Multi-Modal Community Detection in Twitter Datasets

Muhieddine Shebaro and Dr. Jelena Tešić

Department of Computer Science

TEXAS STATE UNIVERSITY

http://DataLab12.github.io/

Motivation

- **Twitter is rich in data modalities: text, images/videos, and connections.**
- **Attributed graph clustering takes into account content of the tweet as well as the connections among users.**
- **Research Question: How well do various modality clusters overlap and can the modalities be combined in a bid to get a better community description?**

State Of The Art

- Use Large Language Models (BERT) for text content features and DNN for image/video features
- Use context: user profile, and location features of geo-tagged tweets for sentiment analysis.
- Model interactions of the tweeter verse using Bi-GCN and Tail-GNN architectures to capture the underlying structure

Pipeline

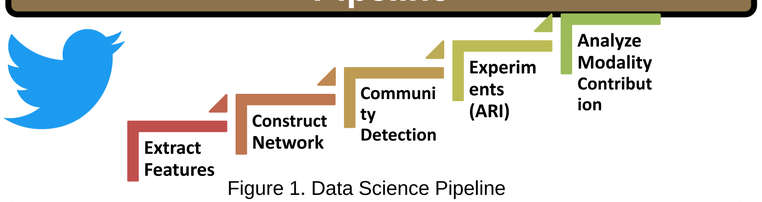


Figure 1. Data Science Pipeline

Datasets

COVID+ Dataset

- MediaEval2020 connection baseline extended and augmented
- 3.2million+ usets and 8+ million tweets
- Hashtags mined: #Coronavirus, #Covid19, and #Covid-19
- Data collected from Mar5ch to September 2020.
- pytwanalysis: Twitter Data Management And Analysis at Scale, IEEE SNAMS 2021.

MuMIN Dataset

Dataset	#Claims	#Threads	#Tweets	#Users	#Articles	#Images	#Languages	%Misinfo
MuMiN-large	12,914	26,048	21,565,018	1,986,354	10,920	6,573	41	94.79%
MuMiN-medium	5,565	10,832	12,659,371	1,150,259	4,212	2,510	37	94.20%
MuMiN-small	2,183	4,344	7,202,506	639,559	1,497	1,036	35	92.71%

Table 1. MuMIN Dataset

Feature Extraction

Textual Features	Visual Features	Network Features
Pretrained BERTweet on COVID-19 Tweets embeddings	OCR	User Attributes (verified…)
State-of-the-art text normalizations beforehand	Type of Image (B&W, Fake)	Replies
No “Fine-tuning of the Transformer” is necessary	Generic DNN (VGG16)	Quotes
	Image Captions (Captioner Locally Trained on MSCOCO)	Retweets

Table 2. Features per modalities used

- COVID(+): we extracted textual features using BERTweet and visual features using DNN
- MuMIN: Visual and Textual features provided

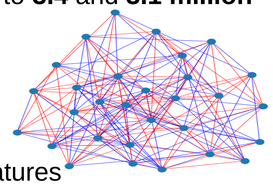
Network Construction, Pre-processing and Augmentation

	0	1
0	1240082784749793286	1239974296040005634
1	1240082784745644033	1240051311971794949
2	1240082784477155328	1240067789655769088
3	1240082784397357058	1240073334714273792
4	1240082784389120006	1239718726288568322
...	...	...
1307744486186004486	1304596764457132033	
1307744509434920960	1307743803122515969	
1307759524934283264	1307658266609176576	
1307774519231361025	1307758755464052742	
1307774692636450821	1307694487838363650	

COVID(+): Replies, Quotes, Retweets.

- Rmoved any edges in the Replies.
- Every target node should be connected to at least **10** nodes.
- Isolated nodes and duplicate edges were eliminated.
- The total number of nodes and edges dropped to **3.4** and **3.1 million**

MuMiN: Quotes, Replies.



Augmentation of COVID+ Dataset

- Original network was augmented with visual similarity graph
- New edges added from vertex to 5 similar vertices
- Similarity was computed using cosine distance between DNN features
- Number of Edges increased from 3.4 million to 4.1 million

Modeling

Graphic Neural Network Training for Community Discovery

- Leverage all modalities and aggregate features from nodes (Message Passing)
- GraphSage produces an embedding of size 50 dimensions (unsupervised).
- Epoch = 1 , batch size = 50, layer size = 50, LR = 10^-3, Adam Optimizer.
- It utilizes the neighborhood sampling improving the scalability and memory efficiency.

Louvian Clustering Algorithm

- Low Execution Time
- Ability to find communities in disconnected networks

HDBSCAN

- PCA dim 10 for BERTweet
- Robust to parameter selection
- Decent HD performance

Experiments

ARI	Network	BERTweet	GNN	Network-V	GNN-V
Network	1.0	0.084	0.0002	0.124	0.001
BERTweet	0.084	1.0	0.0004	0.053	0.0266
GNN	0.0002	0.00036	1.0	0.0001	-0.001
Network-V	0.124	0.0533	0.0001	1.0	0.0138
GNN-V	0.001	0.0265	-0.00091	0.01376	1.0

Table 3. ARI between various multi-modal modes in processed COVID (+)

Mode	# of Communities
Network	91,380
BERTweet	81,252
GNN	30,995
Network-V	67,146
GNN-V	87,505

Table 4. Number of communities in processed COVID (+)

ARI	Network	Text-Emb	GNN	Network-V	GNN-V
Network	1.0	0.00028	0.000052	0.016	0.000052
Text-Emb	0.00028	1.0	0.00066	0.0044	0.00018
GNN	0.000052	0.00066	1.0	0.000052	0.000052
Network-V	0.016	0.0044	0.000052	1.0	0.99
GNN-V	0.000052	0.00066	0.00012	0.99	1.0

Table 5. ARI between various multi-modal modes in large MuMIN dataset

Mode	# of Communities
Network	655
Text-Emb	10
GNN	3
Network-V	21
GNN-V	2

Table 6. number of communities in large MuMIN dataset

Conclusion and Next Steps

- Multiple modalities seem to capture specific information
- Not relevant for community discovery at global scale
- Have value for specific discovery and mining tasks
- Ground truth labeling missing in COVID+ to make a conclusion

Acknowledgments

The work has been supported by NAVAIR, NVIDIA @ Data Lab (DataLab12.github.io) @ TXST