Detecting Opinion Spammer Groups

Lars Stelzer
lars.stelzer@studium.uni-hamburg.de
7346178

Niklas von Boguszewski
nvboguszewski@googlemail.com
6790872

Knowledge Processing in Intelligent Systems: Practical Seminar
Knowledge Technology, WTM, Department of Informatics, University of Hamburg

***Abstract*—This paper aims to detect opinion spammer groups on the Yelp dataset.**
Methode nennen
results zeigen (conclusion)

I. Introduction

Consumers' decisions are significantly influenced by product reviews. Due to the ongoing digitalization more and more consumers rely their decision making on online reviews. Online product reviews have by now become the second must trusted source of product information, followed by recommendations from family and friends. This happens due to the belief that online product reviews reflect recommendations of "real" people. A rising online marketing budget has led to a larger influence and importance of online reviews in purchase and decision making. By the fact that positive/negative reviews are able to enhance/defame products organised opinion spamming started to take place. Nowadays it is estimated that up to 20% of online reviews could be fake.

These fake reviews are not the work of alone acting individuals, but more of organised payed group spammers [1]. More often the media reports of such groups. For example in 2013 nineteen companies got fined for their practice of writing fake online reviews. These companies systematically paid freelance writers from Philippines, Bangladesh and Eastern Europe between $1 to $10 per review. (https://www.marketwatch.com/story/20-of-yelp-reviews-are-fake-2013-09-24) With the growing popularity of these platforms and the direct influence on the businesses success the practice of leaving fake reviews for its own or the competitors business is becoming more common.Luca, M., and Zervas, G. (2016). Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. Management Science, 62(12), 34123427. doi:10.1287/mnsc.2015.2304 The *New York Times* reported in 2011 on the case of businesses hiring workers to post fake 5-star Yelp reviews on their behalfs for as little as 25 cent per review. https://www.nytimes.com/2011/05/22/your-money/22haggler.html In another case Samsung got a heavy penalty for paying people to post messages online that attacked HTC, their competitor, products while praising Samsungs. https://abcnews.go.com/Technology/samsung-fined-paying-people-criticize-htcs-products/story?id=20671547

Yelp filters roughly 16 % of its business reviews on its own. /cite Luca, M., and Zervas, G. (2016). Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. Management Science, 62(12), 34123427. doi:10.1287/mnsc.2015.2304 Still reportedly Yelps spokesman says 20-25 % of the submitted reviews are suspicious. https://www.marketwatch.com/story/20-of-yelp-reviews-are-fake-2013-09-24

In Fig. 1 examples of spam reviews we researched are shown. These reviews were posted by the same person but for two different businesses.

In this paper we propose a method to determine suspicious spammer groups accomplished by three steps. All these steps are based on the freely accessible Yelp.com dataset. Here we summarize the contributions of this paper as follows:

1. Firstly we create user groups based on the review time of the written review and the rating score given by the user. With this step we are able to create groups that reviewed the same business at nearly the same time giving the same rating score. Goal is to find groups that potentially work together writing fake reviews.
2. Followed by the creation of groups we determine features that help us to determine suspicious behaviour. In this paper we use three features that give hints if a user or group acts suspicious. We are looking at behavioral data like review time, rating score and textual similarity between reviews.
3. Based on our groups and features we are giving each group a score that determines how likely this group is an actual spammer group, called suspicious score. The suspicious score is calculated out of the key indicators given by the individual feature assigned to each group.

The paper is structured as follows. Section II introduces related work in the field of spam detection. Section III gives more detailed information about the used Yelp.com dataset, platform and review system. Section IV describes our determined features. Section V our model which aims to find opinion spamming groups gets described. Section VI reports the experimental results. Finally, Section VII concludes and discusses the paper.

II. Related Work

Over the last years the field of spam review detection has been experienced plenty of attention. The existing researches can be divided based on the data and features they analyze. Below are presented three different approaches to filter and detect spam.

1. *Review spam detection:* this approach is based on the data of the review (e.g., textual content, meta data, etc.) to detect and filter spam reviews.
2. *Review spammer detection:* this approach is aiming to find the spammer it self. This can be, for example, achieved by identifying atypical behavior, such as posting time or geolocation.
3. *Group spammer detection:* goal with this approach is to identify whole group of spammers due to the fact that group spammer are more influential and harmful than individual spammer.

Was haben andere gemacht
single fake review erkenne (3)
gruppen fake reviews erkennen
ergebnisse ...
gruppenansatz von paper CPM XY 2 andere gruppenansätze (man kann Twist nennen)

—- In this section we show important work which has been made by other researchers like [3] and [1].

In this section we also want to highlight why this approach can be beneficial for this research area.

III. Yelps Review System

In this section we shortly describe how Yelps review system works.
you can give stars to business
free for all
yelp filters reviews on its own
freitext
yelp hat 20 fake reviews
yelp filterd selber

IV. Dataset

The dataset we use is publically available on Yelp.com and consists of real-world reviews about different businesses (e.g., restaurants, hotels, etc.). Yelp is a recommendations platform based on user-generated reviews. The dataset contains information about reviews, businesses, user data, and geographical data. More than 8 million reviews and more than 200.000 businesses are included in the dataset. For this paper, we look at reviews of 2016 that appeared in the area of Charlotte in North Carolina USA. In total, we examine 51261 reviews and 5954 businesses.

V. Feature

note: add 0 non spamming 1 suspiscous spamming

In the field of spam, detection features can be divided into reviewer-centric and review-centric features. Review-centric features are built upon the raw text review itself and perceive what is written while reviewer-centric features are built upon metadata and describe the behavior of the reviewer within the given review system.

Burstiness (BST): This feature reflects the behavior of spammers who are only active in a short timeframe with a given account. Within that timeframe, they try to write multiple reviews to maximize their impact.

$$BST(r) = \begin{cases} 0 & , L(r) - F(r) > t \\ 1 - \frac{L(r)-F(r}{t} & , \text{otherwise} \end{cases}$$

The BST is described in [2] where $L(r)$ is the last review of a reviewer and $F(r)$ is the first one. We choose a threshold of $t = 28$ which refers to 28 days. The BST value of the group is the mean value of all given BST values of each unique reviewer in that group.

Extreme Rating (EXT): Given the 5-star rating scale of Yelp, spammers are likely to give extreme ratings due to the nature of impacting the general opinion in the given review system about the target product the most up or downwards [3].

$$EXT(r) = \begin{cases} 0 & , R(r) \in \{2, 3, 4\} \\ 1 & , R(r) \in \{1, 5\} \end{cases}$$

Given all reviews of a reviewer $R(r)$ the EXT is 1 if all of them are either 1 or 5-star rating reviews. The EXT value of a group is then calculated by the mean value of all EXT values within a certain group.

Cosine Similarity (CS): We assume that reviewers which are organized within a group of spammers are likely to post multiple duplicate/near duplicate reviews distributed among their target products. We define the CS of a group by comparing each review with one other:

$$CS(g) = \begin{cases} mean(cosine(v_j, v_k), v_j \in \{R(r_j)\}, v_k \in \{R(r_k)\}) \end{cases}$$

Given a group g the CS is defined as the mean value of each pairwise calculated cosine similarity of two reviews v_j and v_k among all reviews within that group.

VI. Model

Our approach to find groups/communities who are working together

1. undirected Graph (G) 2. We look at the reviews of each business 3. Then, if 2 reviewer commented on the same business within in specific time window (alpha=6, in days) and the same star rating, then we add them to the Graph G 4.

Twist nennen (+ Grenzwertvariablen)
Image of Groups

VII. Results/Analysis

In this section we show our plots and describe the results/interpretation.

look at artecicial groups

- wie gut finden wir unseren suspicious score - reasoning über correlation - reasoning über vermutlichen prozentualen anteil -

VIII. Conclusion

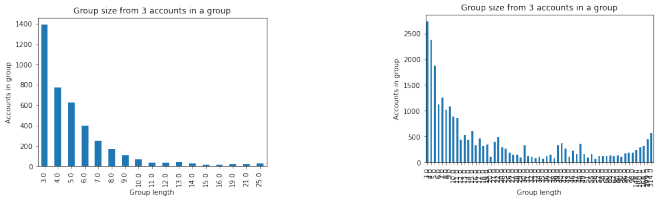
Twist nennen und sagen ob es eine gute Idee war

Finale interpretations der Ergebnisse - Was haben wir eigentlich gemacht - was war das ziel - gruppen erkennen, die fake reviews schreiben - haben wir das erreicht

Was machen wir nächstes mal anders + Empfehlungen

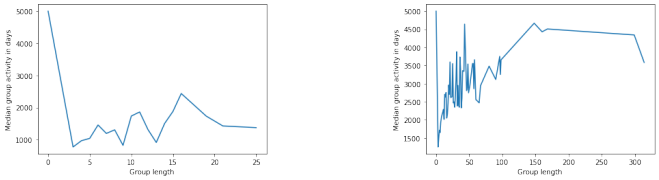
IX. Experimental plots

200.000 rows of the review json



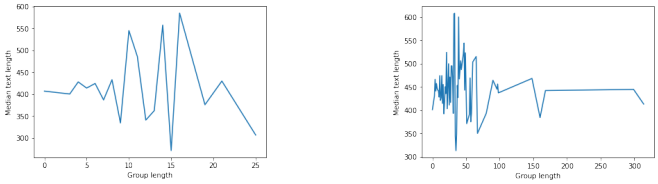
(a) Positive (b) Negative

Figure 1: Groups



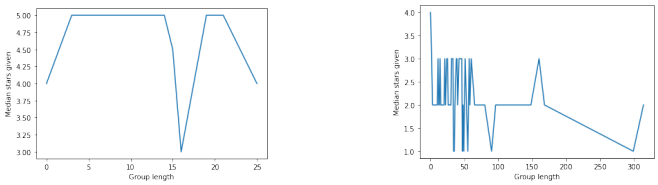
(a) Positive (b) Negative

Figure 2: Activity



(a) Positive (b) Negative

Figure 3: Length



(a) Positive (b) Negative

Figure 4: Stars

References

[1] Euijin Choo, Ting Yu, and Min Chi. Detecting opinion spammer groups through community discovery and sentiment analysis. In *IFIP annual conference on data and applications security and privacy*, pages 170–187. Springer, 2015.

[2] Chuang Ma Guangxia Xu, Mengxiao Hu. Gscpm: Cpm-based group spamming detection in online product reviews. In *IEEE International Conference on Communications (ICC)*, 2019.

[3] Arjun Mukherjee, Abhinav Kumar, Bing Liu, Junhui Wang, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. Spotting opinion spammers using behavioral footprints. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 632–640, 2013.