Semantic Text Analysis for Detection of Compromised Accounts on Social Networks

Dominic Seyler (dseyler2@Illinois.edu)

Lunan Li (lunanli3@illinois.edu)

ChengXiang Zhai (czhai@Illinois.edu)

University of Illinois at Urbana-Champaign, USA

Motivation


By ANDREA PARK / CBS NEWS / March 23, 2017, 10:53 AM
ABC News, "Good Morning
TWITTER
McDonald's Twitter account hacked,
Prominent Twitter accounts compromised after third-
POLITICS • HACKING
Inside Russia's Social Media War on America
Twitter hacked; 250,000 accounts affected
By Heather Kelly, CNN

Motivation

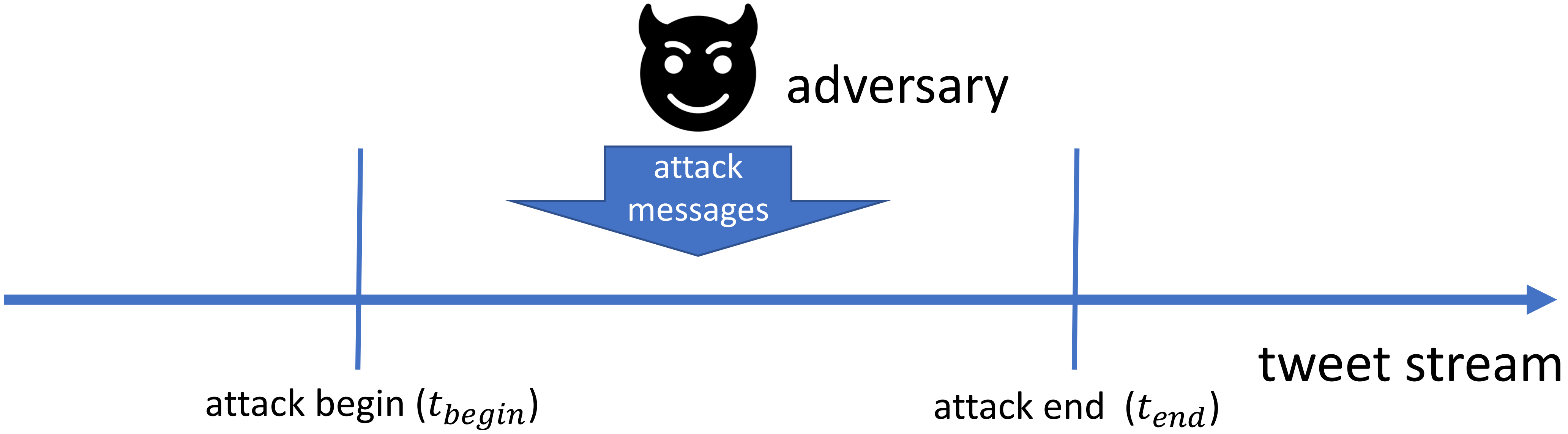
- **Compromised accounts:** legitimate accounts that an adversary takes control over for gaining financial profit or spreading misinformation.
- Compromised accounts are <u>more valuable</u> for hackers:
  - Harder to detect because they show characteristics of legitimate accounts.
  - Hackers can exploit the trust network the legitimate user has created.
- Issues related to compromised accounts:
  - Detection can take up to five days, with 60% of takeovers lasting entire day.
  - In 2013, over 250k Twitter accounts were compromised; issue remains today.
  - 21% of victims of account compromise abandon social media platform.
- **Goal:** Detect compromised accounts on social media platforms.

Talk Outline

1. Threat Model
2. Detection Framework
3. Creating Ground Truth Dataset
4. Feasibility Analysis
5. Experiments
6. Conclusion

Threat Model

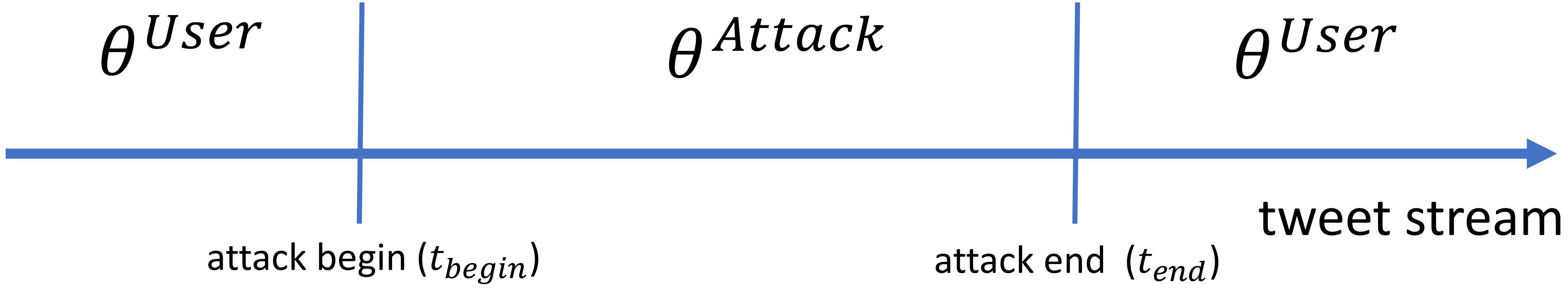
- Adversary's goal: Inject textual output into a benign account to mask its origin and leverage the user's influence network.



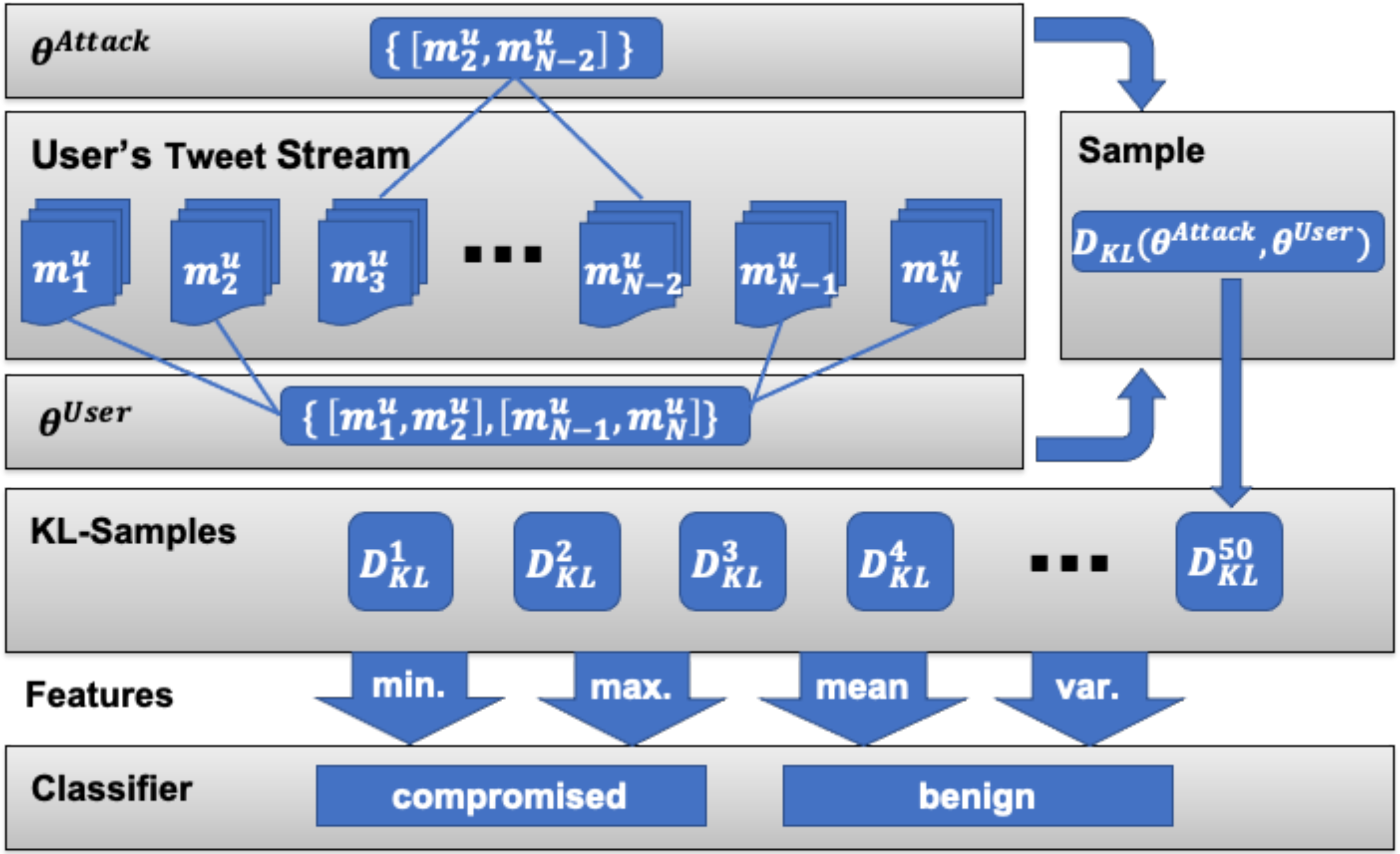
- Observation: adversary's textual output will deviate from user's output.

Detection Framework

- Create language model for attacker (θ^{Attack}) and user (θ^{User}).
- Sample random t_{begin}, t_{end} and measure difference in distributions.
- Use difference as a feature in classification framework.



Instantiation of Detection Framework



Creating Ground Truth Dataset

- No dataset available: Simulate account attacks according to our threat model.
- Use Twitter crawl [1] and switch part of a user's twitter stream with tweets from another user to artificially create a compromised account.
- Begin and end of account take-over are chosen at random.
- Harder than the ”real” problem, since two regular twitter users will use less discriminative language than a user and an adversary.

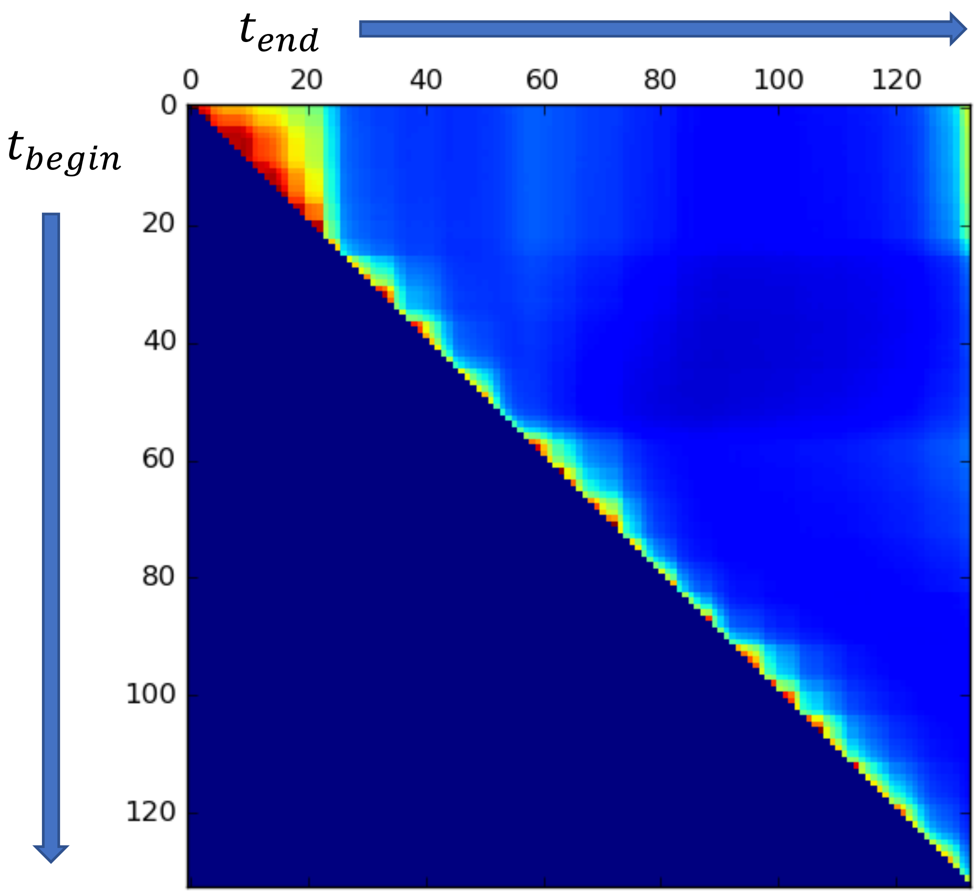
[1] Yang and Leskovec 2011. “Patterns of temporal variation in online media.” In WSDM.

Feasibility Analysis

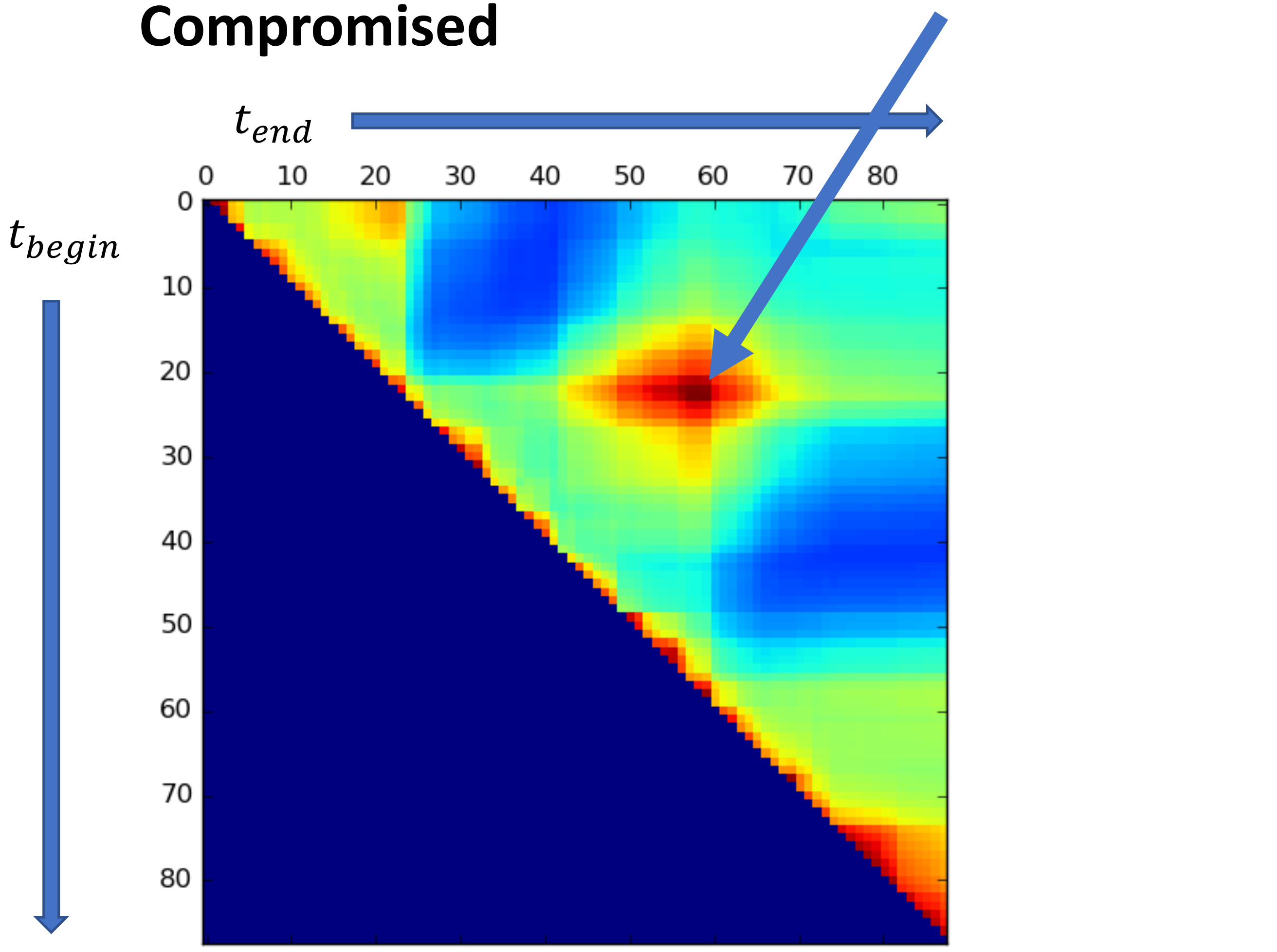
- Find evidence:
  - (1) compromised user accounts do exhibit higher KL-divergence compared to benign accounts.
  - (2) average KL-divergence can be estimated by randomly sampling a certain number of points with different begin-end dates.
- Methodology:
  - Select 495 users at random.
  - Calculate KL-divergence for all possible combinations of t_{begin} and t_{end}.

(1) Compromised User Accounts Exhibit Higher KL-divergence

Benign

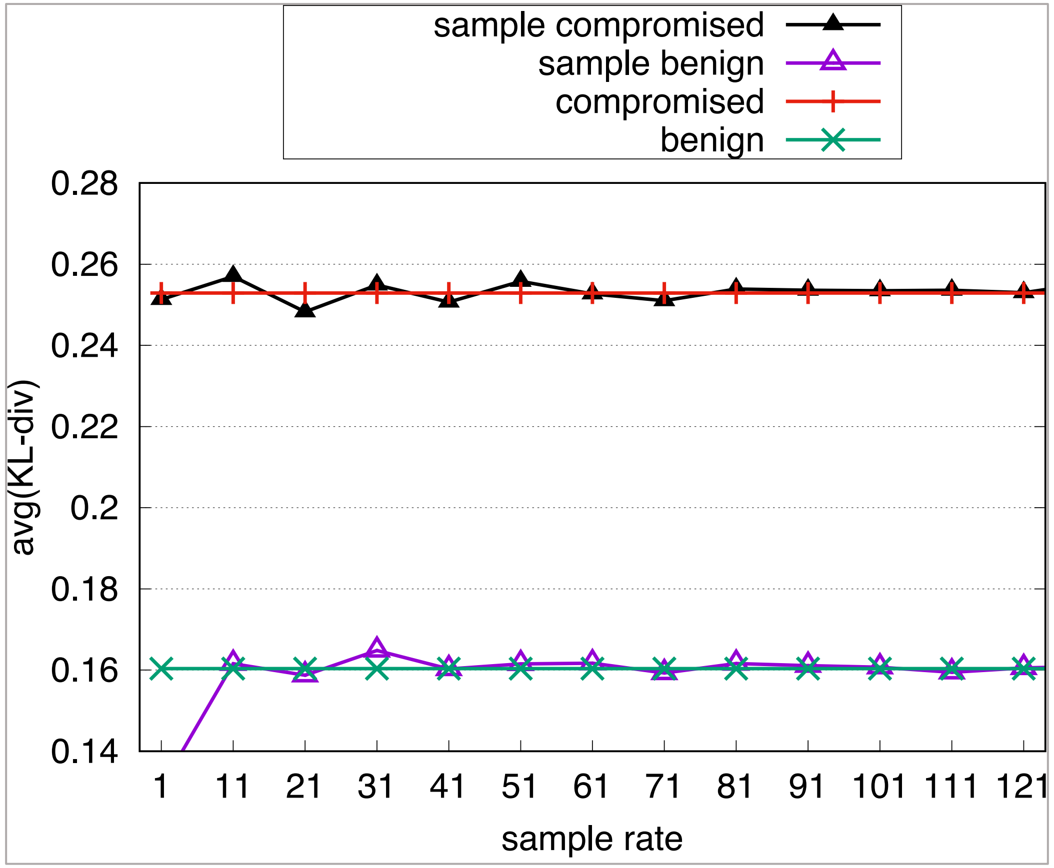


Compromised



(2) Estimate Average KL-divergence Using Random Sampling

- Plot actual average KL-divergence against the average sampled KL-divergence.
- Average KL-div. higher for compromised accounts.
- For sample rates < 81 minimal deviations in approximation (+/- 0.01).

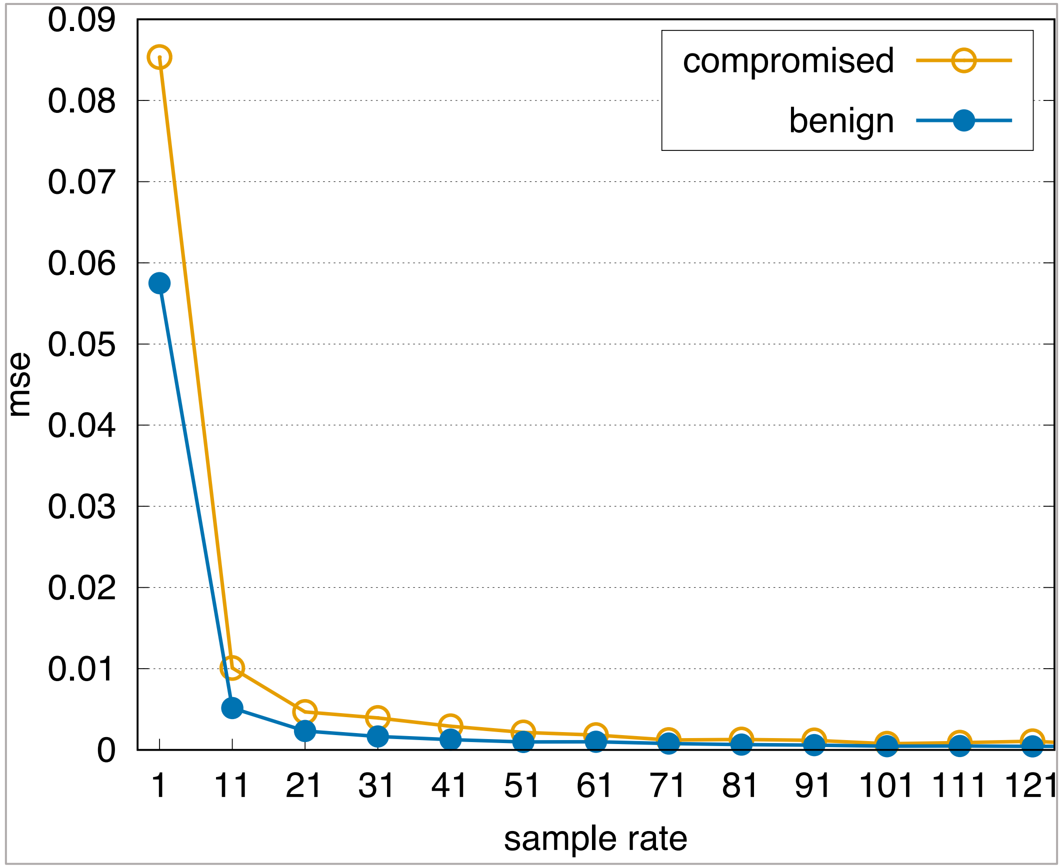


(2) Estimate Average KL-divergence Using Random Sampling

- Measure Mean Squared Error (mse) as:

  $\frac{1}{n}\sum_{u \in U^{test}}(sampled\_avg(u) - actual\_avg(u))^2$

- For sample rates < 50 errors over 0.07 and 0.06.
- For sample rate < 101 mse is close to 0.
- Conclusion: Sample rate of 50 to 100 is sufficient for experiments.



Experiments

- Research Questions
  1. How does the proposed language model feature compare to general text classification features? Can they be combined?
  2. How does the language model feature perform in comparison to other compromised account detection methods?
  3. How effective is our method on a real (non-simulated) dataset, when trained using simulated data?

Experimental Design

- Dataset:
  - Simulation dataset using a Twitter crawl [1] based on our thread model.
  - Resulting dataset contains 99,912 user accounts with close to 129.5 million tweets (dataset is balanced).
- Baselines:
  - General text representations: (1) word count, (2) TF*IDF and (3) Doc2Vec.
  - Existing compromised account detection methods: COMPA [2]; VanDam [3]
- Classification Framework:
  - Support Vector Machine (SVM) with ten-fold cross validation.

[1] Yang and Leskovec 2011. “Patterns of temporal variation in online media.” In WSDM.
[2] Egele et al. 2013. “Compa: Detecting compromised accounts on social networks.” In NDSS.
[3] VanDam et al. 2017. “Understanding compromised accounts on twitter.” In Web Intelligence.

Ablation Study

Ablation study using different measures.

Measure	All	Max	Min	Mean	Var.
Accuracy	**0.80**	0.59	0.76	0.75	0.48
F_1	**0.78**	0.57	0.72	0.72	0.35
Precision	**0.90**	0.61	0.87	0.83	0.47
Recall	**0.68**	0.53	0.61	0.64	0.27

- Maximum performance over all metrics achieved when all features are used.
- High precision (0.9) and accuracy (0.8).
- *Minimum* and *Mean* seem to be strongest features.

Comparison to General Text Representations

Accuracy for different features and their combinations.

%	*COUNT*	*TF*IDF*	*Doc2Vec*	*Doc2Vec + TF*IDF*	*LM*	*LM + TF*IDF*	*LM + Doc2Vec*	all
50	0.53	0.56	0.71	0.72	0.80	0.81	**0.87**	**0.87**
25	0.53	0.55	0.69	0.69	0.75	0.75	**0.82**	**0.82**
10	0.52	0.54	0.63	0.63	0.65	0.65	0.70	**0.71**
5	0.52	0.53	0.59	0.59	0.59	0.59	**0.62**	**0.62**
RND	0.53	0.55	0.68	0.68	0.74	0.74	**0.80**	**0.80**

- LM stand-alone outperforms all general text representations.
- Adding Doc2Vec to LM results the highest improvements.
- Best performance is achieved when features are combined.

Comparison to Related Methods

Model	Accuracy	F_1	Precision	Recall
COMPA	0.62	0.60	0.64	0.56
VanDam	0.50	0.47	0.50	0.45
LM	**0.74**	**0.70**	**0.81**	**0.61**
improvement *LM* over best baseline	19.4%	16.7%	26.6%	8.9%
LM + COMPA	0.75	**0.73**	0.81	0.66
LM + VanDam	0.74	0.71	**0.82**	0.62
LM + COMPA + VanDam	**0.76**	**0.73**	0.81	**0.67**
improvement over *LM*	2.7%	4.3%	1.2%	9.8%
LM + *Doc2Vec* + *TF*IDF* + COMPA + VanDam	**0.81**	**0.79**	**0.85**	**0.75**
improvement when adding standard features	6.6%	8.2%	4.9%	11.9%

- LM stand-alone outperforms all baseline methods.
- Combining methods is beneficial.
- Best performance is achieved, when standard features are added.

Effectiveness on Non-Simulated Data

- Manual Analysis: Apply algorithm to real-world data and investigate accounts with highest probability of being compromised.

Category	Count	Status	Count
News	5	Abandoned	7
Spam	4	Active	6
Re-tweet Bot	2	Deleted	4
Compromised	1	Protected	2
Regular	7	Suspended	1
Unknown	1		

- If trained on real data, detection of more suspended accounts expected.
- Our algorithm can detect “unusual” accounts and users.

Conclusion

- Novel general framework for detecting compromised accounts using semantic text analysis.
- Instantiation of framework was shown to be effective.
- Proposed language model features are most effective and show improvement when added on top of other methods.
- Our features capture signals that exiting methods are missing.
- Model can be trained without any human involvement (using simulation) to detect “unusual” accounts.

Semantic Text Analysis for Detection of Compromised Accounts on Social Networks

Dominic Seyler (dseyler2@Illinois.edu)

Lunan Li (lunanli3@illinois.edu)

ChengXiang Zhai (czhai@Illinois.edu)

University of Illinois at Urbana-Champaign, USA

Thank You!

more info at: https://dominicseyler.com