Lightning Talk - Think Outside the Dataset: Finding Fraudulent Reviews using Cross-Dataset Analysis

Shirin Nilizadeh
University of Texas, at Arlington
shirin.nilizadeh@uta.edu

Hojjat Aghakhani
University of California Santa Barbara
hojjat@ucsb.edu

Eric Gustafson
University of California Santa Barbara
edg@ucsb.edu

Christopher Kruegel
University of California Santa Barbara
chris@ucsb.edu

Giovanni Vigna
University of California Santa Barbara
vigna@ucsb.edu

ABSTRACT

Many crowd-sourced review platforms, such as Yelp, TripAdvisor, and Foursquare, have sprung up to provide a shared space for people to write reviews and rate local businesses. With the substantial impact of businesses' online ratings on their selling [2], many businesses add themselves to multiple websites to more easily be discovered. Some might also engage in reputation management, which could range from rewarding their customers for a favorable review, or a complex review campaign, where armies of accounts post reviews to influence a business' average review score.

Most of previous work use supervised machine learning, and only focus on textual and stylometry features [1, 3, 4, 7]. Their obtained ground truth data is not large and comprehensive [4–8, 10]. These works also assume a limited threat model, *e.g.,* an adversary's activity is assumed to be found near sudden shifts in the data [8], or focused on positive campaigns.

We propose OneReview , a system for finding fraudulent content on a crowd-sourced review site, leveraging correlations with other independent review sites, and the use of textual and contextual features. We assume that an attacker may not be able to exert the same influence over a business' reputation on several websites, due to increased cost. OneReview focuses on isolating anomalous changes in a business' reputation across multiple review sites, to locate malicious activity without relying on specific patterns. Our intuition is that a business's reputation should not be very different in multiple review sites; *e.g.,* if a restaurant changes its chef or manager, then the impact of these changes should appear on reviews across all the websites. OneReview utilizes *Change Point Analysis* method on the reviews of every business independently on every website, and then uses our proposed *Change Point Analyzer* to evaluate change-points, detect those that do not match across the websites, and identify them as suspicious. Then, it uses supervised machine learning, utilizing a combination of textual and metadata features to locate fraudulent reviews among the suspicious reviews.

We evaluated our approach, using data from two reviewing websites, Yelp and TripAdvisor, to find fraudulent activity on Yelp. We obtained Yelp reviews, through the Yelp Data Challenge [9], and used our Change Point Analyzer to correlate this with data crawled from TripAdvisor. Since realistic and varied ground truth data is not currently available, we used a combination of our change point analysis and crowd-labeling to create a set of 5,655 labeled reviews. We used k-cross validation (k=5) on our ground truth and obtained 97% (+/- 0.01) accuracy, 91% (+/- 0.03) precision and 90% (+/- 0.06) recall. The model was used on the suspicious reviews, which classified 61,983 reviews, about 8% of all reviews, as fraudulent.

We further detected fraudulent campaigns that are actively initiated by or targeted toward specific businesses. We identified 3,980 businesses with fraudulent reviews, as well as, 14,910 suspected spam, where at least 40% of their reviews are classified as fraudulent. We also used community detection algorithms to locate several large astroturfing campaigns. These results show the effectiveness of OneReview in detecting fraudulent campaigns.

CCS CONCEPTS

• **Information systems → Trust**; **Reputation systems**.

KEYWORDS

Fraudulent Reviews; Cross-Dataset Change-Point Analysis;

ACM Reference Format:
Shirin Nilizadeh, Hojjat Aghakhani, Eric Gustafson, Christopher Kruegel, and Giovanni Vigna. 2019. Lightning Talk - Think Outside the Dataset: Finding Fraudulent Reviews using Cross-Dataset Analysis. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3316477

REFERENCES

[1] Hojjat Aghakhani, Aravind Machiry, Shirin Nilizadeh, Christopher Kruegel, and Giovanni Vigna. 2018. Detecting deceptive reviews using generative adversarial networks. In *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE.

[2] Michael Anderson and Jeremy Magruder. 2012. Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal* 122, 563 (2012), 957–989.

[3] Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, 171–175.

[4] Nitin Jindal and Bing Liu. 2008. Opinion Spam and Analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM '08)*. ACM, New York, NY, USA, 219–230. https://doi.org/10.1145/1341531.1341560

[5] Jiwei Li, Myle Ott, Claire Cardie, and Eduard H Hovy. 2014. Towards a General Rule for Identifying Deceptive Opinion Spam.. In *ACL (1)*. Citeseer, 1566–1576.

[6] Yuming Lin, Tao Zhu, Hao Wu, Jingwei Zhang, Xiaoling Wang, and Aoying Zhou. 2014. Towards online anti-opinion spam: Spotting fake reviews from the review sequence. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*. IEEE, 261–264.

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316477

[7] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 309–319.

[8] Mahmudur Rahman, Bogdan Carbunar, Jaime Ballesteros, George Burri, Duen Horng, et al. 2014. Turning the Tide: Curbing Deceptive Yelp Behaviors.. In *SDM*. SIAM, SIAM, 244–252.

[9] Yelp. 2016. Yelp Dataset Challenge. https://www.yelp.com/dataset_challenge.

[10] Kyung-Hyan Yoo and Ulrike Gretzel. 2009. Comparison of deceptive and truthful travel reviews. *Information and communication technologies in tourism 2009*.