Tweezers: A Framework for Security Event Detection via Event Attribution-centric Tweet Embedding

Jian Cui, Hanna Kim, Eugene Jang, Dayeon Yim, Kicheol Kim, Yongjae Lee, Jin-Woo Chung, Seungwon Shin, Xiaojing Liao*

Indiana University Bloomington, KAIST, S2W Inc., Retrv Inc.

Social Media (Twitter) is a critical source of threat intelligence

*"Social Media is a **Viable Source of Threat Intelligence**"*

- Hunting Threats on Twitter

TREND MICRO

*"Social media platforms are **a treasure trove of information** that can **give early warnings on emerging threats**"*

- What is Social Media Threat Intelligence?

Recorded Future

Social Media (Twitter) is a critical source of threat intelligence

Social Media

Social Media (Twitter) is a critical source of threat intelligence

Security event

Heartbleed 2.0? OpenSSL warns of a second-ever critical flaw…

SystemBC: The Multipurpose Proxy Bot Still Active…

Another phishing campaign distributing malicious APKs via fake Google Play sites

Social Media

Security event provides actionable threat intelligence

Social Media

Security event	Actionable Threat intelligence	
Heartbleed 2.0? OpenSSL warns of a second-ever critical flaw…	`CVE-2022-3602` `CVE-2022-3786`	CVE
SystemBC: The Multipurpose Proxy Bot Still Active…	**IP:** `1xx.1xx.1xx.1...` **Hash:** `0077***5b691`	IP, Hash
Another phishing campaign distributing malicious APKs via fake Google Play sites	`http://h**.in/` `http://h2**.in/go**`	

Challenges in security event detection from Twitter

➢ **Process Overwhelming Volume of Tweets**

The sheer volume and noise in tweets complicate accurate security event detection.

➢ **Ensure Complete Coverage of Security Events**

Existing methods cover only 2.7%–29.8% of events, missing many critical security incidents.

Existing Approaches

SONAR: Automatic Detection of Cyber Security Events over the Twitter Stream

Quentin Le Sceller
Security Research Centre
Concordia University
Montreal, Quebec, Canada
q_lescel@encs.concordia.ca

On the Automated Assessment of Open-Source Cyber Threat Intelligence Sources

Max Mühlhäuser[1]

Cybersecurity Event Detection with New and Re-emerging Words

Hyejin Shin
hyejin1.shin@samsung.com
Samsung Research

WooChul Shim*
woochul.shim@samsung.com
Samsung Research

Jiin Moon
jiin.moon@samsung.com
Samsung Research

Existing Approaches

1. Leverage Text Embedding to embed tweets

- *Word2Ve*
- *GloVe*

2. Apply Clustering Algorithms

- *K-means*
- *DBSCAN*

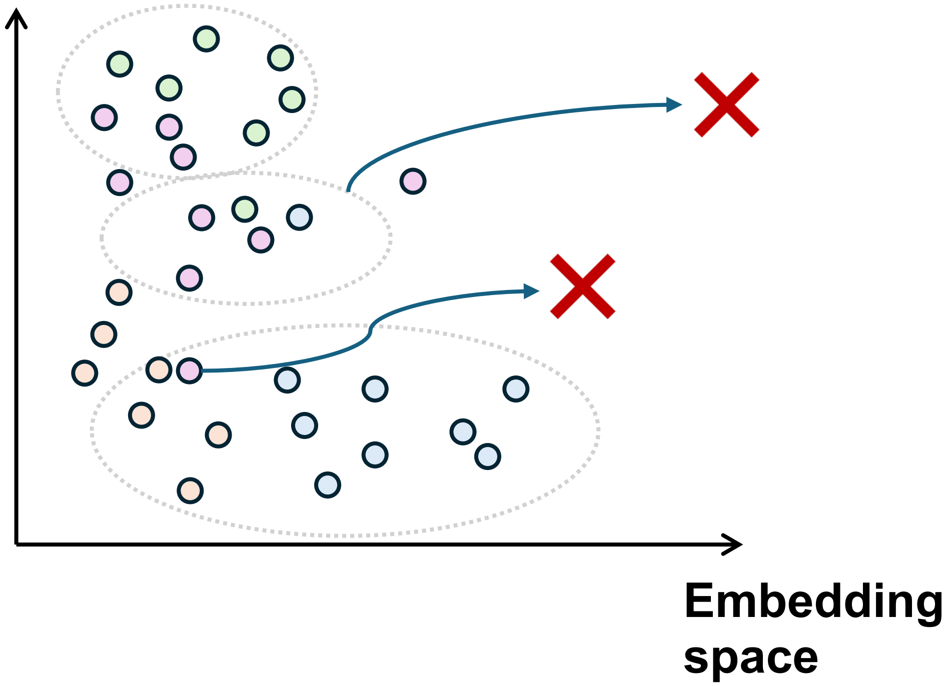
Existing Approaches

1. Leverage Text Embedding to embed tweets

- *Word2Ve*
- *GloVe*

2. Apply Clustering Algorithms

- *K-means*
- *DBSCAN*



Existing Approaches

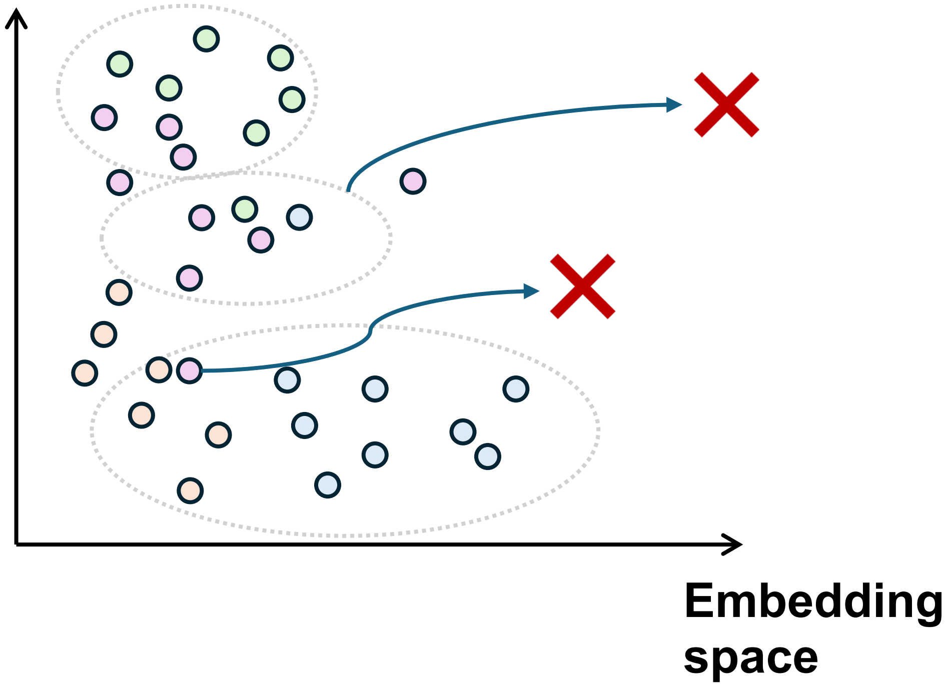
1. Leverage Text Embedding to embed tweets

- *Word2Ve*
- *GloVe*

2. Apply Clustering Algorithms

- *K-means*
- *DBSCAN*

This observation is also noted in **transformer-based models such as BERT and LLaMA**



Existing Approaches

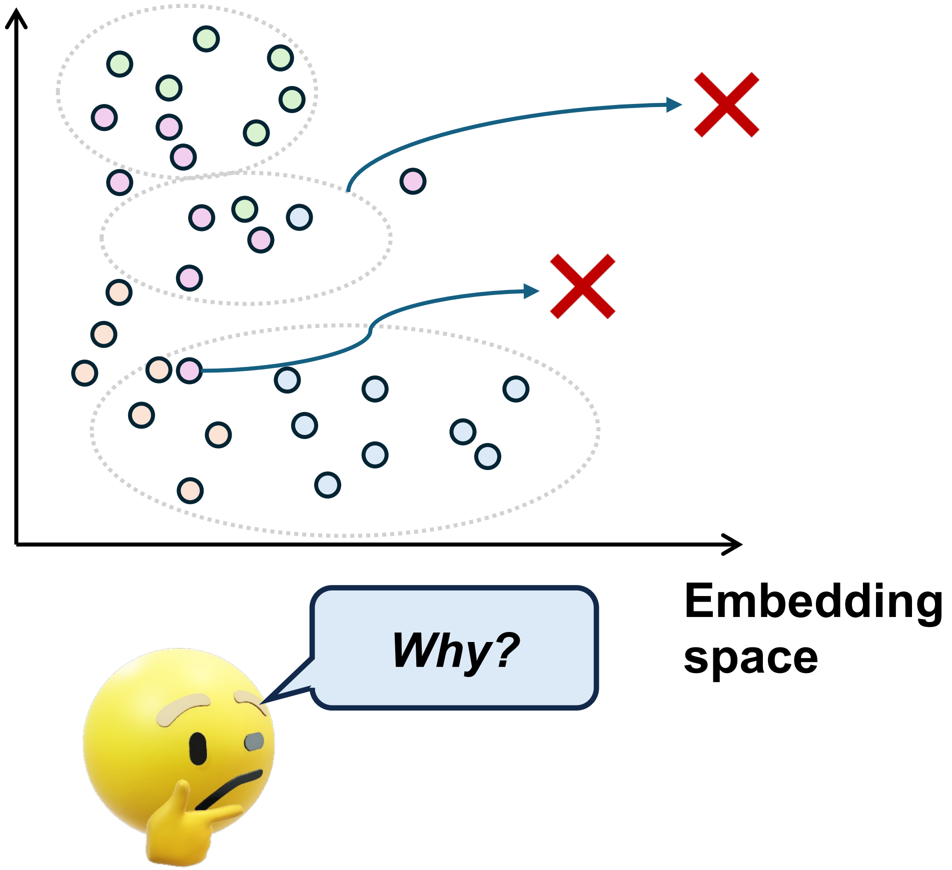
1. Leverage Text Embedding to embed tweets

- *Word2Ve*
- *GloVe*

2. Apply Clustering Algorithms

- *K-means*
- *DBSCAN*

This observation is also noted in **transformer-based models such as BERT and LLaMA**



False similarity in text embedding

WhatsApp 0-Day Bug Let Hackers Execute an Arbitary Code Remotely …. fixed two critical zero-day bugs that …

T_{e_1}

T_{e_2}

Microsoft Exchange zero-days reportedly exploited in attacks … allowing for remote code execution …

Severe WhatsApp bug (CVSS 9.8) and no one is talking about it. #RCE over … Update your WhatsApp

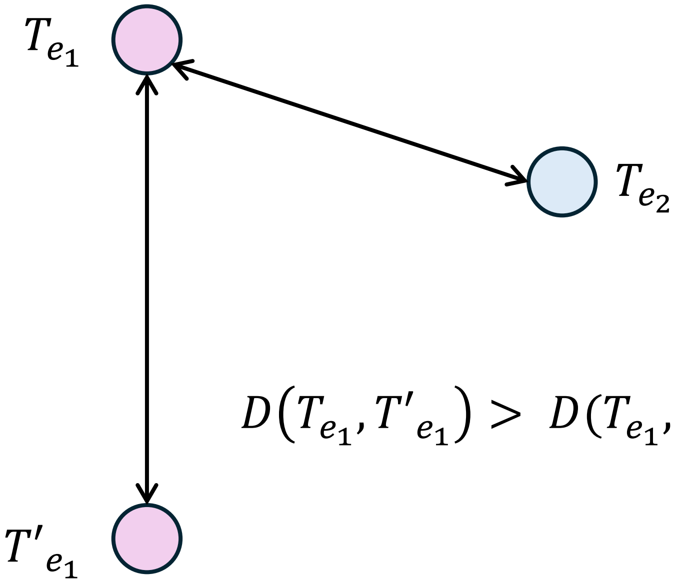
T'_{e_1}

Event1: *WhatsApp 0-day*

Event2: *Exchange 0-day*

False similarity in text embedding

WhatsApp 0-Day Bug Let Hackers Execute an Arbitary Code Remotely …. fixed two critical zero-day bugs that …



Microsoft Exchange zero-days reportedly exploited in attacks … allowing for remote code execution …

$$D(T_{e_1}, T'_{e_1}) > D(T_{e_1}, T_{e_2})$$

Severe WhatsApp bug (CVSS 9.8) and no one is talking about it. #RCE over … Update your WhatsApp

The distance between tweets from the same event is larger!

Different security event has different *security attributes* !

WhatsApp 0-Day Bug Let Hackers Execute an Arbitary Code Remotely …. fixed two critical zero-day bugs that …

Microsoft Exchange zero-days reportedly exploited in attacks … allowing for remote code execution …

Severe WhatsApp bug (CVSS 9.8) and no one is talking about it. #RCE over … Update your WhatsApp

Different security event has different *security attributes* !

WhatsApp 0-Day Bug Let Hackers Execute an Arbitary Code Remotely …. fixed two critical zero-day bugs that …

Microsoft Exchange zero-days reportedly exploited in attacks … allowing for remote code execution …

Severe WhatsApp bug (CVSS 9.8) and no one is talking about it. #RCE over … Update your WhatsApp

Different Exploited Vulnerability!

Different security event has different *security attributes* !

Swachh City platform hacked, data of 16 million users leaked https://t.co/xx

Observed Data

"Indian banks' customers … infect their #Android devices with a fake REWARD app to steal their personal data.

Malware

"NEW CYBERSECURITY NEWS: Palestinian Hacktivist Group GhostSec Compromises 55 Berghof PLCs Across Israel

Threat Actor

Different security event has different *security attributes* !

Swachh City platform hacked, data of 16 million users leaked https://t.co/xx

Observed Data

"Indian banks' customers … infect their #Android devices with a fake REWARD app to steal their personal data.

Malware

"NEW CYBERSECURITY NEWS: Palestinian Hacktivist Group GhostSec Compromises 55 Berghof PLCs Across Israel

Threat Actor

STIX™

Structured Threat Information Expression (STIX™)

- **Structured language** for describing, sharing, and **analyzing cyber threat information consistently.**
- *De facto* standard for **Cyber Threat Intelligence (CTI).**
- Defines **18 objects (entities).**

Security attributes are key to distinguish different events!

Security attributes are key to distinguish different events!

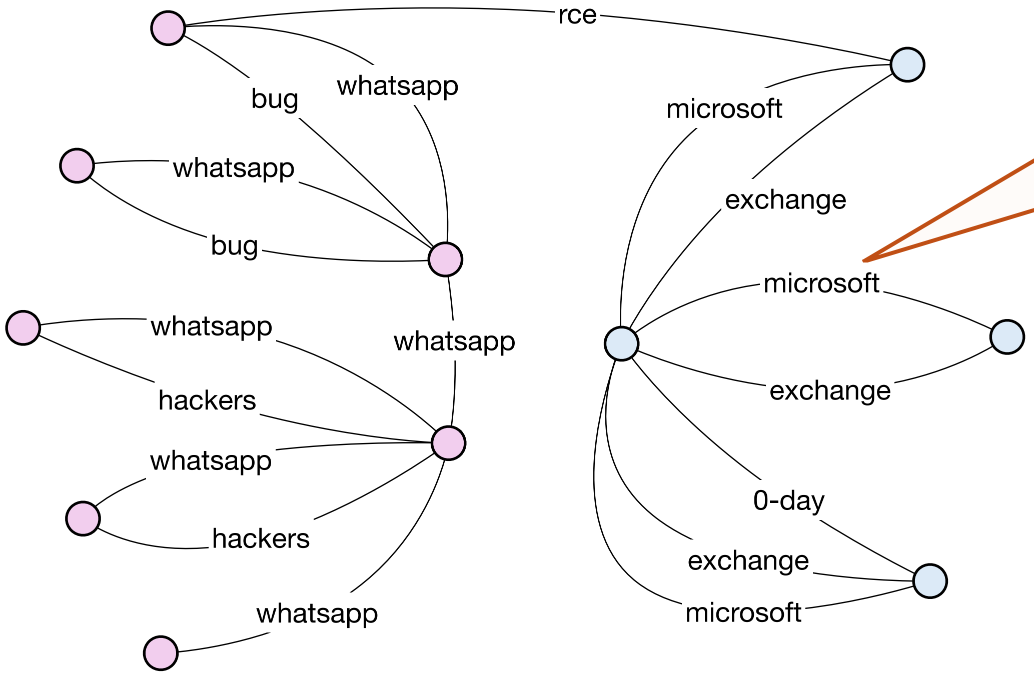
Then, how to fully leverage this information for security event detection?

Security attributes are key to distinguish different events!

Then, how to fully leverage this information for security event detection?

Graph Representation!

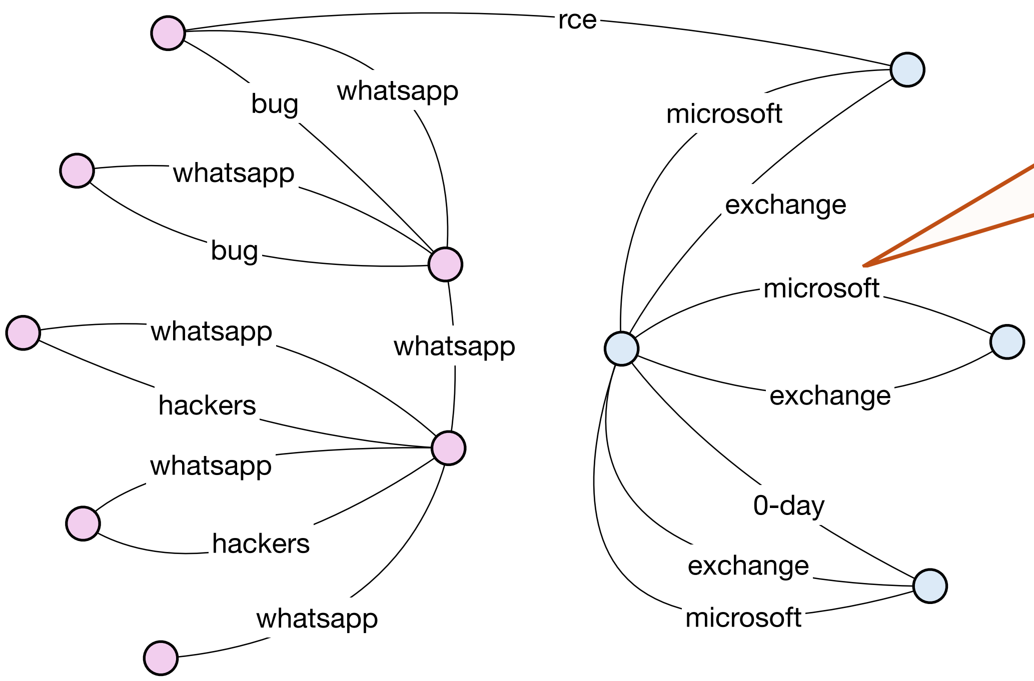
Tweet Relation Graph Representation



Tweet Relation Graph

Tweets with common security attributes are connected

Tweet Relation Graph Representation

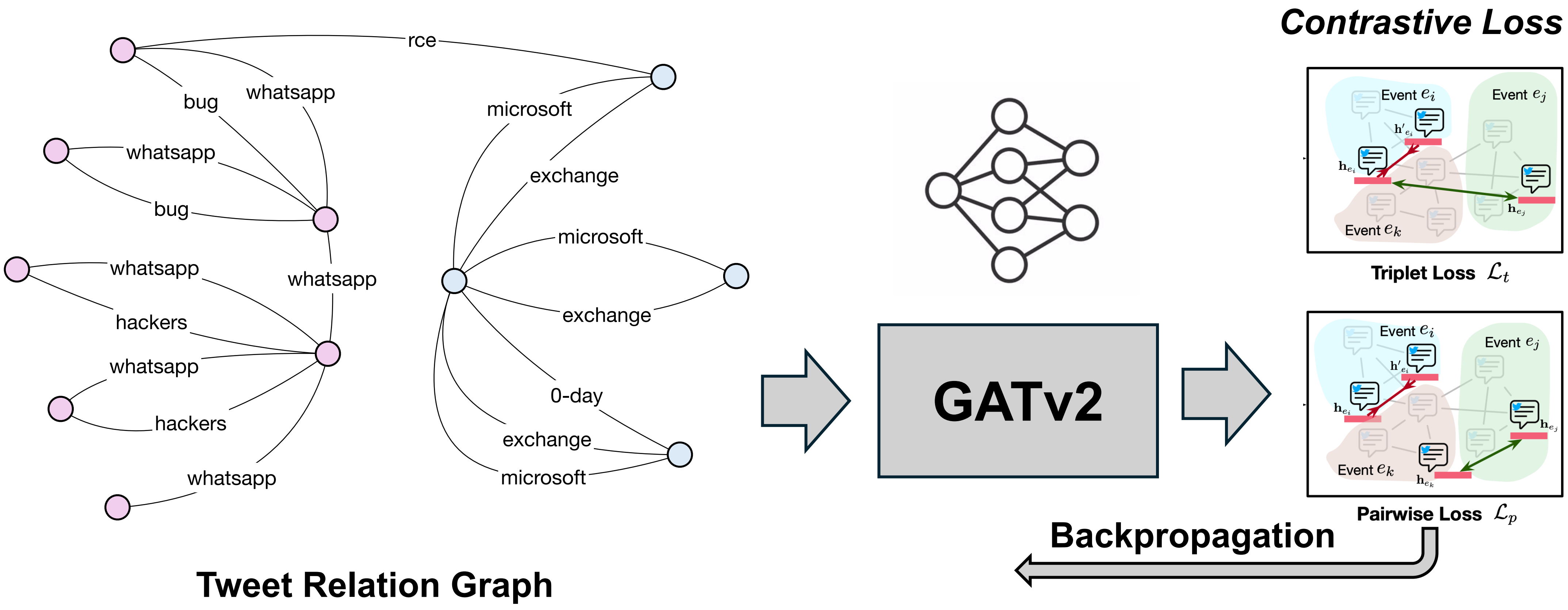


Tweet Relation Graph

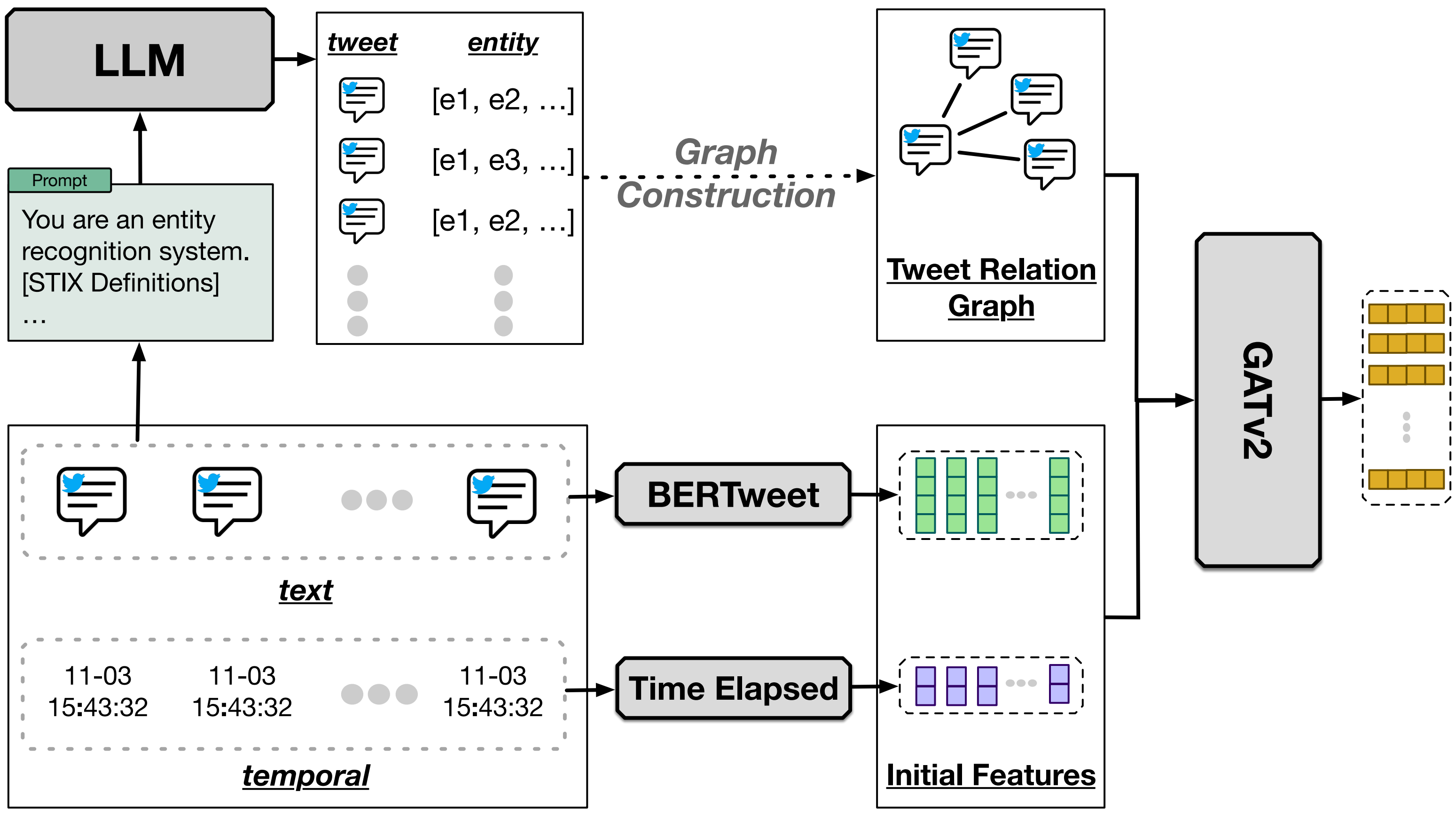
Tweets with common security attributes are connected

Represents how security attributes are shared across tweets

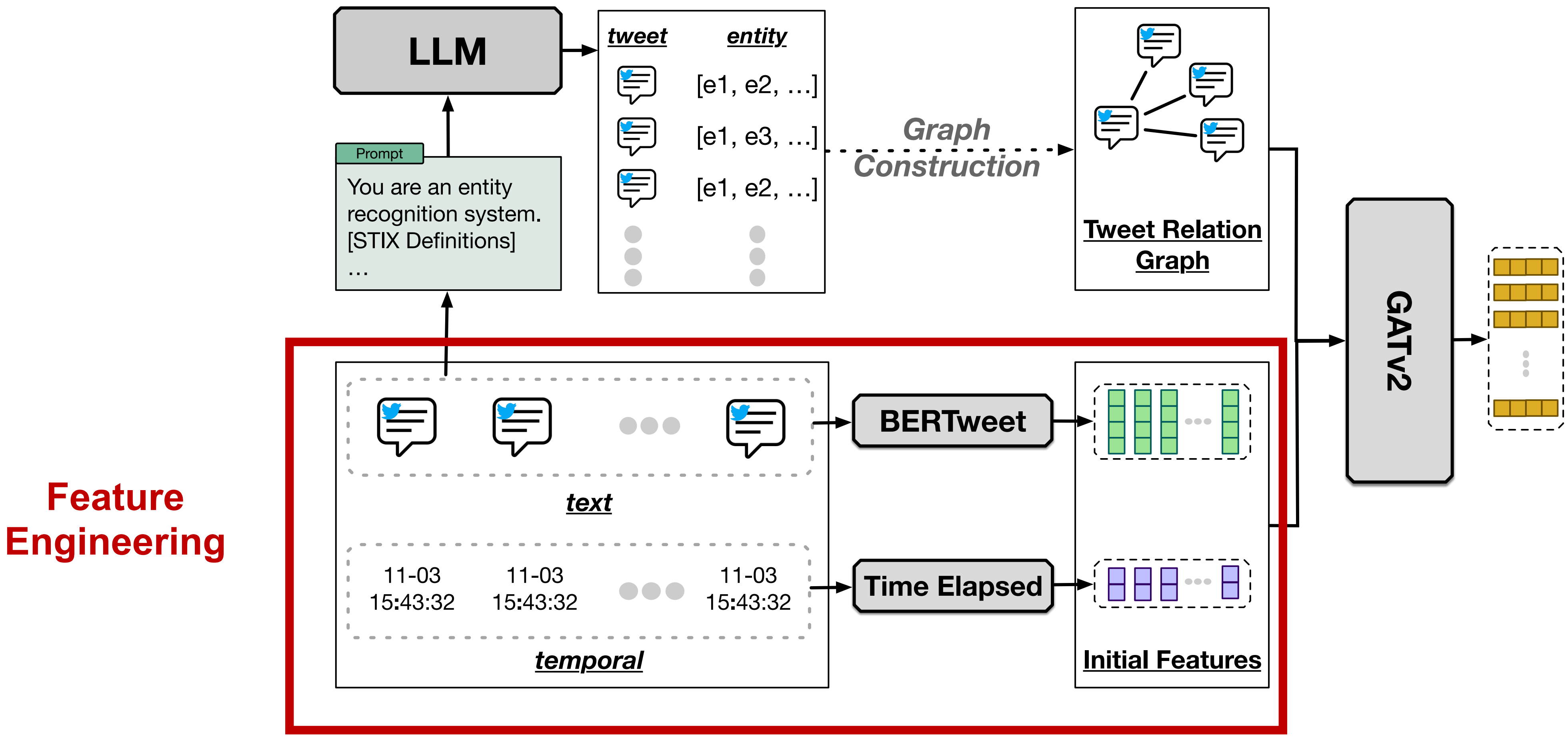
Embedding Security Attributes with Graph Neural Networks



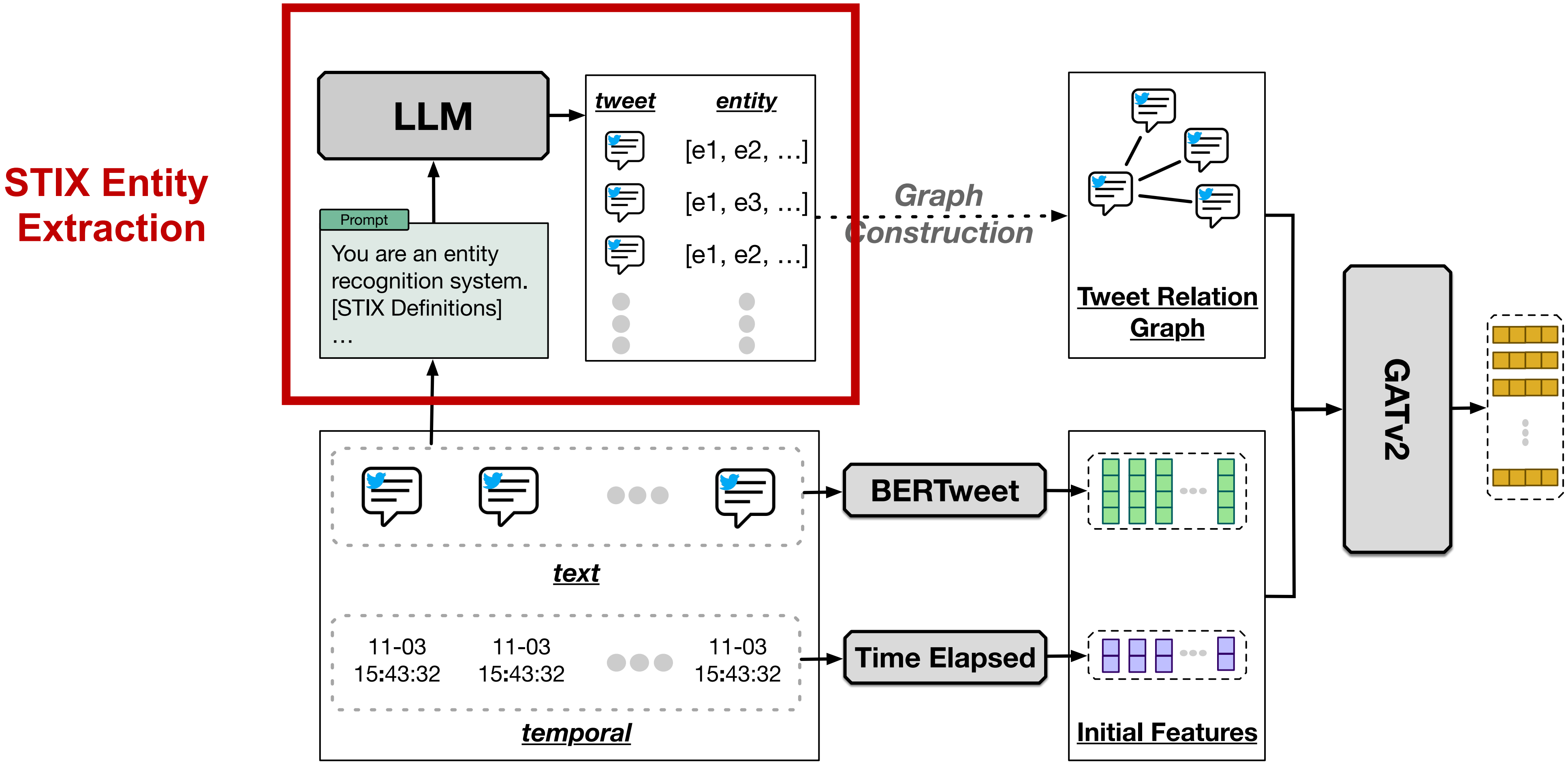
Tweet Embedding Workflow



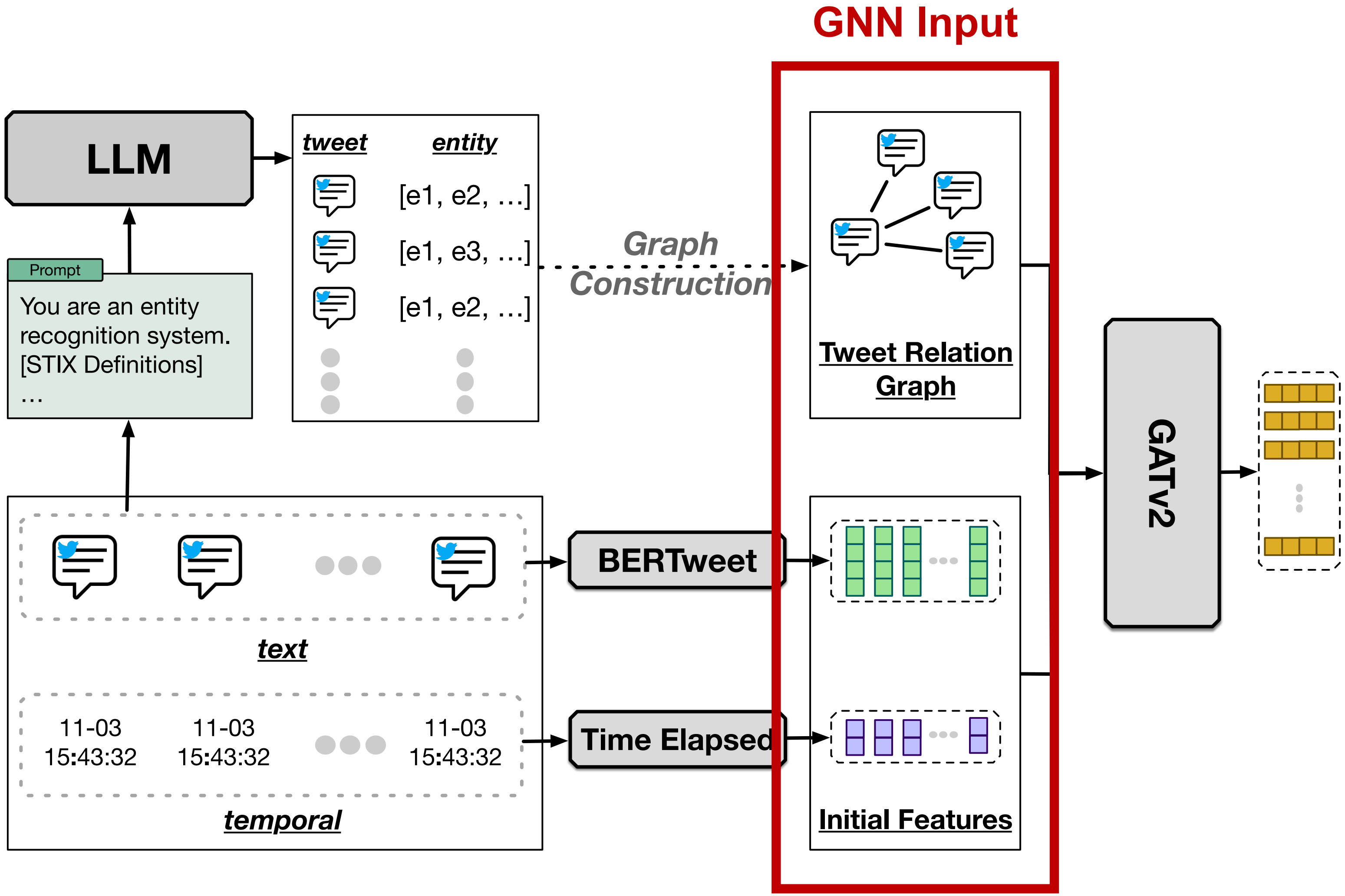
Tweet Embedding Workflow



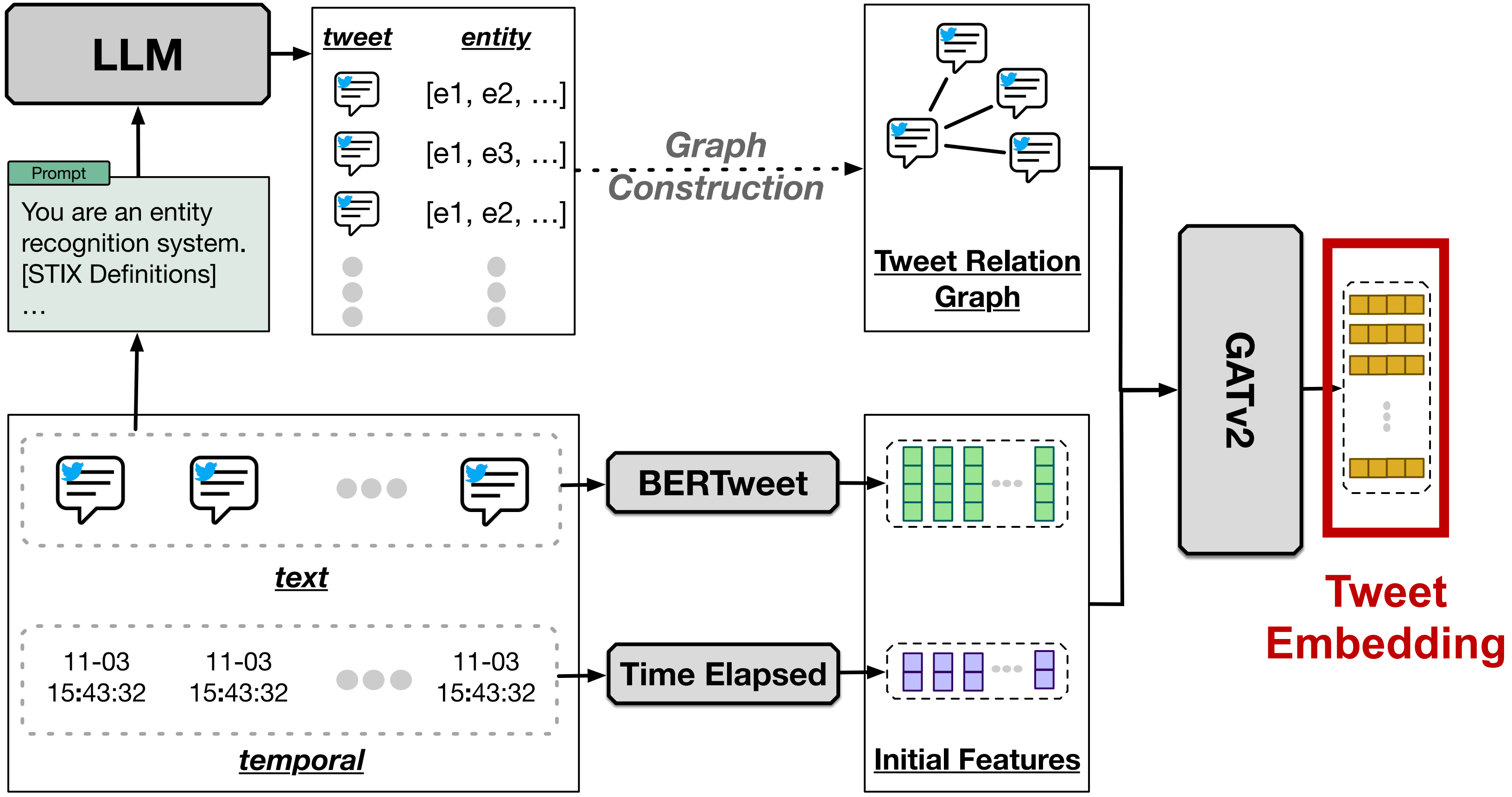
Tweet Embedding Workflow



Tweet Embedding Workflow



Tweet Embedding Workflow



Embedding Effectiveness

- Experimental Setup

Dataset

- **Time period:** 06/2022 – 02/2024
- **167 security events (2,054 tweets)** for training/validation/testing
- Distinct Time Periods for testing

Baselines

- **Text-Based Embedding:** TF-IDF, Word2Vec
- **PLM:** BERT, BERTweet, SecureBERT, LLaMA2
- **Graph-Based:** GCN, GATv2, GraphSAGE

Metrics

- Normalized Mutual Information **(NMI)**
- Adjusted Mutual Information **(AMI)**
- Adjusted Rand Index **(ARI)**

Embedding Effectiveness

- Experiment Results

	Model	Test-1 (2022.10)			Test-2 (2024.01)			Test-3 (2024.02)		
		AMI (↑)	**ARI (↑)**	**NMI (↑)**	**AMI (↑)**	**ARI (↑)**	**NMI (↑)**	**AMI (↑)**	**ARI (↑)**	**NMI (↑)**
Keyword	TF-IDF	0.3036	0.0552	0.5147	0.4559	0.0989	0.6150	0.4669	0.0941	0.6218
	Word2Vec	0.0463	0.0060	0.1135	0.2469	0.0389	0.3876	0.1380	0.0246	0.2367
PLM	BERT	0.2389	0.0203	0.4395	0.2671	0.0291	0.4544	0.2716	0.0264	0.4573
	BERTweet	0.3466	0.0676	0.5211	0.2777	0.0312	0.4618	0.1729	0.0143	0.3372
	SecureBERT	0.3046	0.0324	0.5020	0.2555	0.0259	0.4434	0.1927	0.0161	0.3763
	Llama2	0.1138	0.0090	0.3041	0.2324	0.0271	0.4375	0.2127	0.0254	0.4118
Graph	GCN	0.2806	0.0869	0.4455	0.3771	0.1253	0.5454	0.3335	0.1171	0.5111
	GATv2	0.3396	0.0988	0.5274	0.4065	0.1430	0.5476	0.3440	0.1112	0.5077
	GraphSAGE	0.3164	0.0912	0.5019	0.3666	0.1305	0.5350	0.3210	0.1048	0.4866
Our Embedding		**0.5919**	**0.3384**	**0.7344**	**0.6561**	**0.4470**	**0.7763**	**0.5950**	**0.3387**	**0.7404**

Embedding Effectiveness

- Experiment Results

	Model	Test-1 (2022.10)			Test-2 (2024.01)			Test-3 (2024.02)		
		AMI (↑)	**ARI (↑)**	**NMI (↑)**	**AMI (↑)**	**ARI (↑)**	**NMI (↑)**	**AMI (↑)**	**ARI (↑)**	**NMI (↑)**
Keyword	TF-IDF	0.3036	0.0552	0.5147	0.4559	0.0989	0.6150	0.4669	0.0941	0.6218
	Word2Vec	0.0463	0.0060	0.1135	0.2469	0.0389	0.3876	0.1380	0.0246	0.2367
PLM	BERT	0.2389	0.0203	0.4395	0.2671	0.0291	0.4544	0.2716	0.0264	0.4573
	BERTweet	0.3466	0.0676	0.5211	0.2777	0.0312	0.4618	0.1729	0.0143	0.3372
	SecureBERT	0.3046	0.0324	0.5020	0.2555	0.0259	0.4434	0.1927	0.0161	0.3763
	Llama2	0.1138	0.0090	0.3041	0.2324	0.0271	0.4375	0.2127	0.0254	0.4118
Graph	GCN	0.2806	0.0869	0.4455	0.3771	0.1253	0.5454	0.3335	0.1171	0.5111
	GATv2	0.3396	0.0988	0.5274	0.4065	0.1430	0.5476	0.3440	0.1112	0.5077
	GraphSAGE	0.3164	0.0912	0.5019	0.3666	0.1305	0.5350	0.3210	0.1048	0.4866
Our Embedding		**0.5919**	**0.3384**	**0.7344**	**0.6561**	**0.4470**	**0.7763**	**0.5950**	**0.3387**	**0.7404**

Embedding Effectiveness

- Experiment Results

	Model	Test-1 (2022.10)			Test-2 (2024.01)			Test-3 (2024.02)		
		AMI (↑)	ARI (↑)	NMI (↑)	AMI (↑)	ARI (↑)	NMI (↑)	AMI (↑)	ARI (↑)	NMI (↑)
Keyword	TF-IDF	0.3036	0.0552	0.5147	0.4559	0.0989	0.6150	0.4669	0.0941	0.6218
	Word2Vec	0.0463	0.0060	0.1135	0.2469	0.0389	0.3876	0.1380	0.0246	0.2367
PLM	BERT	0.2389	0.0203	0.4395	0.2671	0.0291	0.4544	0.2716	0.0264	0.4573
	BERTweet	0.3466	0.0676	0.5211	0.2777	0.0312	0.4618	0.1729	0.0143	0.3372
	SecureBERT	0.3046	0.0324	0.5020	0.2555	0.0259	0.4434	0.1927	0.0161	0.3763
	Llama2	0.1138	0.0090	0.3041	0.2324	0.0271	0.4375	0.2127	0.0254	0.4118
Graph	GCN	0.2806	0.0869	0.4455	0.3771	0.1253	0.5454	0.3335	0.1171	0.5111
	GATv2	0.3396	0.0988	0.5274	0.4065	0.1430	0.5476	0.3440	0.1112	0.5077
	GraphSAGE	0.3164	0.0912	0.5019	0.3666	0.1305	0.5350	0.3210	0.1048	0.4866
Our Embedding		**0.5919**	**0.3384**	**0.7344**	**0.6561**	**0.4470**	**0.7763**	**0.5950**	**0.3387**	**0.7404**

Embedding Effectiveness

- Experiment Results

	Model	Test-1 (2022.10)			Test-2 (2024.01)			Test-3 (2024.02)		
		AMI (↑)	**ARI (↑)**	**NMI (↑)**	**AMI (↑)**	**ARI (↑)**	**NMI (↑)**	**AMI (↑)**	**ARI (↑)**	**NMI (↑)**
Keyword	TF-IDF	0.3036	0.0552	0.5147	0.4559	0.0989	0.6150	0.4669	0.0941	0.6218
	Word2Vec	0.0463	0.0060	0.1135	0.2469	0.0389	0.3876	0.1380	0.0246	0.2367
PLM	BERT	0.2389	0.0203	0.4395	0.2671	0.0291	0.4544	0.2716	0.0264	0.4573
	BERTweet	0.3466	0.0676	0.5211	0.2777	0.0312	0.4618	0.1729	0.0143	0.3372
	SecureBERT	0.3046	0.0324	0.5020	0.2555	0.0259	0.4434	0.1927	0.0161	0.3763
	Llama2	0.1138	0.0090	0.3041	0.2324	0.0271	0.4375	0.2127	0.0254	0.4118
Graph	GCN	0.2806	0.0869	0.4455	0.3771	0.1253	0.5454	0.3335	0.1171	0.5111
	GATv2	0.3396	0.0988	0.5274	0.4065	0.1430	0.5476	0.3440	0.1112	0.5077
	GraphSAGE	0.3164	0.0912	0.5019	0.3666	0.1305	0.5350	0.3210	0.1048	0.4866
Our Embedding		**0.5919**	**0.3384**	**0.7344**	**0.6561**	**0.4470**	**0.7763**	**0.5950**	**0.3387**	**0.7404**

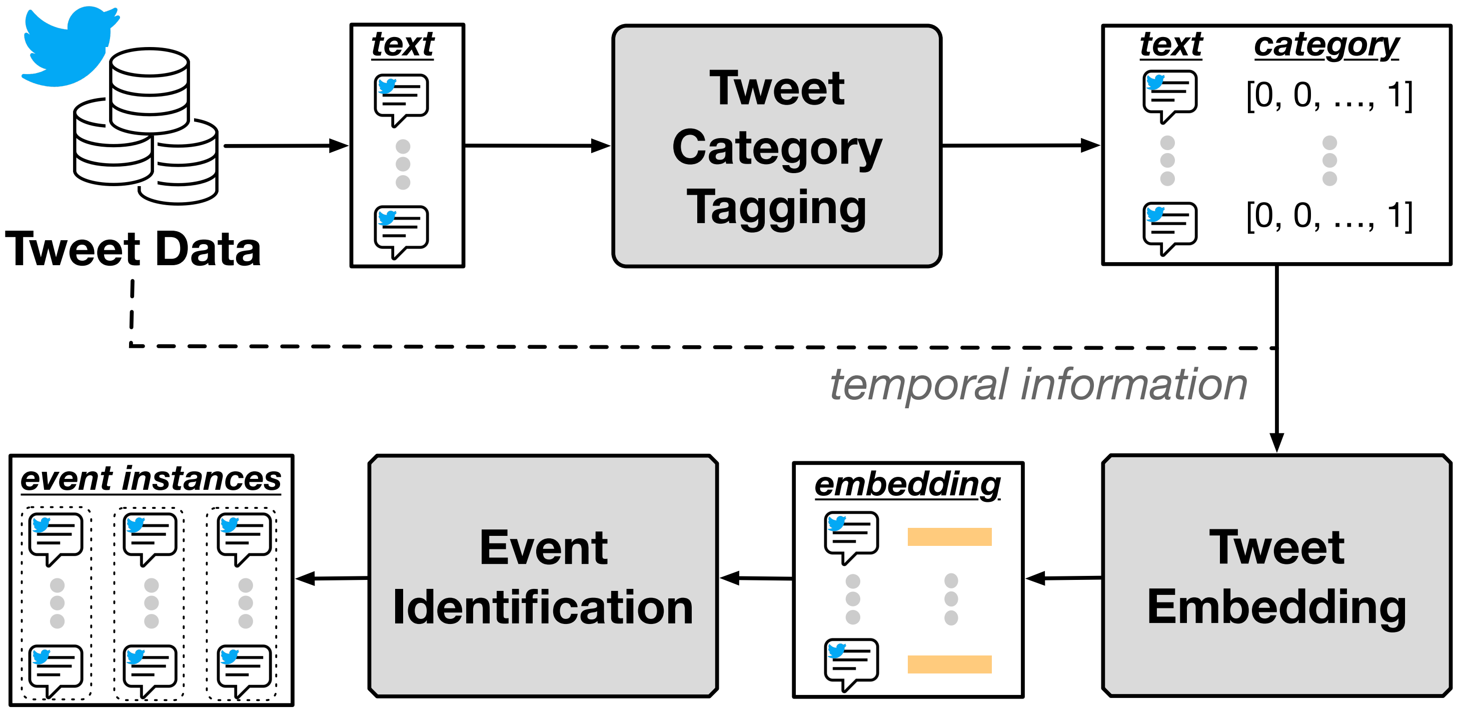
**Further trained on Twitter and Security corpus*

Embedding Effectiveness

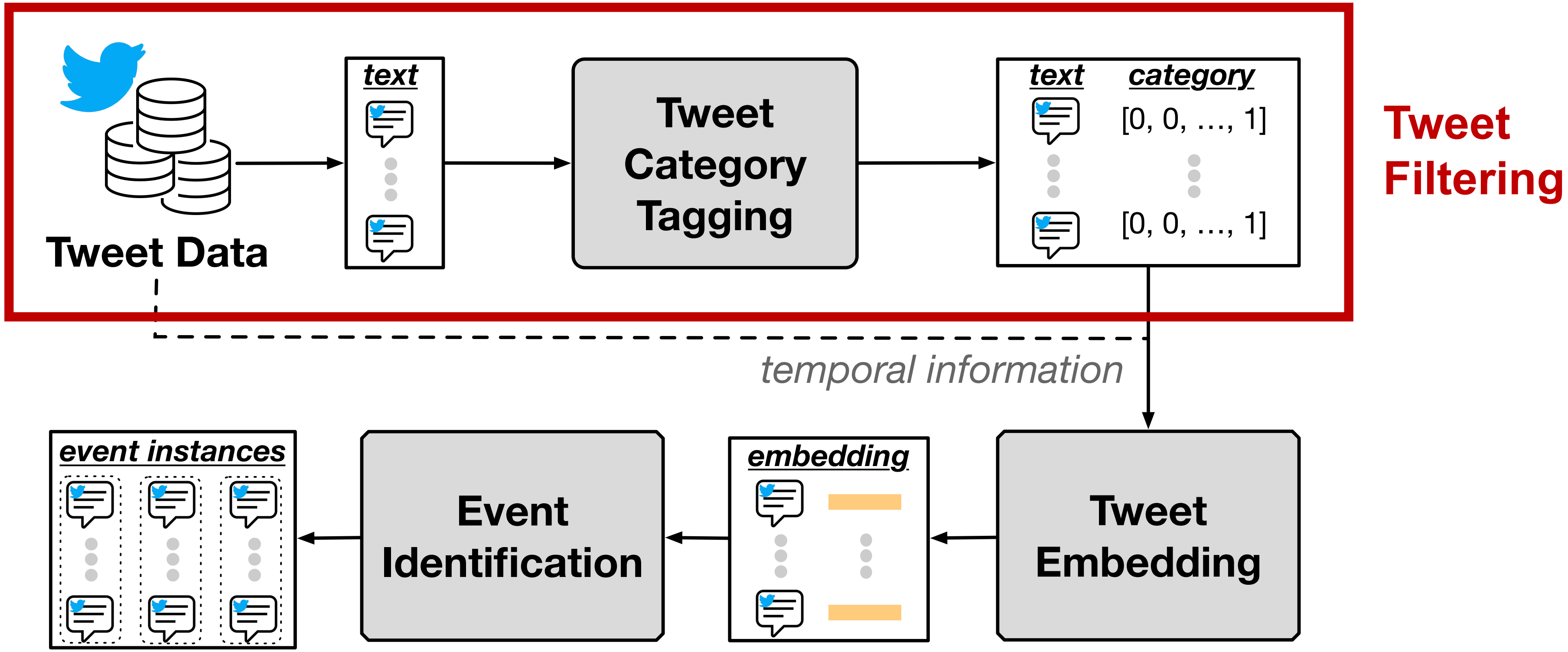
- Experiment Results

	Model	Test-1 (2022.10)			Test-2 (2024.01)			Test-3 (2024.02)		
		AMI (↑)	**ARI (↑)**	**NMI (↑)**	**AMI (↑)**	**ARI (↑)**	**NMI (↑)**	**AMI (↑)**	**ARI (↑)**	**NMI (↑)**
Keyword	TF-IDF	0.3036	0.0552	0.5147	0.4559	0.0989	0.6150	0.4669	0.0941	0.6218
	Word2Vec	0.0463	0.0060	0.1135	0.2469	0.0389	0.3876	0.1380	0.0246	0.2367
PLM	BERT	0.2389	0.0203	0.4395	0.2671	0.0291	0.4544	0.2716	0.0264	0.4573
	BERTweet	0.3466	0.0676	0.5211	0.2777	0.0312	0.4618	0.1729	0.0143	0.3372
	SecureBERT	0.3046	0.0324	0.5020	0.2555	0.0259	0.4434	0.1927	0.0161	0.3763
	Llama2	0.1138	0.0090	0.3041	0.2324	0.0271	0.4375	0.2127	0.0254	0.4118
Graph	GCN	0.2806	0.0869	0.4455	0.3771	0.1253	0.5454	0.3335	0.1171	0.5111
	GATv2	0.3396	0.0988	0.5274	0.4065	0.1430	0.5476	0.3440	0.1112	0.5077
	GraphSAGE	0.3164	0.0912	0.5019	0.3666	0.1305	0.5350	0.3210	0.1048	0.4866
Our Embedding		**0.5919**	**0.3384**	**0.7344**	**0.6561**	**0.4470**	**0.7763**	**0.5950**	**0.3387**	**0.7404**

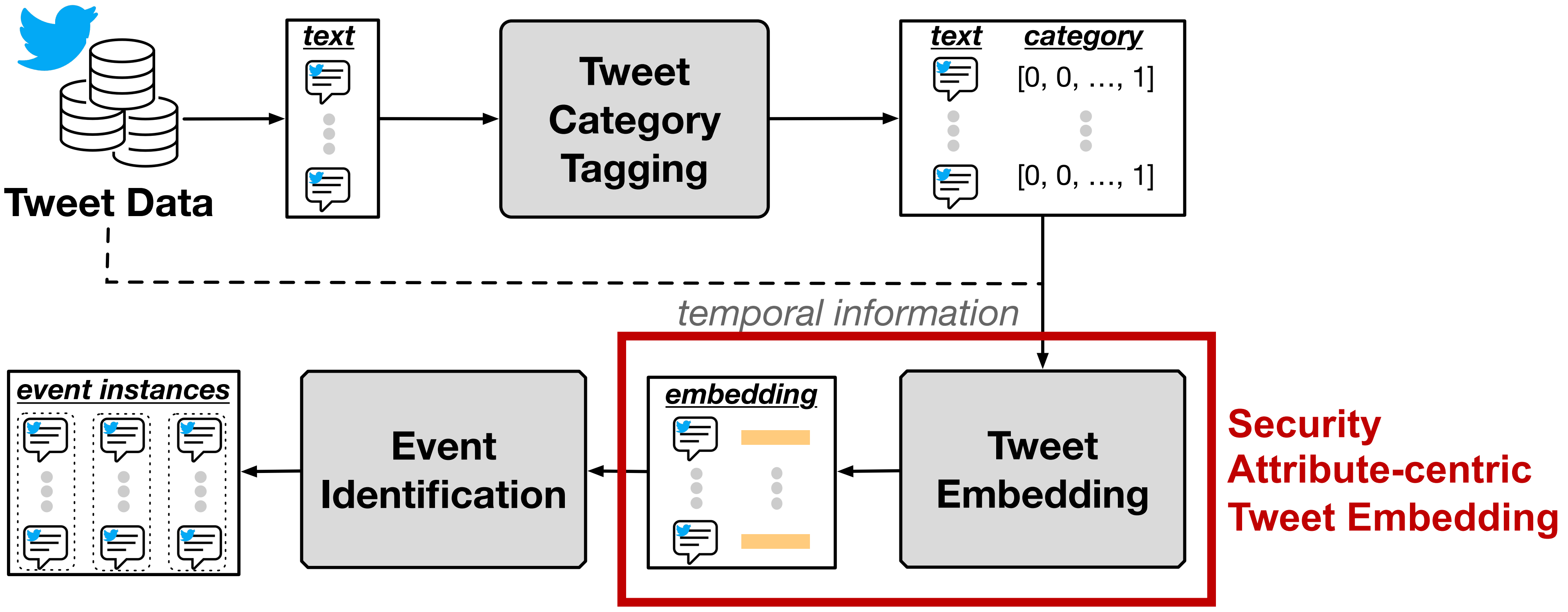
Tweezers: A Framework for Security Event Detection



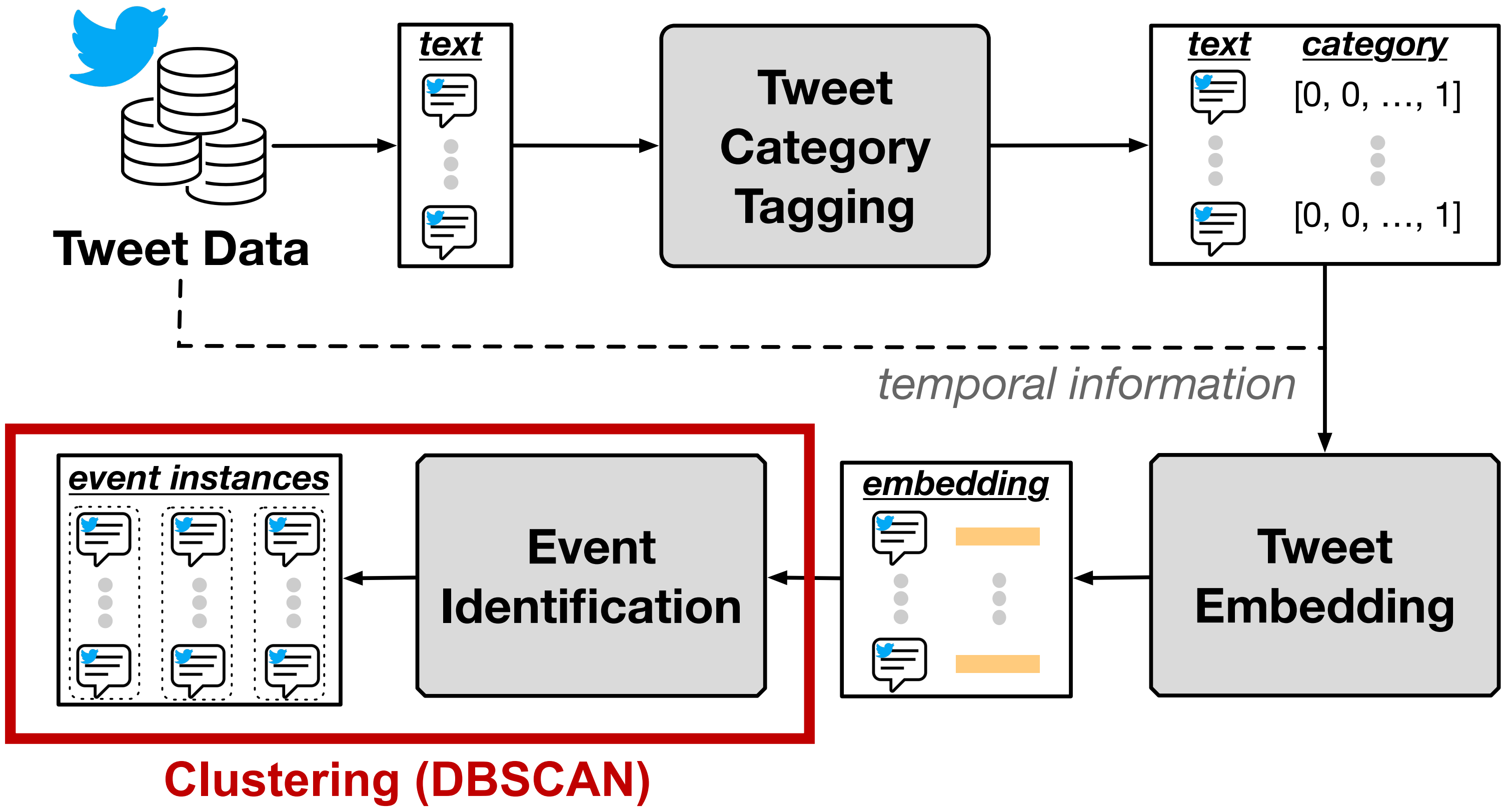
Tweezers: A Framework for Security Event Detection



Tweezers: A Framework for Security Event Detection

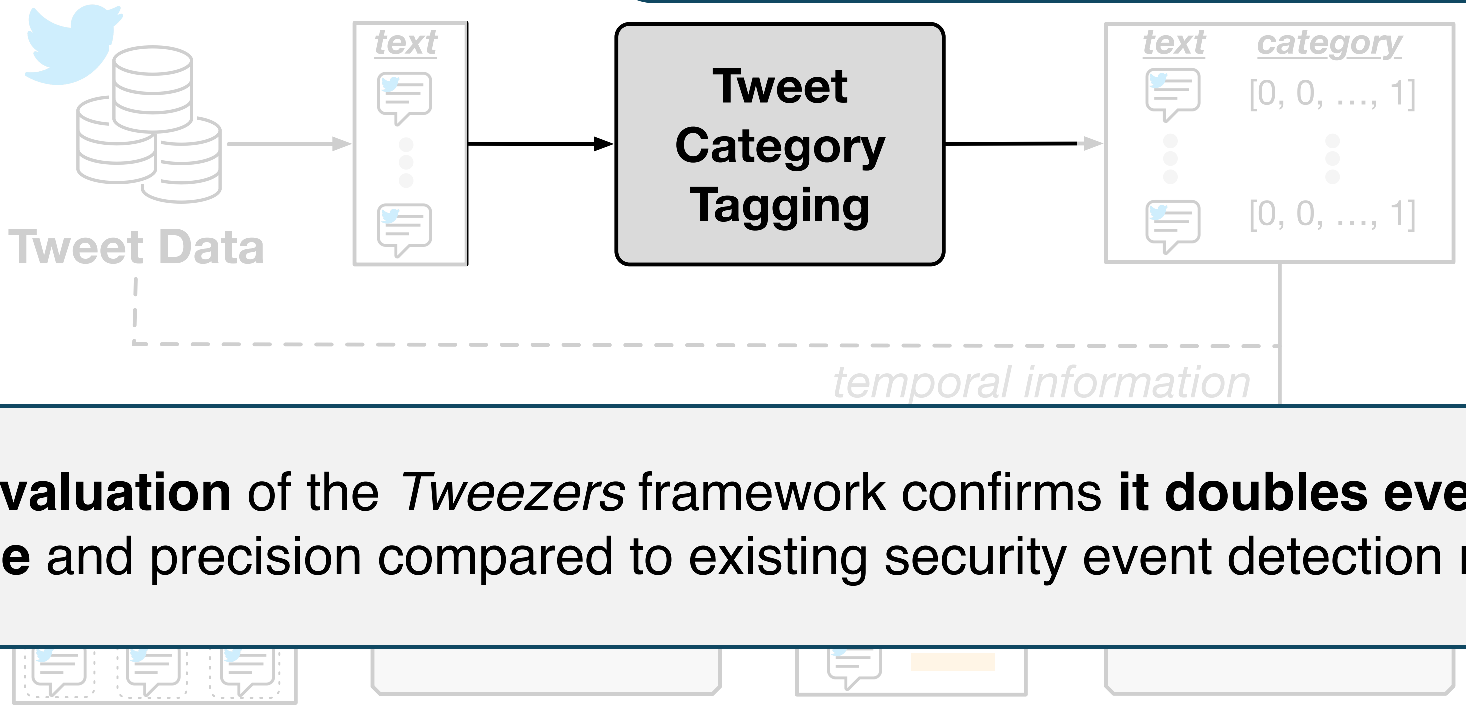


Tweezers: A Framework for Security Event Detection



More in the paper

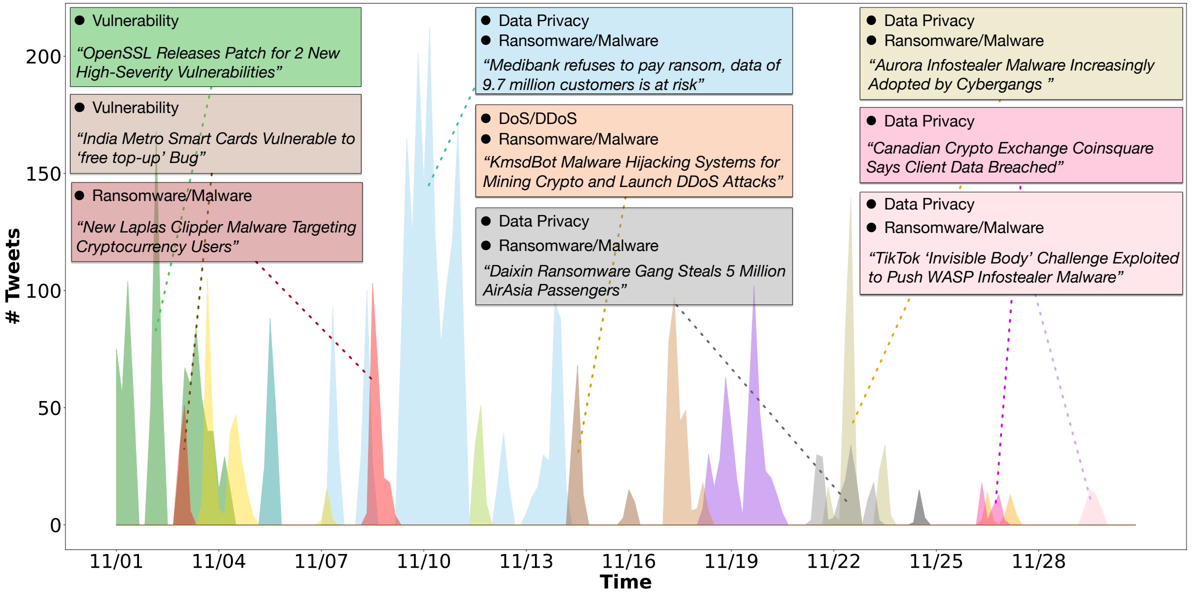
- Introduced seven categories to enhance correctness of labeling and tweet classification process.
- Enabled multi-category classification for accurate event categorization.



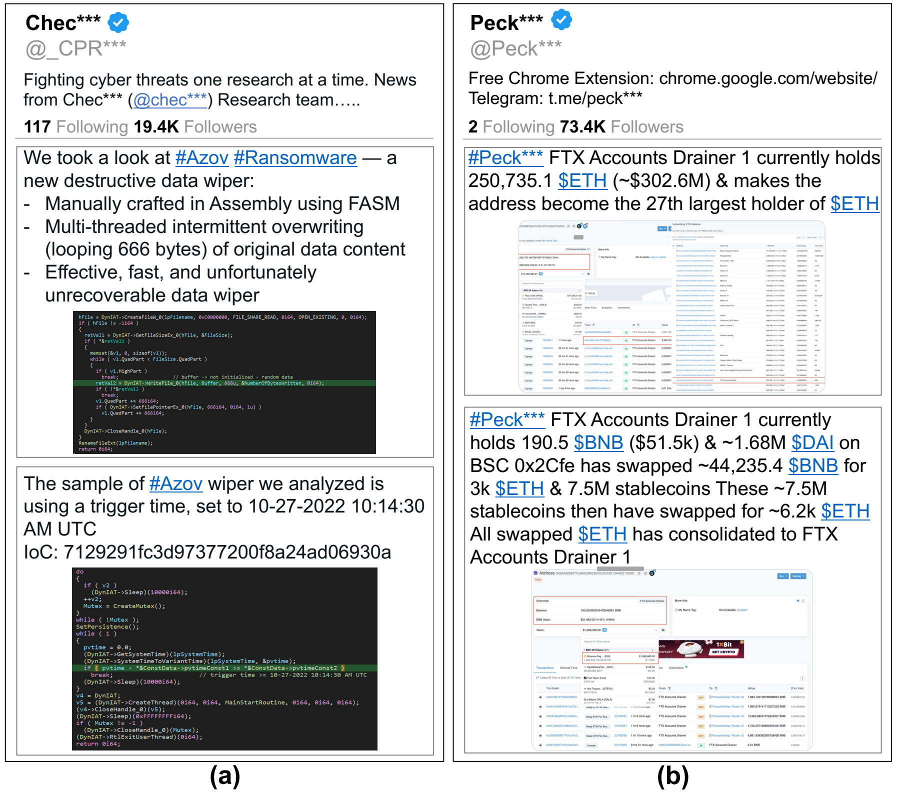
End-to-end evaluation of the *Tweezers* framework confirms **it doubles event detection coverage** and precision compared to existing security event detection method.

More in the paper

- Two use cases of *Tweezers*



Security Event Trend Analysis



Finding Informative Security Users

Summary

- *Introduced a novel event attribution-centric tweet embedding method for precise and comprehensive security event detection.*

- *Developed the Tweezers framework, which outperforms baselines by doubling event detection coverage and precision.*

- *Demonstrated real-world applications in security trend analysis and identifying informative security users.*

Paper

Code

Thank You!

INDIANA UNIVERSITY

KAIST

S2W