(Dis)information Wars

Adrian Casillas Maryam Farboodi Layla Hashemi

Maryam Saeedi Steven Wilson

August 14, 2024

Disinformation and Propaganda

- Disinformation is rampant
- A major tool of oppressive regimes
- Social media has increased the power of disinformation and propaganda
    - Accounts can hide their allegiance
    - Speed of transmission is much higher

Disinformation and Propaganda

- Disinformation is rampant
- A major tool of oppressive regimes
- Social media has increased the power of disinformation and propaganda
  - Accounts can hide their allegiance
  - Speed of transmission is much higher
- **How to find disinformation and combat its effects?**

Disinformation and Propaganda

- Disinformation is rampant
- A major tool of oppressive regimes
- Social media has increased the power of disinformation and propaganda
    - Accounts can hide their allegiance
    - Speed of transmission is much higher
- **How to find disinformation and combat its effects?**
- Limitations of current approaches:
    - Real-time content moderation is slow
    - Ex-post debunking has limited impact

Our Approach

We propose a network-based methodology to identify disinformation:

1. Classify accounts using network characteristics
2. Label news based on initiators' categories

Key advantages:

- Early detection before viral spread
- Uses limited data from first few initiators
- Effective without continuous updates

Data

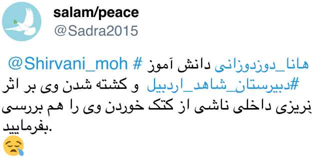
- Twitter/X data from "Woman, Life, Freedom" protests in Iran
- Period: September 2022 - March 2023
- 9.5 million accounts, 1.7 million active accounts
  - At least 10% of their engagements in Farsi
  - At least 10 engagements in Farsi from September 2022
- Complete network construction of active accounts
  - Following
  - Follower
  - Repost
  - Reposted

Categories of Account

- We classify all these active farsi accounts into 3 groups
  - Ordinary: normal accounts
  - Unsafe: accounts that actively participate or will participate in dispersion of disinformation
  - Pro-regime: accounts that are openly promoting the regime's propaganda

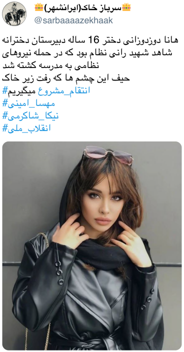
An Example: Disinformation

- Hana Douzdouzani was killed by regime (this did not happen!)



salam/peace
@Sadra2015

@Shirvani_moh #هانا_دوزدوزانی دانش آموز
#دبیرستان_شاهد_اردبیل و کشته شدن وی بر اثر
نریزی داخلی ناشی از کتک خوردن وی را هم بررسی
بفرمایید.
😪

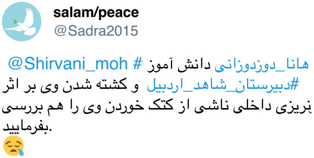


👑سرباز خاک(ایرانشهر)👑
@sarbaaaazekhaak

هانا دوزدوزانی دختر 16 ساله دبیرستان دخترانه
شاهد شهید رانی نظام بود که در حمله نیروهای
نظامی به مدرسه کشته شد
حیف این چشم ها که رفت زیر خاک
#انتقام_مشروع میگیریم
#مهسا_امینی
#نیکا_شاکرمی
#انقلاب_ملی

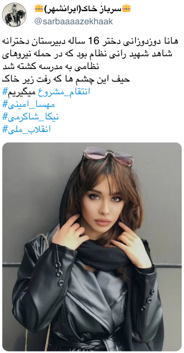
An Example: Disinformation

- Hana Douzdouzani was killed by regime (this did not happen!)



salam/peace
@Sadra2015

@Shirvani_moh # هانا_دوزدوزانی دانش آموز
#دبیرستان_شاهد_اردبیل و کشته شدن وی بر اثر
نریزی داخلی ناشی از کتک خوردن وی را هم بررسی
بفرمایید.
😪



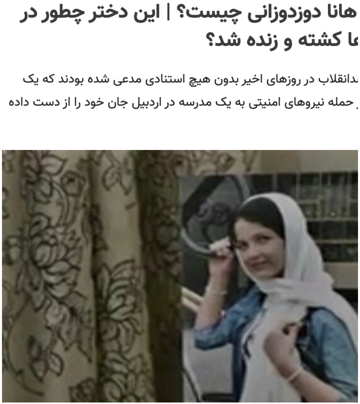
👑سرباز خاک(ایرانشهر)👑
@sarbaaaazekhaak

هانا دوزدوزانی دختر 16 ساله دبیرستان دخترانه
شاهد شهید رانی نظام بود که در حمله نیروهای
نظامی به مدرسه کشته شد
حیف این چشم ها که رفت زیر خاک
#انتقام_مشروع میگیریم
#مهسا_امینی
#نیکا_شاکرمی
#انقلاب_ملی

- Why this can be good for the regime?

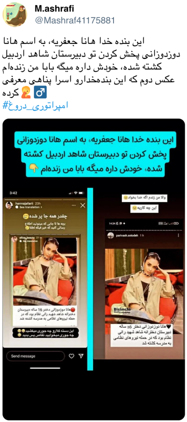
Benefit: Rebuttal

- Rebuttal done in a few days
  - government back account/government news agencies/forced confessions
- Benefit: discrediting opposition
  - claiming all killed were similarly fake news



هانا دوزدوزانی چیست؟ | این دختر چطور در
ا کشته و زنده شد؟

دانقلاب در روزهای اخیر بدون هیچ استنادی مدعی شده بودند که یک
حمله نیروهای امنیتی به یک مدرسه در اردبیل جان خود را از دست داده



M.ashrafi
@Mashraf41175881

این بنده خدا هانا جعفریه، به اسم هانا
دوزدوزانی پخش کردن تو دبیرستان شاهد اردبیل
کشته شده، خودش داره میگه بابا من زنده‌ام
عکس دوم که این بنده‌خدارو اسرا پناهی معرفی
کرده 🤦 ♂
#امپراتوری_دروغ

این بنده خدا هانا جعفریه، به اسم هانا دوزدوزانی
پخش کردن تو دبیرستان شاهد اردبیل کشته
شده، خودش داره میگه بابا من زنده‌ام 👇

3:42
hannajafarii 23h
See translation >

چقدر همه جا پر شده 🥲

بچه ها تا جایی که میتونید اطلاع
رسانی کنید که خبر فیکه لطفا 🥲

eliza.beauty 21 m
See translation

🔴هانا دوزدوزانی دختر 16 ساله دبیرستان
دخترانه شاهد شهید رائی نظام بود که در
حمله نیروهای نظامی به مدرسه کشته شد

این دسته کلارو چه جوری میکشید 😢😢
چه جوری میخوابید تقاص پس بدید 😢

Send message

والا من زندم اگه خدا بخواد 😐

این چه کاریه 😐

14:34
parivash.sotodeh 1 h

Bisimchi
Media

🖤هانا دوزدوزانی دختر 16 ساله
دبیرستان دخترانه شاهد شهید رائی
نظام بود که در حمله نیروهای نظامی
به مدرسه کشته شد

Labeled Set

- Total of 1,435 labeled accounts:
  - 476 unsafe accounts, 14 disinformation campaigns
  - 489 ordinary accounts
  - 470 pro-regime accounts
- All verified by 3 independent journalists

Labeled Set

- Total of 1,435 labeled accounts:
  - 476 unsafe accounts, 14 disinformation campaigns
  - 489 ordinary accounts
  - 470 pro-regime accounts
- All verified by 3 independent journalists
- Labeled Set divided into Training (70%) and Test (30%) groups

Network Proximity Measure

- Constructed network proximity measures:
  - Following
  - Follower
  - Repost
  - Reposted
- For both unsafe and pro-regime accounts
- Proximity scores capture network structure and how close accounts are to labeled training accounts within each set

Account Classification

- Multinomial logistic regression
- Elastic net regularization
- Features:
  - Network proximity measures
  - Basic account characteristics
  - Variables related to activities
- Outputs propensity scores for each category
- Classify an account to the category with highest propensity score

Account Classification Results and Highlights

- The elastic net keeps all network proximity measures
- Shows that the accounts within each category are highly connected
- Unsafe accounts try to follow ordinary accounts but they cannot get a lot of followback from them

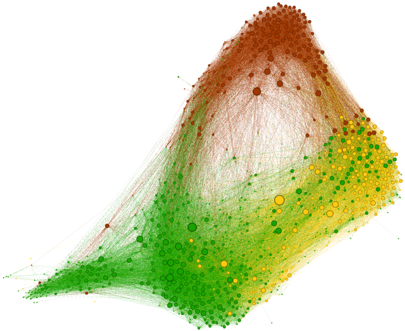
Account Classification Results and Highlights

- The elastic net keeps all network proximity measures
- Shows that the accounts within each category are highly connected
- Unsafe accounts try to follow ordinary accounts but they cannot get a lot of followback from them
- 95.13% overall accuracy on test set
- Precision and Sensitivity:
  - Ordinary: 92.6% / 93.9%
  - Unsafe: 93.6% / 92.3%
  - Pro-regime: 99.3% / 99.3%

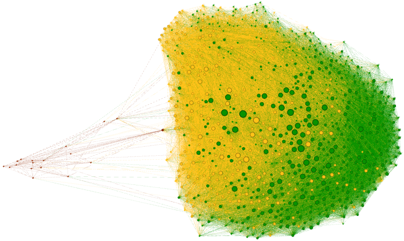
Account Distribution

- 16% of active accounts classified as unsafe
- 8% classified as pro-regime
- 76% classified as ordinary

Network Comparison of 1,000 Most Active Farsi-Twitter Accounts



Follower-Following Network



Retweets Network

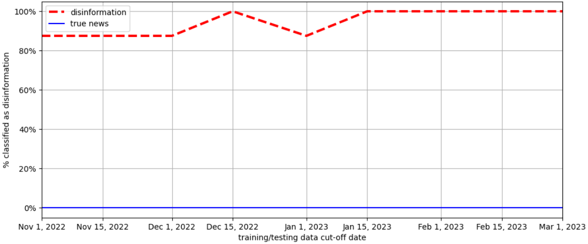
Green: Ordinary ; Yellow: Unsafe ; Red: Pro-regime

Disinformation Labeling

- Analyze first 10 initiators of a news event
  - Retrain the algorithm using only the first 6 disinformation campaigns
  - Include 10 real news events
  - The last 8 out of our 14 disinformation campaigns
- Label as disinformation if 7 or more classified as unsafe
- Balances early detection with accuracy

Disinformation Labeling

- Analyze first 10 initiators of a news event
  - Retrain the algorithm using only the first 6 disinformation campaigns
  - Include 10 real news events
  - The last 8 out of our 14 disinformation campaigns
- Label as disinformation if 7 or more classified as unsafe
- Balances early detection with accuracy
- Effective even with limited training data

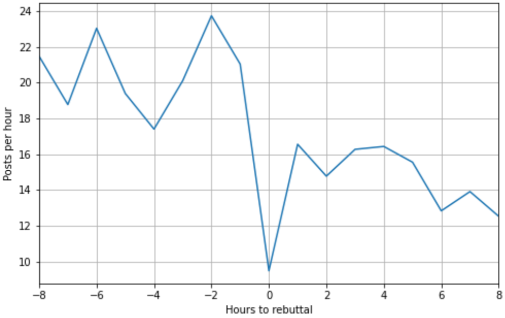


Impact of Rebuttal on Disinformation Spread

- We use our data once more to get an estimate of the effect of rebuttals
- For each disinformation, we tag all the events of rebuttals by other users, either as new tweet or as a response to previous ones.
- Run a diff-in-diff to estimate the impact of rebuttals
- Modeled as Poisson process with varying mean

Impact of Rebuttal on Disinformation Spread

- We use our data once more to get an estimate of the effect of rebuttals
- For each disinformation, we tag all the events of rebuttals by other users, either as new tweet or as a response to previous ones.
- Run a diff-in-diff to estimate the impact of rebuttals
- Modeled as Poisson process with varying mean
- Rebuttals significantly reduce disinformation spread
- Rebuttal leads up to ~80% decline in post volume



Impact on Disinformation Spread

Estimated effects of implementing our approach:

- 3x reduction in number of original posts
- 2x reduction in maximum user engagement rate
- At least 2x reduction in effective lifespan of disinformation

Robustness

- Reducing manipulation by only using network variables
  - Will decrease the accuracy rate slightly
  - The detection of disinformation campaigns varies between 75% and 100% depending on data length
- Value of training data
  - Increasing training data improves the algorithm greatly, but adding length of data to improve variables have minimal impact
- Expert validation
  - Three independent journalists assessed account categories
  - Model outperformed human experts (93% vs 87% accuracy)
  - Journalists often uncertain (65% of cases)
  - Strong correlation between model and expert assessments

Concluding Remarks

- **Disinformation wars**
    - Intentional spread of disinformation on social media platforms
- Propose network-based approach for early disinformation detection
- High effectiveness on real-world data
- Has significant potential to mitigate disinformation spread

Thank You!

Score Manipulation

- Account characteristics can be susceptible to manipulation
- Right up imposters' alley!
- Make policy design using scores hard
- Which characteristics are more difficult to manipulate?
  - Network-based characteristics!

Non-Manipulable Classifier

	Ordinary		Unsafe		Propaganda	
	coef	std err	coef	std err	coef	std err
log(Unsafe Followers Measure)	-2.64	0.54	4.06	0.67	-0.42	0.73
log(Unsafe Following Measure)	1.84	0.50	0.00	0.68	-1.23	0.71
log(Unsafe Retweets Measure)	1.02	0.38	0.00	0.46	-1.61	0.69
log(Unsafe Retweeted Measure)	-2.02E-03	0.40	2.52	0.53	-1.53	0.96
log(Propaganda Followers Measure)	-0.95	0.95	-0.35	0.71	2.29	0.95
log(Propaganda Following Measure)	0.00	0.76	-2.81	0.83	2.46	0.90
log(Propaganda Retweets Measure)	-0.70	0.81	-0.41	1.32	2.12	0.47
log(Propaganda Retweeted Measure)	-1.48	0.96	0.00	0.43	1.25	0.57
Degree Followers Centrality	0.76	4.51	0.00	6.23	-0.17	4.43
Eigen Followers Centrality	-3.02	2.17	-1.90	2.30	5.91	0.91
No. Observations:	556					
Log-Likelihood:	-0.23					
Pseudo R-squ.:	0.85					

Non-Manipulable Classifier

type I-II error trade-off

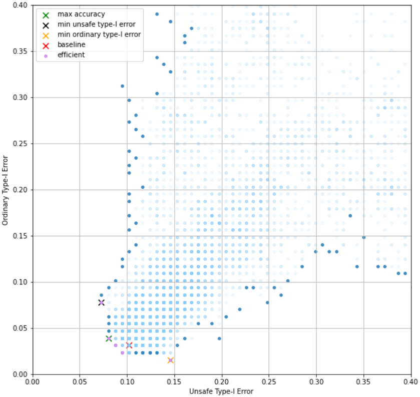
- performance: confusion matrix

$$\begin{bmatrix} 114 & 14 & 0 \\ 25 & 112 & 0 \\ 1 & 0 & 107 \end{bmatrix}$$

Table: Confusion matrix for classifier non-manipulable classifier
Total accuracy: 89.28%

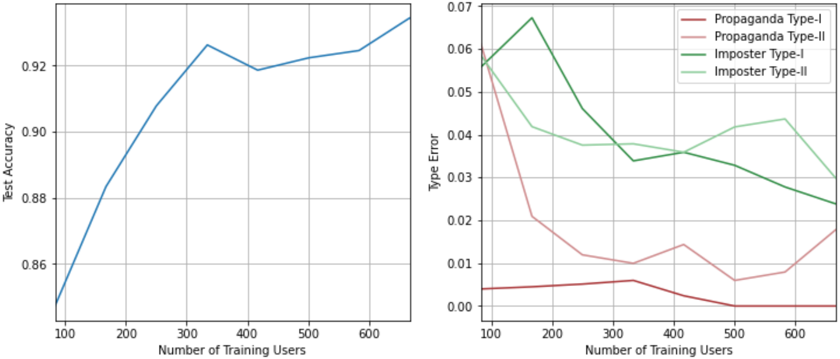
- accuracy
  - 90% total
  - $\sim$ 100% propaganda accounts

Trade-off Between Type I and Type II Error: Ordinary and Unsafe



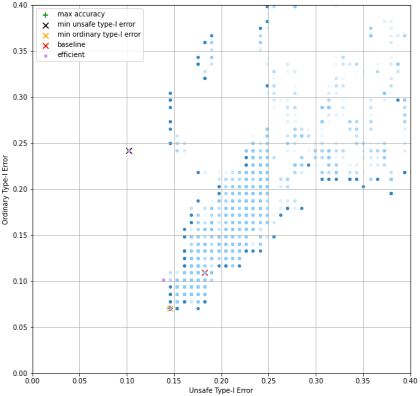
back

Total Classification Accuracy as a Function of Training Sample Size



back

Trade-off Between Type I and Type II Error: Unsafe Ordinary Non-Manipulable Classifier



back

Decreasing Ordinary Type I Error

- If an account is ordinary, most of their first followers will also be ordinary
- It is hard for imposter accounts to convince real people to start following them when they just join

	Baseline	0.5	0.6	0.7	0.8	0.9
Ordinary Type I	3.1%	1.6%	1.6%	2.3%	3.1%	3.1%
Unsafe Type I	10.2%	13.9%	12.4%	11.6%	10.9%	10.2%

Table: Effect of reclassifying the accounts classified as unsafe as ordinary based on the share of their initial ordinary followers on ordinary and unsafe type I errors.