Tackling skewed data in on-line fake review detection

Satheesh Joseph
satheeshrishi@berkeley.edu

Catherine Mou
catherine041616@berkeley.edu

Yi Zhang
yizhang7210@berkeley.edu

***Abstract*—The increasing need and value of the digital text and the trustworthiness behind it inspired our imagination to handle a niche problem within the space of opinion spam analysis. Recently many researches have been conducted to increase the detection accuracy. To the best of our knowledge, an important issue that has been less studied is to tackle imbalanced data in on-line fake review detection. In this paper, we study a methodology to tackle the imbalanced data issue and the accuracy effects on different state-of-the-art models. For this research, we use Yelp review data for hotels and restaurants that is classified as needed.**

I. Introduction

On-line fake review detection is a relatively well studied 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 activities, 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], [12] in this area of research – the lack of substantial body of actually 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 and/or fine-tuned on actual fake reviews. We will look to validate this approach by then training a review detection model on a balanced dataset that includes the generated fake reviews, and achieving comparable results to state-of-the-art research [8].

Section II provides the background on the existing research in fake review detection. Section III lays out our methods of constructing the fake review generator. Section IV discusses the experiments we run to validate the usefulness of the generated fake reviews in training a detection model. And we draw our final conclusions in section V.

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) for 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 can help with the detection model to distinguish between genuine reviews and fake ones. To paraphrase Tolstoy – genuine reviews are all alike; every fake review is fake in its own way.

III. Methods

The general structure of a Generative Adversarial Network consists of a *generative* network that generates candidates as well as a *discriminative* network that evaluates them. Here we apply this technique to generate fake review texts.

For the *generative* network, we here use the well-established, pre-trained GPT-2 language model.

GPT-2 was pre-trained on web scraped industrial-scale corpus, it was trained with a batch size of 512, well-defined sentence length, volcabulary seize of 50,000., and can be used for various NLP tasks including text generation. It is a statistical tool to generate next words in sequence based on preceding words, at every stage, it will take the previously generated data as additional input when generating next output. GPT-2 has outperformed other language models when it comes to generating text based on small input contents like hotel/restaurant reviews. GPT-2 has the ability to adapt to context of text, so it can generate realistic and coherent output.

To generate domain specific fake Yelp reviews, we took the Transformers library from huggingface and modified the language model file [11] to fine-tune the pre-trained model using the actual fake reviews from our training set.

For the *discriminative* network, we trained a simple Neural Network using ELMo embedding with 2 dense layers trained on the original data set.

To maintain some quality of the generated fake reviews, the final set of generated fake reviews that are used for model evaluation in section IV contain only the ones that the discriminative network made an error on.

IV. Experiment & Discussion

A. Experimental Setup

For our experiments, we use the publicly available data set originally obtained by [5] and [4]. This data set, containing 5858 reviews for hotels and 67019 reviews for restaurants on Yelp, is also used by a number of prior research papers for benchmarking, notably [10] and [8]. A more detailed statistics of the raw data set is in Table I. As mentioned above, it is indeed highly imbalanced with a much smaller number of fake reviews.

TABLE I. Summary of Experimental Data

Subject	Hotels	Restaurants
Total # Reviews	5858	67019
Total # Genuine Reviews	5078	58716
Total # Fake Reviews	780	8303
% Fake Reviews	13.3%	12.4%

To validate our hypothesis on the usefulness of the generated fake reviews, we set up a 2-stage experiment.

Firstly, we use the methods laid out in section III to generate fake reviews for both hotels and restaurants. Secondly, we add the generated fake reviews to the training data to obtain a balanced training set, then run a number of classification models on the mixed training set and compare it against our benchmark results.

We construct the data sets in the following ways to be comparable with the prior research papers [10], [8].

For a balanced test set:

- we first limit the pool of reviews to the first review per reviewer after 2012-01-01
- we take all the fake reviews (because there are fewer)
- we sample the same number of reviews from the genuine reviews
- this gives us a balanced, non-duplicated test set

For a balanced training set:

- we first limit the pool of reviews to be the ones prior to 2012-01-01
- we take all the actual fake reviews
- we include all generated fake reviews
- we sample the same number of reviews from the genuine reviews
- this gives us a balanced, non-duplicated training set

TABLE II. Summary of Training/Test Data

Subject	Hotels	Restaurants
Training set size	5070	
# Genuine Reviews	2535	
# Actual Fake Reviews	561	
# Generated Reviews	1974	
% Fake Reviews in training set	50%	
Test set size	432	
% Fake Reviews in test set	50%	

B. Models

We established a number of models all based on Neural Networks to compare our results with the benchmark.

Model 1: Our baseline model with GloVe embedding and 1 layer of LSTM.

Model 2: Our main model with GloVe embedding, 1 layer of Bidirectional LSTM, and a few more dense layers.

Model 3: A BERT based model.

Model 4: Same as Model 2 except using ELMo embedding instead of GloVe.

For each model, we'll be training on 4 different training sets for comparison:

Set 1: raw, imbalanced training set

Set 2: balanced training set by under-sampling genuine reviews to the number of fake reviews

Set 3: balanced training set by over-sampling fake reviews with replacement to the number of genuine reviews

Set 4: balanced training set by including generated fake reviews per Table II

C. Results

We report the best results of each model on each training set after hyper-parameter tuning via grid search. We will also report state-of-the-art classification benchmark from [8] using generated behavior features.

D. Discussion

V. Conclusion

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.

TABLE III. EXPERIMENTAL RESULTS - HOTELS

accuracy (f1)	Set 1	Set 2	Set 3	Set 4
Model 1	0.51/0.35	0.60/0.62	0.57/0.57	0.59/0.59
Model 2	0.56/0.63	0.59/0.59	0.57/0.62	0.59/0.60
Model 3	0.5	0.57	0.53	0.52
Model 4				
bfGAN[8]	83.0/83.4			

TABLE IV. EXPERIMENTAL RESULTS - RESTAURANTS

accuracy (f1)	Set 1	Set 2	Set 3	Set 4
Model 1	-	-	-	-
Model 2	-	-	-	-
Model 3	-	-	-	-
Model 4	-	-	-	-
bfGAN[8]	75.7/75.1			

[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.

[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] Jason Wu. How to build an ai text generator: Text generation with a gpt-2 model, 2020.

[12] 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.