Tackling imbalanced data in on-line fake review detection

Satheesh Joseph Catherine Mou Yi Zhang

***Abstract*—Abstract**

I. Introduction

On-line fake review detection is a relatively well studies subject, especially in recent years, given its outsized impact on consumers engaging in e-commerce activities online. Positive reviews bring a meaningful increase in sales volume to the products[2], and vice versa for negative reviews.

As a result, there has been an increase in opinion spamming activity, and detecting fake reviews has become an essential requirement for on-line marketplaces to maintain the integrity and fairness of their platform.

However, there has been one persisting challenge [7], [8], [9], [11] in this area of research – the lack of substantial body of actual, proven fake reviews, directly leading to significantly imbalanced datasets.

Our work aims to tackle this problem of data imbalance by borrowing ideas from Generative Adversarial Network (GAN). Our **hypothesis** is that it's possible to make up the data imbalance by generating fake reviews from a language model trained/fine-tuned on actual fake reviews. We will validate this approach by then training a review detection model on a balanced dataset that includes the generated negative reviews, and achieving comparable results to state-of-the-art research [8].

II. Background

There is no shortage of research tackling the problem of fake review detection. A recent survey [3] does a great job laying out the landscape of the various techniques and data sets used for fake review detection.

According to this survey, all large datasets (>20k reviews) from Yelp [6] contain less than 15% actual fake reviews. There are a few other widely used public datasets crawled from TripAdvisor, but they are of much smaller scale, with the fake review training set generated via a manual process from Amazon Mechanical Turk.

The only balanced dataset of moderate volume is crawled from Yelp by Barbado et. al. [1], however it has not been widely adopted in the research community as a benchmark for detecting fake reviews.

Given this state of the related work, and inspired by Stanton et. al. [7] who used GAN techniques to generate behavioral features (e.g. number of reviews, percentage of positive reviews) of on-line Yelp reviewers, we believe that similar techniques can be used to generate the reviews themselves.

With sufficient representativeness, we believe the generated fake reviews can serve as additional training examples that helps with the detection model to distinguish between genuine reviews and fake ones – "Happy families are all alike; every unhappy family is unhappy in its own way."

WIP:

- Benchmark control group paper [8]

- Data description [10]

- GAN Paper [7]

- Original data comes from [5] [4]

III. Methods

Methods

IV. Results & Discussion

Results and Discussion

V. Conclusion

Example table

References

[1] Rodrigo Barbado, Oscar Araque, and Carlos A Iglesias. A framework for fake review detection in online consumer electronics retailers. *Information Processing & Management*, 56(4):1234–1244, 2019.

[2] Nga N Ho-Dac, Stephen J Carson, and William L Moore. The effects of positive and negative online customer reviews: do brand strength and category maturity matter? *Journal of marketing*, 77(6):37–53, 2013.

[3] Rami Mohawesh, Shuxiang Xu, Son N. Tran, Robert Ollington, Matthew Springer, Yaser Jararweh, and Sumbal Maqsood. Fake reviews detection: A survey. *IEEE Access*, 9:65771–65802, 2021.

[4] Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Natalie Glance. What yelp fake review filter might be doing? In *Seventh international AAAI conference on weblogs and social media*, 2013.

[5] Arjun Mukherjee, Vivek Venkataraman, Bing Liu, Natalie Glance, et al. Fake review detection: Classification and analysis of real and pseudo reviews. *UIC-CS-03-2013. Technical Report*, 2013.

TABLE I. SUMMARY OF EXPERIMENTAL RESULTS

System	n	γ	Samples	mean	max
A	21	0.2	24236 (11%)	0.083	0.084
		0.15	43307 (20%)	0.081	0.081
		0.1	97440 (44%)	0.074	0.074
B	23	0.2	26332 (2.2%)	0.035	0.036
		0.15	46812 (4.0%)	0.026	0.026
		0.1	105326 (8.9%)	0.0068	0.0069
C	26	0.2	29289 (2.2%)	0.084	0.085
		0.15	52070 (3.9%)	0.084	0.084
		0.1	117156 (8.8%)	0.080	0.082
D	20	0.2	23375 (2.2%)	0.074	0.080
		0.15	41555 (4.0%)	0.034	0.037
		0.1	93497 (8.9%)	0.024	0.024

n = Number of features of system.
γ = User specified maximum error.
Samples = Number & proportion of samples used.
mean = Average actual error from the 10 runs.
max = Maximum actual error from the 10 runs.

[6] Shebuti Rayana and Leman Akoglu. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining*, pages 985–994, 2015.

[7] Gray Stanton and Athirai A Irissappane. Gans for semi-supervised opinion spam detection. *arXiv preprint arXiv:1903.08289*, 2019.

[8] Xiaoya Tang, Tieyun Qian, and Zhenni You. Generating behavior features for cold-start spam review detection with adversarial learning. *Information Sciences*, 526:274–288, 2020.

[9] Jingdong Wang, Haitao Kan, Fanqi Meng, Qizi Mu, Genhua Shi, and Xixi Xiao. Fake review detection based on multiple feature fusion and rolling collaborative training. *IEEE Access*, 8:182625–182639, 2020.

[10] Xuepeng Wang, Kang Liu, and Jun Zhao. Handling cold-start problem in review spam detection by jointly embedding texts and behaviors. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 366–376, 2017.

[11] Chunyuan Yuan, Wei Zhou, Qianwen Ma, Shangwen Lv, Jizhong Han, and Songlin Hu. Learning review representations from user and product level information for spam detection. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1444–1449. IEEE, 2019.