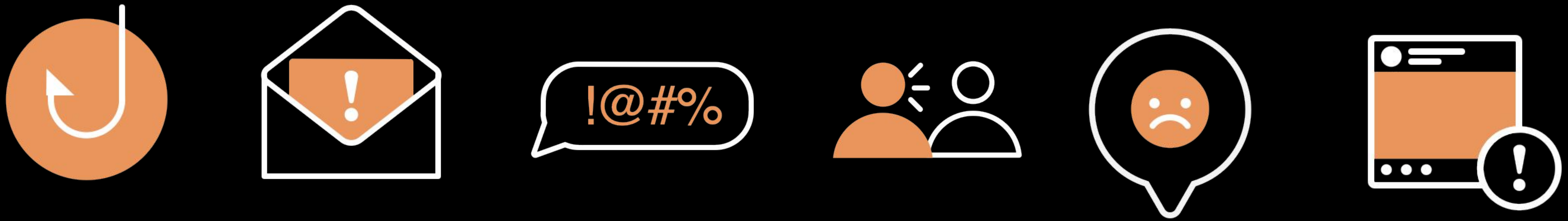
Deep Entity Classification: Abusive Account Detection for Online Social Networks

Teng Xu[1], Gerard Goossen[1], Huseyin Kerem Cevahir[1], Sara Khodeir[1], Yingyezhe Jin[1], Frank Li[1,3], Shawn Shan[1,2], Sagar Patel[1], David Freeman[1], and Paul Pearce[1,3]

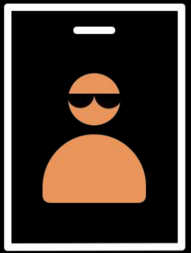
[1]Facebook, Inc [2]University of Chicago [3]Georgia Institute of Technology

30TH USENIX SECURITY SYMPOSIUM

Problem



Clickbait Spam Harassment Bullying Hate Speech Nudity

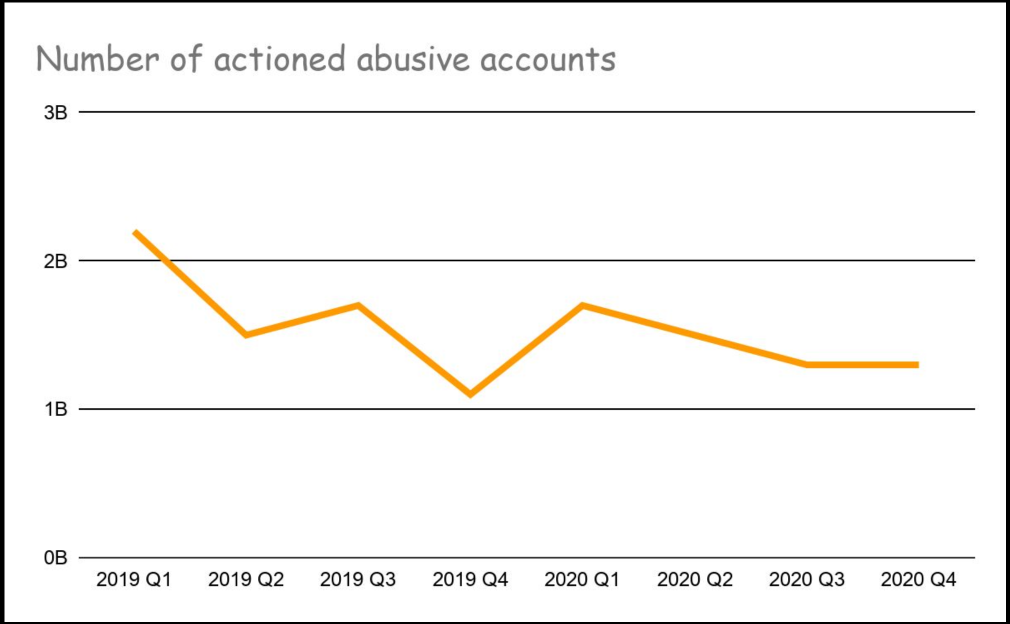


Abusive account: an account created for the purpose of abuse (i.e. activity that goes against Facebook's Community Standards).

Abusive Accounts on Facebook

Estimated 5% of monthly active users are abusive accounts.[1]

Took down **1.3 Billion Abusive** Accounts from 2020 Q4[1], most within minutes of registration, before they could become active users.



1. *Facebook community standards enforcement report: https://transparency.facebook.com/community-standards-enforcement#fake-accounts*

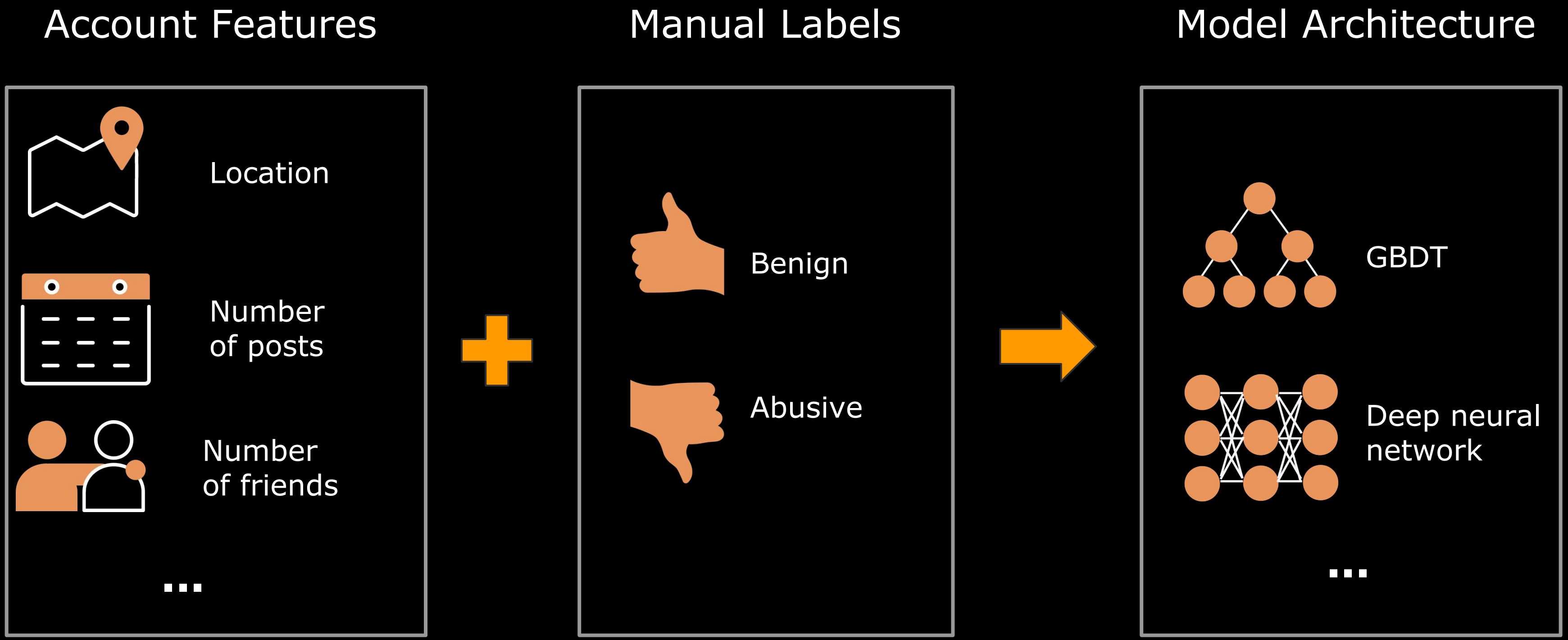
Machine Learning Based Detection

Manual review
does not scale

Heuristic rules are hard to
create and maintain

Adversaries
move fast

ML: Traditional Approach



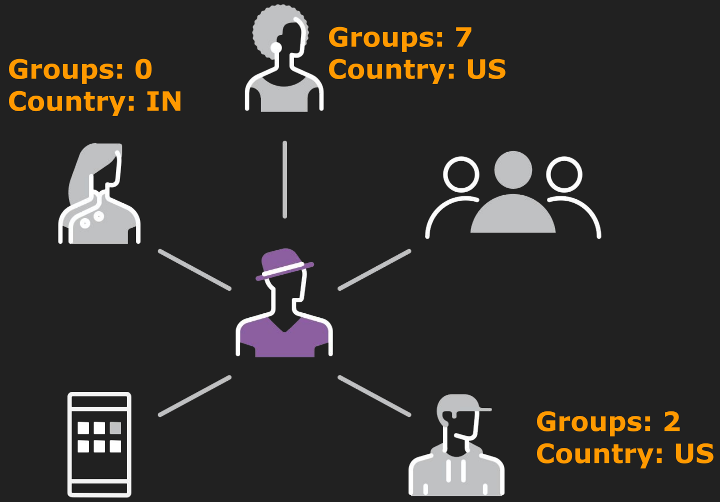
Solution: deep entity classification

Problem	Solution
Features can be gamed by attackers.	Extract "deep features" of accounts by aggregating properties and behavioral features from direct and indirect neighbors in graph.
Features are hand written, which only scales to hundreds of features.	Define dozens of features per edge, apply to all edges, and recursively traverse the graph, resulting in tens of thousands of features.
Obtaining large amounts of ground truth data is difficult.	Use a multi-stage multi-task learning technique using large amounts of low-precision automated labels, and small amounts of high-precision human labels.

Deep Feature Extraction

First order

- Apply aggregation functions to direct features of fanout entities.
- Numeric aggregation functions:
    - *max*
    - *min*
    - *mean*
    - *p75*
    - *p25*
    - *variance*
- Categorical aggregation functions:
    - *percentage of the most common category*
    - *percentage of empty values.*
    - *entropy of the category values.*
    - *number of distinct categories.*



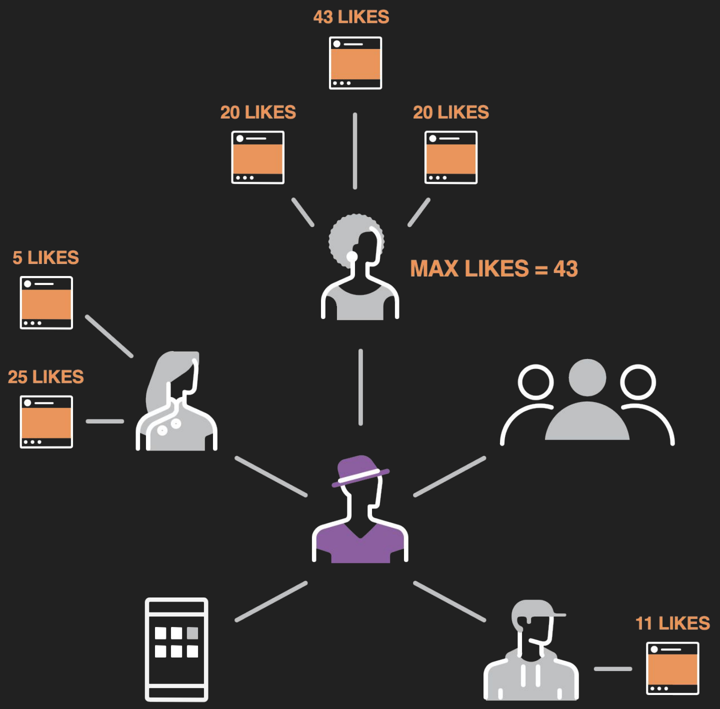
Avg (# of groups per friend) = 3

Most common percentage (friend country) = 0.67

Deep Feature Extraction

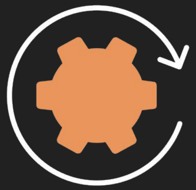
Second order

- Apply aggregation functions to second order fanout entities.
- Aggregate results over first order fan-outs.
- Lots of features, expensive to calculate.



$$\min_{\text{friends}}\left(\max_{\text{posts}}(\#\text{ of likes per post})\right) = 11$$

Training Data



Automated Labels

- Lower precision
- High volume
- Low cost
- Sources: past actioned accounts, user reports

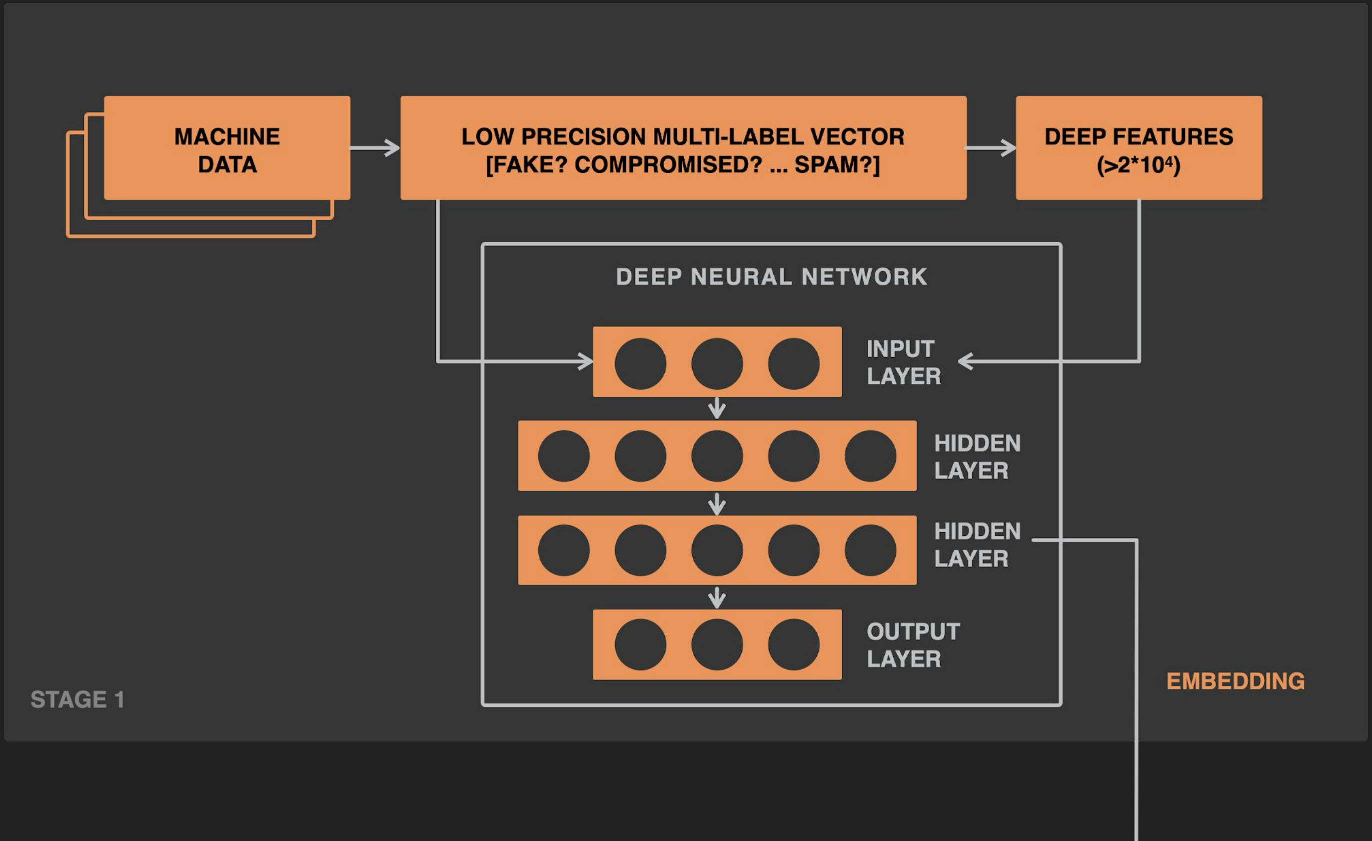
Human Labels

- Higher precision
- Low volume
- High cost
- Sources: manual review by domain experts

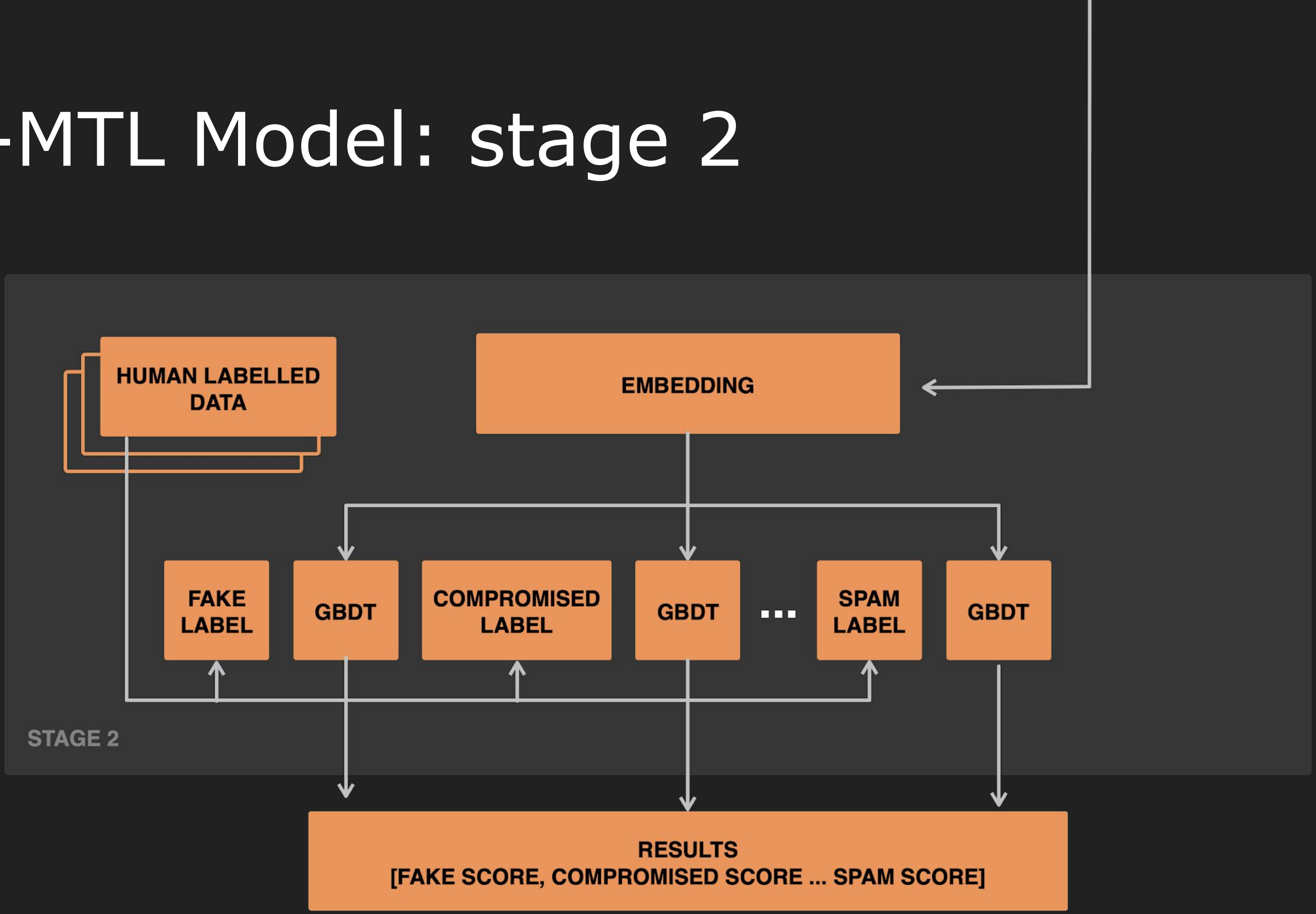
How do we avoid overfitting and also obtain benefit of high-quality labels?

Multi-stage multi-task learning (MS-MTL)

MS-MTL Model: stage 1



MS-MTL Model: stage 2



Model Comparisons

1

Only behavioral features + GBDT

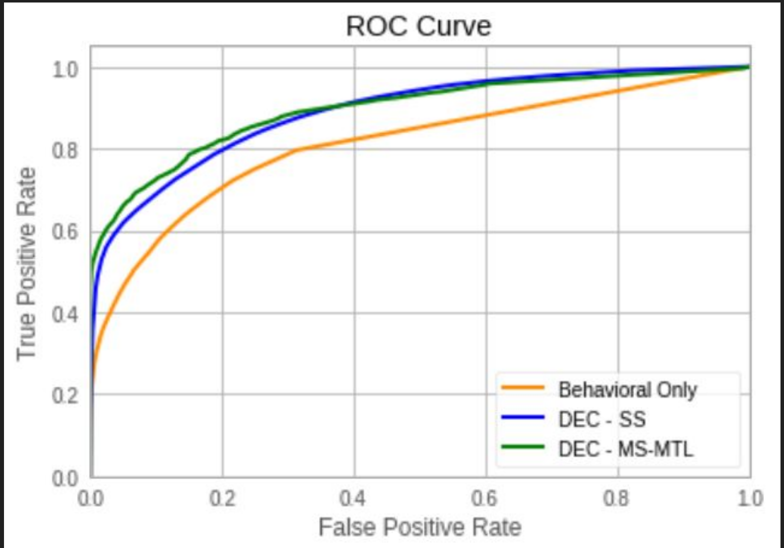
2

DEC features + single stage deep neural network (SS)

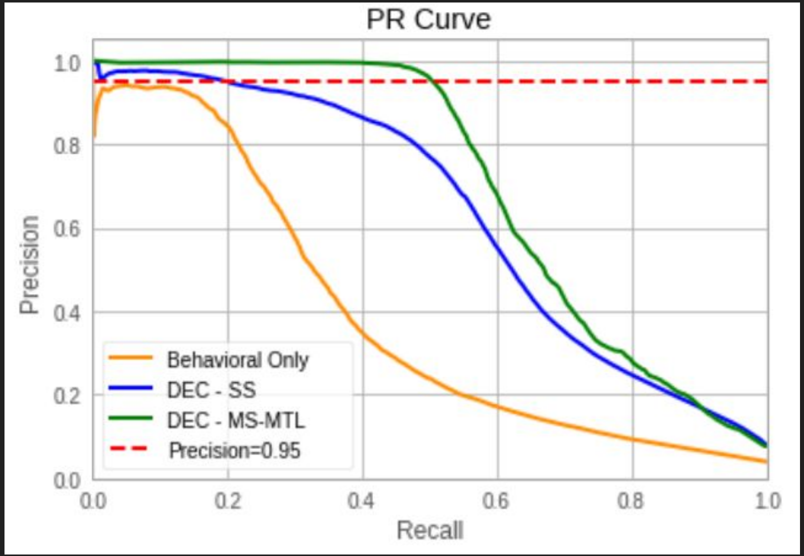
3

DEC features + MS-MTL

Offline evaluation

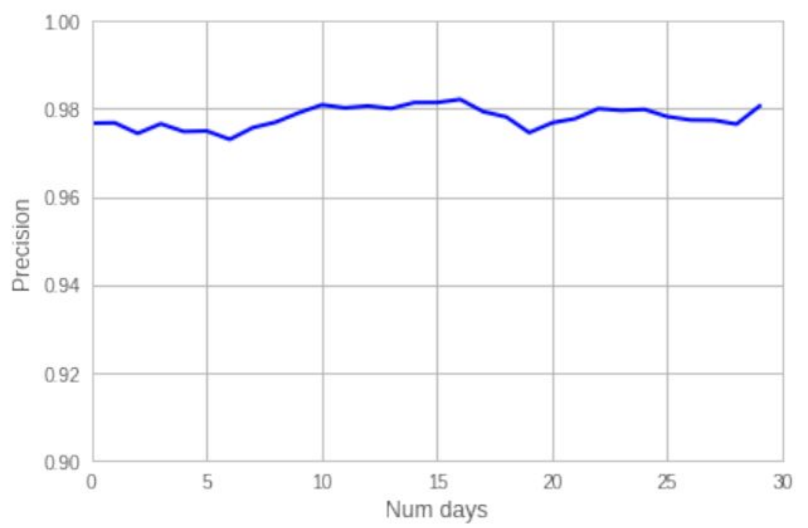


ROC

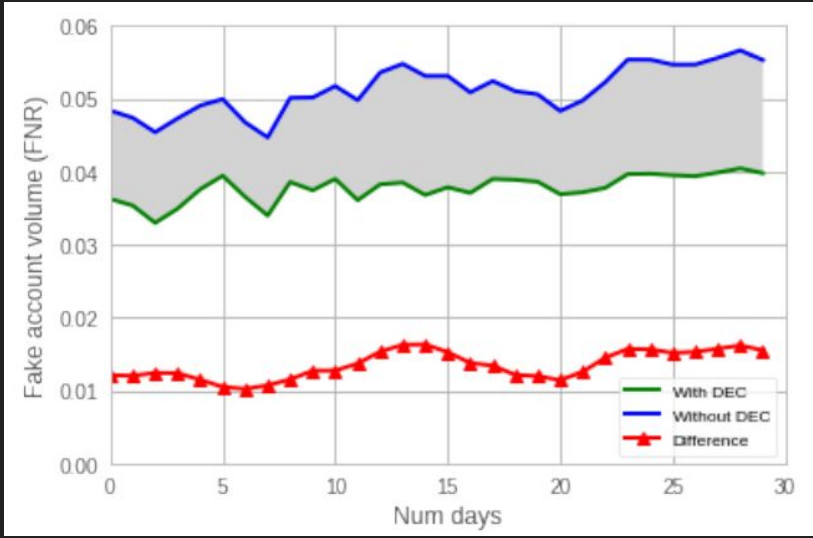


Precision-recall

Online evaluation



In production: precision over 30 days



In production: recall over 30 days

Takeaways

1. **Extracting graph-based "deep features" of accounts allows us to scale features and resist adversarial adaptation.**

2. **MS-MTL training leverages both high quantity-low precision, and low quantity-high precision training data to improve model performance.**

3. **DEC's two-year deployment has resulted in Facebook taking down hundreds of millions of abusive accounts.**

4. **Counterintuitively, the deployment of DEC *reduced* global CPU usage on Facebook despite the high computational load.**

Thank you

Contact: xuteng@fb.com for questions and further information