The Language of Deceivers: Linguistic Features of Crowdfunding Scams

Wafa Shafqat, Seunghun Lee, Sehrish Malik, Hyun-chul Kim*
Sangmyung University, Cheonan, South Korea
{wafashafqat92, mr.leesh90, serrym29, hyunchulk}@gmail.com

ABSTRACT

Crowdfunding sites with recent explosive growth are equally attractive platforms for swindlers or scammers. Though the growing number of articles on crowdfunding scams indicate that the fraud threats are accelerating, there has been little knowledge on the scamming practices and patterns. The key contribution of this research is to discover the hidden clues in the text by exploring linguistic features to distinguish scam campaigns from non-scams. Our results indicate that by providing less information and writing more carefully (and less informally), scammers deliberately try to deceive people; (i) they use less number of words, verbs, and sentences in their campaign pages. (ii) scammers make less typographical errors, 4.5-4.7 times lower than non-scammers.(iii) Expressivity of scams is 2.6-8.5 times lower as well.

Keywords

Crowdfunding Scam; Kickstarter; Linguistic Analysis

1. INTRODUCTION

Crowdfunding has significantly upheaved in recent years in terms of popularity and success. In 2014, global crowdfunding has experienced an explosive growth of 167% with $16.2 billion raised. Kickstarter.com, the largest crowdfunding site, reportedly raised more than $1 billion funding from 7.7 million investors in 2015 [1]. As crowdfunding becomes mainstream, it also creates a great potential for scams, due to its openness, flexible requirements, flexibility in defining a purpose and lack of legal resources for investors [1]. A well-known attempted crowdfunding scam is “Kobe red beef jerky" on Kickstarter, which almost enabled a fraud of $120,309 from 3,252 backers (i.e., investors) in just less than four weeks [1]. Fortunately, it was discovered to be a scam and then suspended, just hours before the fundraising ends.

According to the Financial Crime Enforcement Network, 171% increase in crowdfunding mentions in their Suspicious Activity Report filings between 2013 and 2015, is an indicator that crowdfunding is now confronting the rising and challenging issues of fraud and deception. The growing number of blistering articles on crowdfunding scams and discussions on di□erent platforms like $reddit.com$, proclaim perturbation, and a sign of warning that the general public needs to be protected from the imminent assault of the fraudsters, particularly in order for crowdfunding to become a viable and trustworthy source of capital for new startup companies. However, despite the increasing threat of crowdfunding scams, we know very little about them, as there has been little attempt to provide a detailed analytical study on the possible common properties and practices of those scams. Our study fills this gap by collecting and analyzing scam campaigns on the largest crowdfunding site, Kickstarter, using linguistic structure and cue analysis as previous studies have applied to distinguish deceivers from the truth-tellers [2][3].

As a first step towards an empirically-grounded understanding of crowdfunding scams, this work makes three key contributions to the field: (1) We find that scammers deliberately try to deceive people by intentionally providing less information and writing more carefully, and less informally. For example, when compared with non-scammers, i) scammers use 11% , 56.8%, 57.7% less number of words in their $campaign$ section, $updates$ section and $comments$ section, respectively, of the project page. ii) scammers use 8.34%, 55.11% and 50.31% less number of verbs in their $campaign$, $updates$ and $comments$ respectively. (2) We also find scammers make less typographical errors; non-scammers make around 4.5-4.7 times more spelling errors than scammers, consistently in all the $campaign$, $updates$ and $comments$ sections; (3) Expressivity (which will be defined in Section 2) of non scammers, particularly in $comments$ section of the project campaign is 8.5 times higher than scammers, which implies, expressiveness of the language of scammers is low, due to over-control and less conviction about what is being said [3].

2. METHODOLOGY

Kickstarter campaigns primarily include (i) description of the project known as $campaign$, (ii)$updates$ where project creators report their progress, and (iii) $comments$ where both backers and creator freely leave and share their posts. As Kickstarter does not make any o□cial records of scam cases publicly available, we collected data in July 2015, for campaigns accused of being scams by public forums or media such as $kickscammed.com$, $reddit.com$, or $Facebook\ page$ (Crowdfunding Projects that Never Delivered)[1], etc., and composed a list of 140 fraud cases that fall under the suspicion of large community, along with the disputed details or claims. These campaigns, in total, successfully have deceived 175,260 backers by raising $11.5 million. This list was then conservatively refined into our list of 25 most suspicious scam campaigns based on the following criteria: i) No promised deliveries were made to the backers for more than 7 months after the expected delivery date (though admittedly there still might be a possibility someone might

*Corresponding author: Hyun-chul Kim (hyunchulk@gmail.com)

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2889356.

[1]https://www.facebook.com/groups/1380253912299062/

have received the product but never left a comment at any place we looked for). ii) Project creator has not made any new *updates* for the last 7 months i.e., since Dec. 2014. iii) covered by press media as a fraud case, *forbes.com*, *CNN-Money.com*, etc.

We also collected data of 150 Non-Scam campaigns (i.e., successfully delivered projects) in July 2015 from (i) CNN Money's list of delivered projects and (ii) campaigns listed at *outgrow.me*, a marketplace for successfully crowdfunded products. We admit that neither of our datasets for scams and non-scams are that large, yet they were still good enough for us to find out meaningful patterns, as shown below.

To investigate the way scammers use language, we used linguistic structure and cue analysis [2][3], adopting Zhou et al.'s constructs and their respective definitions for each variable [3]. These linguistic constructs are divided into Quantity (# of words, verbs, sentences), Complexity (Average # of words per sentence , Average # of characters per word, Pausality - Average # of punctuation marks per sentence, Average # of clauses per sentence), Diversity (Redundancy - # of function words per sentence), Non-immediacy (Group references - first person plural pronoun, Self references - first person singular pronoun), Expressivity (ratio of adjectives plus adverbs to nouns plus verbs), Informality (Typo Ratio - Average # of misspelled words), etc. We extracted all the linguistic cues using LIWC[2] and tested separately for *campaign*, *updates*, and *comments* sections of each project. Results are shown in Table 1, where we have only included cues with the most notable results, given space limitations.

Linguistic Cues	Campaign Scam (Non-Scam)	Updates Scam (Non-Scam)	Comments Scam (Non-Scam)
Word Count	1109.3 (1247.7)	2848 (6584)	2491 (5887.1)
Sentence Count	60 (70)	39.7 (77.8)	156.8 (319.5)
Verb Count	55 (60)	154 (343)	138.2 (278.1)
Typo Ratio	0.28 (1.28)	0.8 (3.7)	1.2 (5.7)
Expressiveness	1.3 (3.4)	1.9 (4.9)	1.2 (10.2)

Table 1: Avg. Linguistic Cues of Scams and Non-scams

3. PRELIMINARY RESULTS

From Table 1, when compared with non scammers we observe that scammers use i) 11% fewer words in *campaign*, 56.75% fewer words in *updates* and 57.69% fewer words in *comments*, ii) 14.3%, 49% and 50.93% fewer sentences in *campaign*, *updates* and *comments* respectively, and iii) 8.34%, 55.11% and 50.31% fewer verbs in *campaign*, *updates* and *comments* respectively. With an average of 1.3, 3.7 and 5.7, Typo Ratio of non-scam *campaign*, *updates* and *comments* respectively is 4.5-4.7 times higher than scammers, as also shown in six scattered plots in Figure 1. Similarly, the Expressivity of scammmers is lower by 63% in *campaign*, 61% in *updates* and 88% in *comments*. Also in Figure 1, we observe expressivity of non-scammers is much higher particularly in *comments*, thus forming a di□erent cluster for each group of projects, in the two-dimensional feature space of Typo Ratio and Expressivity. Our results are corroborated with previous studies on the behavior of scammers in online dating profiles, in TA-CMC (Text-based Asynchronous Computer Mediated Communications) and in non-interactive situations [2][3], where scammers are less forthcoming than truth-tellers as they use fewer words to curtail the information that could later be verified. In TA-CMC, expressivity of deceitful senders is low because of their over-control and less conviction in what they say.

[2]http://liwc.wpengine.com

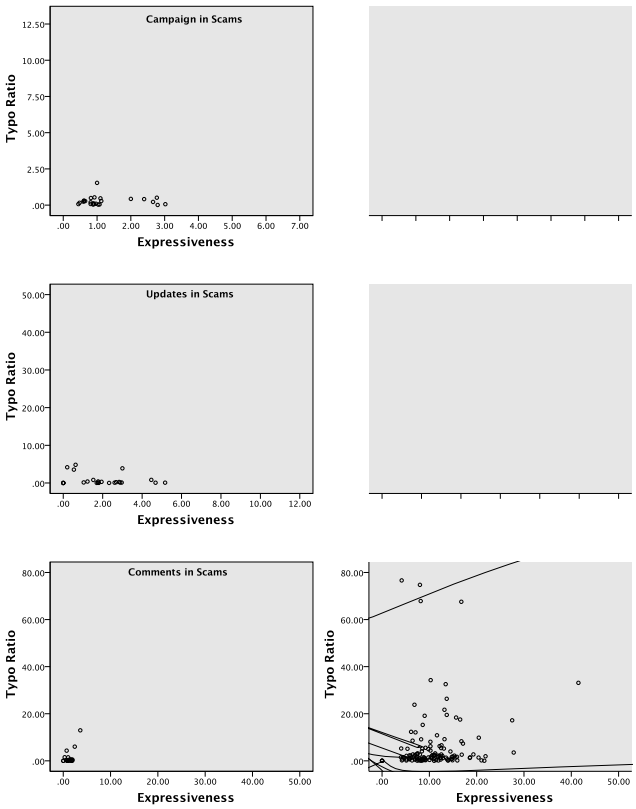

Figure 1: Scam (left) and Non-Scam (right) Campaigns, Updates, and Comments over Expressivity & Typo Ratio

Our ongoing work includes (i) collecting, validating, analyzing, and publicly releasing a ground truth dataset consisting of hundreds to thousands of real crowdfunding scam cases. Researchers need data to progress in this field to understand, detect, and prevent crowdfunding scams, particularly at an earlier phase of such campaigns, (ii) in-depth investigation on more features useful for detecting scams, such as communication behavior of scammers (and backers), their contents, as well as temporal and spatial attributes, etc.

4. ACKNOWLEDGMENTS

This research was supported by Basic Science Research Program through the NRF (National Research Foundation of Korea) funded by the Ministry of Education (NRF-2013R1 A1A2010474).

References

[1] T. H. Ho. Social purpose corporations: The next targets for greenwashing practices and crowdfunding scams. *Seattle J. Soc. Just.*, 13:935–1015, 2015.

[2] C. L. Toma and J. T. Hancock. Reading between the lines: linguistic cues to deception in online dating profiles. In *ACM CSCW*, pages 5–8, 2010.

[3] L. Zhou, J. K. Burgoon, J. F. Nunamaker, and D. Twitchell. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group decision and negotiation*, 13(1):81–106, 2004.