Misinformation Detection with AIP

Tanish Tyagi

Cornell University

April 26, 2025

Problem Statement

- College students heavily rely on fast, unverified information from social media, messaging apps, and news feeds
- Rapid rumor spread about campus safety, health, or politics can heighten tensions, cause disruptions, and erode trust in leadership
- University administrators face growing risks and need a comprehensive, data-driven response to misinformation
- Centralizing social media feeds, official announcements, and threat intelligence enables early rumor detection and timely corrections
- Machine learning can identify patterns of false information and flag recurring sources for appropriate intervention

Demo Scenario

- In October 2023, violent antisemitic threats targeting Jewish students and locations at Cornell University caused widespread fear and disruption
- Cornell faced heavy criticism for delayed and limited communication—taking over three days to respond and failing to issue a university-wide alert
- A real-time monitoring system would have immediately flagged the threats, alerting administrators and security teams for faster action
- The system can also address other urgent campus issues (e.g., health hazards) and be scaled to other universities

Data Source 1 of 3*: Social Media Posts

- 3,000 Cornell-related posts from platforms like Twitter and Reddit during the last week of October 2023
- Preprocessed using a pipeline builder to clean text and extract relevant entities (e.g., sentiment, location)
- Cleaned data was input into a "Use LLM" block to identify posts making threats or related to threats

LATITUDE	LONGITUDE	PLATFORM	POST ID	TEXT	TIMESTAMP	USER ID
-39.9665335	-7.820518	twitter	P000143	Is it true Morrison switched to p…	Oct 29, 2023, 9:46 AM	12d54cc5-63ed-4988-b207-
42.45767057829…	-76.47627736750…	twitter	P000218	Is the CTB line ever <30 min? Ser…	Oct 29, 2023, 4:23 PM	d6d6beb7-ff47-4c4c-9710-
41.674434	127.778911	reddit	P000176	Count down: 3 hours until I gas l…	Oct 29, 2023, 12:22 PM	71a3cd7e-712b-40fc-8756-
59.325304	36.025562	reddit	P000057	Count down: 3 hours until I gas l…	Oct 30, 2023, 9:14 AM	827f8098-94ba-4859-a5d4-

Figure: Social Media Post Ontology

*All data for this project has been pro-grammatically generated based on real-world patterns

Data Source 2 of 3: Social Media Accounts

- 1000 Cornell-related accounts from social media websites that had posted in last week of October 2023
- Data was imputed and standardized using Code repository
- Logistic Regression classifier was regressed on features to identify whether an account should be flagged as propagating misinformation
- Model deployed to Foundry using Adapters so it could be integrated into Workshop application

ACCOUNT ID	PRIMARY PLATFORM	ACCOUNT AGE DAYS	VERIFIED FLAG	FOLLOWERS CNT	BASE POST RATE	PRIOR FLAGS CNT	STRIKE COUNT	AVG DAILY LIKES	AVG DAILY RESHARES	AVG DAILY COMMENTS
445	2	166	0	41	1.576586570233…	0	0	18	5	4
895	2	317	1	251	0.681437142045…	0	0	21	1	24
485	3	446	1	2,949	1.269867012615…	0	1	128	3	10
210	3	1,644	1	180	1.132546619622…	0	0	32	2	16
730	2	59	1	1,338	1.211329107388…	0	0	37	10	6
734	2	697	0	537	1.557805172689…	0	0	18	6	45
788	4	494	1	132	0.804191602449…	0	0	14	3	3
109	0	67	1	65	2.711699976819…	0	0	453	32	12
631	1	357	1	117	1.349115920597…	1	0	17	2	4

Figure: Social Media Account Ontology

Data Source 3 of 3: Police Reports

- 150 police reports made to Cornell University Police Department (CUPD) in last week of October 2023
- Preprocessed using pipeline builder to clean text
- Cleaned data + output from LLM in Data Source 1 went into "Use LLM" block, which analyzed police reports to see if any threats made on social media could be corroborated

CAD EVENT ID	CALL TEXT	GEO LAT	GEO LON	INCIDENT TYPE
CU-20231029-022	Information reported at Kosher …	42.4498	-76.4821	Information
CU-20231029-126	Assistance Request reported at …	42.4498	-76.4821	Assistance Request
CU-20231029-042	Suspicious Activity reported at D…	42.4448	-76.4847	Suspicious Activity
CU-20231029-008	Suspicious Activity reported at D…	42.4448	-76.4847	Suspicious Activity
CU-20231029-148	Resolved reported at North Cam…	42.4551	-76.4729	Resolved
CU-20231029-038	Information reported at Anabel …	42.4493	-76.4835	Information
CU-20231029-048	Follow-up patrol related to earli…	42.4498	-76.4821	Medical Aid

Figure: Police Report Ontology

Next Steps

- This platform serves as a proof-of-concept for a Misinformation Detection system that can provide university administrators with the ability to issue campus-wide corrections during situations that could impact university operations
- To further improve the accuracy of the system and the scope of issues to monitor, additional data sources can be incorporated. For example, Threat-intelligence OSINT feeds, student newspapers, and anonymous apps like Sidechat
- This system can be scaled for other universities and educational institutions, as well as for local towns and city administrators

Github:

https://github.com/anaconda121/Semester-In-Palantir