Ephemeral Astroturfing Attack: The Case of Fake Twitter Trends

Tuğrulcan Elmas (tugrulcan.elmas@epfl.ch)
Rebekah Overdorf (rebekah.overdorf@epfl.ch)
Ahmed Özkalay (afozkalay@gmail.com)
Karl Aberer (karl.aberer@epfl.ch)
EPFL
Problem

- Novel Attack: Ephemeral Astroturfing
  - Goal is to create a trend (traditional astroturfing)
  - Inauthentic actions are removed with no effect on the impact

- Systematic manipulation of Twitter trends in Turkey by bots since 2015

- Unaddressed
  - No prior study
  - Media reports
  - Twitter takedown in 2020
  - No action despite two disclosures
Ephemeral Astroturfing

- Target: A voting based «What’s popular?» system
  - Twitter trends, Instagram Explore

- Coordinated voting by inauthentic accounts having impact (astroturfing)

- Actions are removed quickly, impact continues (ephemerality)
  - Time-of-Check-Time-of-Use problem
Case of Fake Twitter Trends

- Target: Twitter Trends
  - What’s popular **now**

- Coordinated posting & deleting using bots

- Premade / generated tweets plus **the target keyword**

- Compromised accounts as bots
Attack Behaviour: Coordinated Voting & Deletions

EURO S&P 2021 | EPHEMERAL ASTROTURFING ATTACK: THE CASE OF FAKE TWITTER TRENDS

Tuğrulcan Elmas

[Graphs showing the timeline of tweets, deletions, and retweets for two hashtags: #SuriyelilerDefolsun and #SamsununAdresiMacellanCafe.]
Attack Behaviour: Generated Tweets

- Content that would **pass spam filters**
  - Not necessarily the Turing test
  - Target keyword + random words
  - Infrequent and not repetitive
  - No other pattern observed
Attack Behaviour: Compromised Accounts

- Compromised accounts
  - Genuine activity except trend manipulation
  - Users are unaware or helpless
- Phishing websites on Twitter trends
  - Gain free followers
  - See who look at your profile
- Signed up a honeypot account
  - 560 attacks in 5 months
- Pass Quality Filters
Classifying Fake Trends

- Data Collection Challenge:
  - We do not know what is trending before it becomes trending

- Solution: 1% Realtime Sample – Provided by archive.org
  - Classify generated tweets with a rule based method
  - A trend is attacked if contained in at least x such tweets
  - x is 4 in 1% sample
Impact on Top Trends

- Same attack since 2015 July
- %47 of top 5 trends in Turkey are poisoned
- %20 of top 10 trends in the World are poisoned

Bars: Top 10
Lines: Top 5
- 108,000 bots active in 2015-2019
  - %45 of them are suspended or deleted now

- 33,000 still active in 2019

- 6 communities
  - 2 communities taken down

- Same community of bots promote different political parties
What are the Trends About?

- Over 6,500 fake trends in 2019

- Illicit Advertisements (32%)
  - Phishing & Trend Service
  - Gambling

- Politics (12%)
  - Political Slogans
  - Disinformation Campaigns

- Appeals to Government (18%)

- Religious Cults (9%)

- Misc (26%)
  - Marriage Proposals
Implications

- Data Integrity
  - Hides what made the topic popular
  - Poisoned many political science studies investigating #SyriansGetOut

- Security
  - Trending algorithm
  - Market for compromised accounts

- Politics
  - Disinformation & hate speech
  - Twitter as a fake proxy
What Now?

Half of Twitter trending topics in Turkey found to be fake
Top subjects for fake trends in 2019 include illicit ads, politics, and people seeking policy changes
Beyza Binnur Donmez | 10.06.2021

How hackers have manipulated Twitter’s trending algorithm for years
Researchers say the algorithm can be deceived by “ephemeral astroturfing”

ANKARA
At least 47% of the Twitter trends in Turkey are fake, and 20% of those trends make up worldwide trends, according to a study by researchers in Switzerland.

Manipulating Twitter Trends is possible, and it's dangerous