: A collection of english-language twitter posts about covid-19 vaccines.," in *ICWSM*, pp. 992–999, 2021.

[29] M. K. Elhadad, K. F. Li, and F. Gebali, "Covid-19-fakes: A twitter (arabic/english) dataset for detecting misleading information on covid-19," in *International Conference on Intelligent Networking and Collaborative Systems*, pp. 256–268, Springer, 2020.

[30] E. Chen, K. Lerman, E. Ferrara, *et al.*, "Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set," *JMIR public health and surveillance*, vol. 6, no. 2, p. e19273, 2020.

[31] A. Z. Klein, A. Magge, K. O'Connor, J. I. F. Amaro, D. Weissenbacher, and G. G. Hernandez, "Toward using twitter for tracking covid-19: a natural language processing pipeline and exploratory data set," *Journal of medical Internet research*, vol. 23, no. 1, p. e25314, 2021.

[32] S. Liu and J. Liu, "Public attitudes toward covid-19 vaccines on english-language twitter: A sentiment analysis," *Vaccine*, vol. 39, no. 39, pp. 5499–5505, 2021.

[33] K. Garcia and L. Berton, "Topic detection and sentiment analysis in twitter content related to covid-19 from brazil and the usa," *Applied soft computing*, vol. 101, p. 107057, 2021.

[34] G. Muric, Y. Wu, E. Ferrara, *et al.*, "Covid-19 vaccine hesitancy on social media: building a public twitter data set of antivaccine content, vaccine misinformation, and conspiracies," *JMIR public health and surveillance*, vol. 7, no. 11, p. e30642, 2021.

[35] G. Katsimpras, D. Vogiatzis, and G. Paliouras, "Determining influential users with supervised random walks," *WWW 2015 Companion - Proceedings of the 24th International Conference on World Wide Web*, pp. 787–792, 5 2015.

[36] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[37] Q. Cao, H. Shen, J. Gao, B. Wei, and X. Cheng, "Popularity prediction on social platforms with coupled graph neural networks," *WSDM 2020 - Proceedings of the 13th International Conference on Web Search and Data Mining*, pp. 70–78, 1 2020.

[38] C. Zheng, Q. Zhang, S. Young, and W. Wang, "On-demand influencer discovery on social media," *Proceedings of the 29th ACM international conference on information & knowledge management*, vol. 1, pp. 2337–2340, 10 2020.

[39] W. Chen, J. Wang, Y. Zhang, H. Yan, and X. Li, "User based aggregation for biterm topic model," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 489–494, 2015.

[40] C. Li, Y. Duan, H. Wang, Z. Zhang, A. Sun, and Z. Ma, "Enhancing topic modeling for short texts with auxiliary word embeddings," *ACM Transactions on Information Systems (TOIS)*, vol. 36, no. 2, pp. 1–30, 2017.