TABLE IV: Comparison to features from related methods. The dataset has random percentages of tweets compromised (RND).

Model	Accuracy	F_1	Precision	Recall
COMPA [8]	0.62	0.60	0.64	0.56
VanDam [13]	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%

TABLE V: Statistics of manually evaluated accounts.

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		

that if our algorithm were applied at much larger scale to all the Twitter users, it would most likely be able to detect many more compromised accounts.

VI. Conclusion and Future Work

We proposed a novel general framework based on semantic text analysis for detecting compromised social media accounts. Following the framework, we proposed a specific instantiation based on uni-gram language models and KL-divergence measure, and designed features accordingly for use in a classifier that can distinguish compromised from benign accounts. We conclude that (1) the proposed *LM* feature is most effective, even when used as a single feature-based detection method. (2) *LM* captures new signals that haven't been captured in the existing methods and features, which is shown by the further improvement when added on top of the baselines. (3) The best performing method would combine the proposed LM with all the existing features. Although *LM* is motivated by a security problem, our general idea of performing differential semantic analysis of text data may be applicable to other domains where incohesion (or outlier) in text data needs to be captured.

References

[1] S. Cresci *et al.*, "Better safe than sorry: an adversarial approach to improve social bot detection," in *Web Sci'*, 2019.

[2] C. Grier, K. Thomas, V. Paxson, and C. M. Zhang, "@spam: the underground on 140 characters or less," in *CCS*, 2010.

[3] J. Martinez-Romo and L. Araujo, "Detecting malicious tweets in trending topics using a statistical analysis of language," *Expert Syst. Appl.*, 2013.

[4] K. Thomas *et al.*, "Consequences of connectivity: Characterizing account hijacking on twitter," in *SIGSAC*, 2014.

[5] E. Zangerle and G. Specht, ""sorry, I was hacked": a classification of compromised twitter accounts," in *SAC*, 2014.

[6] H. Kelly, "Twitter hacked; 250,000 accounts affected," February 2013. [Online]. Available: https://cnn.it/2XIpLYZ

[7] K. Olmstead and A. Smith, "Americans and cybersecurity," *Pew Research Center*, vol. 26, 2017.

[8] M. Egele *et al.*, "COMPA: detecting compromised accounts on social networks," in *NDSS*, 2013.

[9] X. Ruan *et al.*, "Profiling Online Social Behaviors for Compromised Account Detection," *IEEE Trans. Information Forensics and Security*, 2016.

[10] H. Karimi *et al.*, "End-to-end compromised account detection," in *ASONAM*, 2018.

[11] C. VanDam *et al.*, "Cadet: A multi-view learning framework for compromised account detection on twitter," in *ASONAM*, 2018.

[12] ——, "You have been caute! early detection of compromised accounts on social media," in *ASONAM*, 2019.

[13] ——, "Understanding compromised accounts on twitter," in *WI*, 2017.

[14] S. Gupta *et al.*, "Modeling and detecting anomalous topic access," in *ISI*, 2013.

[15] D. Trang *et al.*, "Evaluating Algorithms for Detection of Compromised Social Media User Accounts," in *ENIC*, 2015.

[16] B. Viswanath *et al.*, "Towards detecting anomalous user behavior in online social networks," in *USENIX*, 2014.

[17] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *Ann. Math. Statist.*, 1951.

[18] C. Zhai and S. Massung, *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. Morgan & Claypool, 2016.

[19] J. Yang and J. Leskovec, "Patterns of temporal variation in online media," in *WSDM*, 2011.

[20] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *ICML*, 2014.

[21] Y. Wang *et al.*, "A study of feature construction for text-based forecasting of time series variables," in *CIKM*, 2017.

[22] J. Pennington *et al.*, "Glove: Global vectors for word representation," in *EMNLP*, 2014.