Soft Computing and Machine Intelligence

Soft Computing and Machine Intelligence Journal, Vol (1), Issue (1), 2021

Sentiment Analysis to Classify Scams in Crowdfunding

Wafa shafqat [a] and Yung-cheol byun [a] *

(a) Department of Computer Engineering, Jeju National University, Jeju 63243, Korea

* Corresponding author: ycb@jejunu.ac.kr

Abstract: The accelerated growth of the internet and the enormous amount of data availability has become the primary reason for machine learning applications for data analysis and, more specifically, pattern recognition and decision making. In this paper, we focused on the crowdfunding site Kickstarter and collected the comments in order to apply neural networks to classify the projects based on the sentiments of backers. The power of customer reviews and sentiment analysis has motivated us to apply this technique in crowdfunding to find timely indications and identify suspicious activities and mitigate the risk of money loss.

Keywords: Crowdfunding; Sentiment Analysis; Machine Learning

1. Introduction

The tremendous increase in the data availability over the past few years has resulted in the emergence of different techniques and tools, which help us understand and analyze the available online data in the best way possible. Data is everywhere, in different forms, e.g., text, audio, video, etc. Primarily, almost every other site gives its customers a right to leave their comments or reviews on their products so that they can reevaluate and improvise their experiences. There have been many studies on sentiment classification on Social networks, e.g., Twitter [1], Facebook, etc. Crowdfunding sites are gaining popularity at an accelerated rate over the past few decades. Analysis of user data on crowdfunding sites can be helpful in many ways; it can help improve the user experience. User's comments related to a specific product can help others decide on that product. This analysis eventually helps mitigate the risk of loss of money or frauds as crowdfunding sites are facing these sorts of challenges at the same time [2].

This study also focuses on the crowdfunding site, kickstaer.com; one of the most famous and leading crowdfunding sites. We have come up with other project categories, too, including the successful and failed projects. Regardless of the status of a project, it can be a fraudulent campaign or a genuine one. It is challenging for a user to identify that at first. Figure 1, shows a screenshot of the project which got very famous within days, but later it was suspended by Kickstarter as they recorded some suspicious activities. It was not easy for a user to identify that project is not genuine initially as it seemed very interesting, and people were loving it. We have come up with other project categories too including the successful and failed projects. Regardless of the status of a project, it can be a fraudulent campaign or a genuine one.

The rest of the paper has related work in section 2, data description and analysis in section 3, next section 4, explains the methodology, and in section 5, experimental results are discussed.

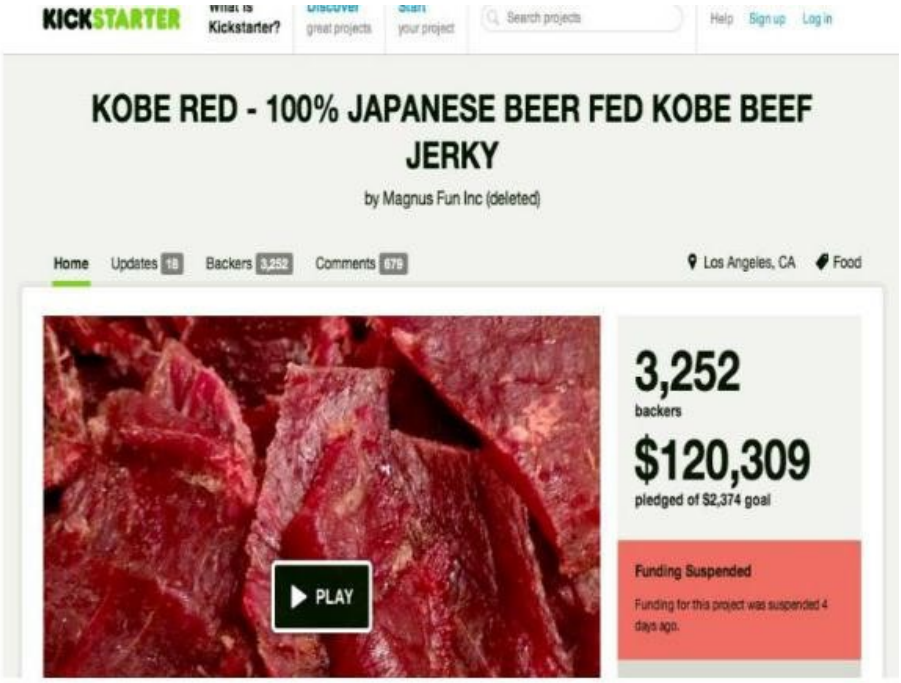

Figure 1. Screenshot of a Scam Campaign.

2. Related work

With the advancements and rapid growth of social networks, machine learning and NLP have attracted many researchers, mainly towards sentiment analysis and opinion mining. In [3], authors have used overall sentiments of a document for its classification using standard machine learning algorithms like Naïve Bayes and Support Vector Machine, etc. In another survey paper from the Liu et al. [4] focused on sentiment aware applications to overcome and find the solution to the new challenges

3. Data Analysis

The above Table 1 describes our data set briefly. We targeted kickstarter.com as our primary data source. Data for both categories were collected and stored separately. The above-mentioned table categories refer to the projects that managed to take public funds but never delivered are called Scams. Similarly, projects that were successfully delivered are referred to as Non-scams.

We used a python-based scroller for data collection, which, when given a project's URL or ID, fetches all the relevant fields related to that project. These fields include; Project's generic features, e.g., its title, creator's information, Date, category etc. The challenging task was the collection of comments as it is spanned over multiple pages. There was a high variation in commenting styles; as backers can show their emotions through emoticons and use slang, there is no fixed writing pattern. It was challenging to extract that data like the length of comment can vary from just a few letters to a few hundred words. There is no specific separator between comments, and it is hard to distinguish where one comment ends and the next starts, especially when these comments are from the same backer.

For this, though many checks had to be incorporated to get comments into a specified format (with an accuracy of 9%), some comments still had to be parsed manually as they usually contain many unique characters.

Table 1. Data description.

Data	Explanation
Total Projects	300
Categories	Scam/Non scam
Data Type	Text
Source	Kickstarter.com

4. Methodology

The percentage of successfully funded Kickstarter projects as of November 2020 is just 38.21 % [5]. The primary and most important part of our methodology is data collection. We first selected project IDs for both categories. Then these IDs were accessed and data was collected. This data was stored separately for comparison purposes.

From the data collection and extraction of relevant data, how we stepped forward to find useful analytical results; is elaborated in Figure 2. Our neural

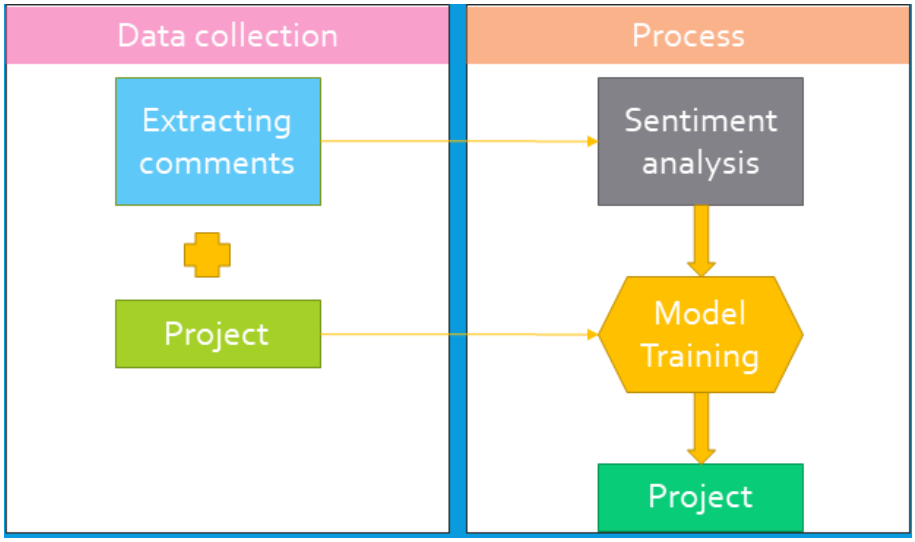

Figure 2. Process details to classify a project based.

The network model was trained based on the sentiment analysis results along with the project features mentioned above. We collected the commenter's name, the date when it was published, and the text itself for each comment. This data was stored to be later used for training. We performed sentiment analysis on this data.

The result of sentiment analysis was a number ranging between +3 and -3, where +3 represents the highest satisfaction (positive comment) and -3, being the lowest level of satisfaction (negative comment) and 0 represents a neutral comment.

Category	comment	Money	Amount	Status	SentimentScore	Location	Type	Date
Scam	0	1,500	0	canceled	-2	Chatha	Video Games	Jul 9, 201
Non scam	13	500,000	685	canceled	0	Atlant	Video Games	Feb 4, 20
Scam	153	5,000	6,007.01	successful	-1	Salt Lake (	Illustration	Jul 30, 201
Scam	76	50,000	14,984	failed	0	San Franc	Design	Apr 9, 20
Scam	123	20,000	12,803	suspende	-3	Vancouve	Hardware	Oct 31, 20
Non scam	83	80,000	4,739	canceled	2	Hollywood	Video Games	Apr 26, 20
Scam	88	6,700	7,583	suspende	1	Los Ange	Product Design	Aug 6, 20
Scam	157	1,000	4,738.70	successful	-2	North W	Tabletop Games	Aug 12, 20
Non scam	72	25,000	78,481	suspende	2	Ashevil	Technology	Jan 19, 20
Non scam	304	1,500	4,329.46	successful	2	Santa Bar	Video Games	Feb 29, 20

Figure 3. Screen shot of Data.

Our data looks like as shown in above Figure 3. This is just a screenshot of a small chunk of our data. It has many other fields as well, which have been described above.

These sentiment scores and other project features were fed to our machine learning model to predict the project class as scam or non-scam. The Neural network model takes the data saved in a .csv file and process it. As we have different types of training features, we perform normalization before the training process, as shown below in Figure 4.

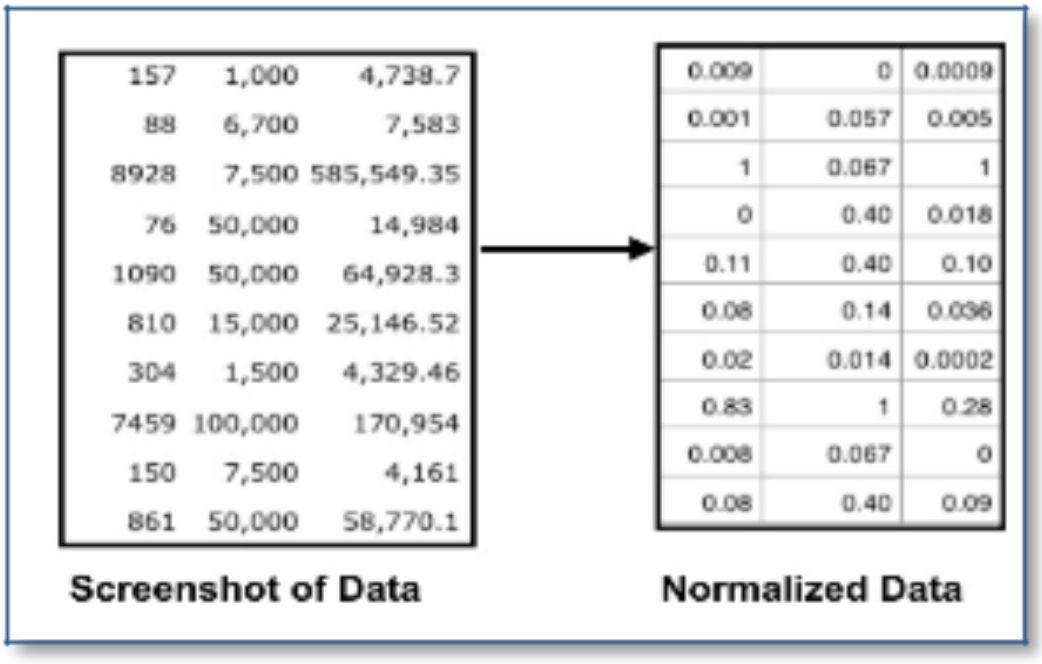

Figure 4. Screenshot of Normalized data.

5. Experimental Results

The following graph in Figure 5, shows the comments by backers over time for a very successful and non-scam project. The x-axis shows the timeline, and all comments were recorded for the fundraising period. This analysis was performed to check how backers' involvement over time in terms of their comments or pledging amount. The blue bars represent the amount of money pledged, and the red bars represent when a creator updates the project. We can see that the behaviors and involvement of backers change with the activities of the creator. An update from the creator can affect the sentiments of his backers positively or negatively. It can be noticed that as this project was genuine

and the creator was updating over time, that is why we can see more positive and neutral comments by backers. On the other side, in Figure 6, we have an example of a scam project, where it can be observed that negative comments are increasing over time compared with positive comments.

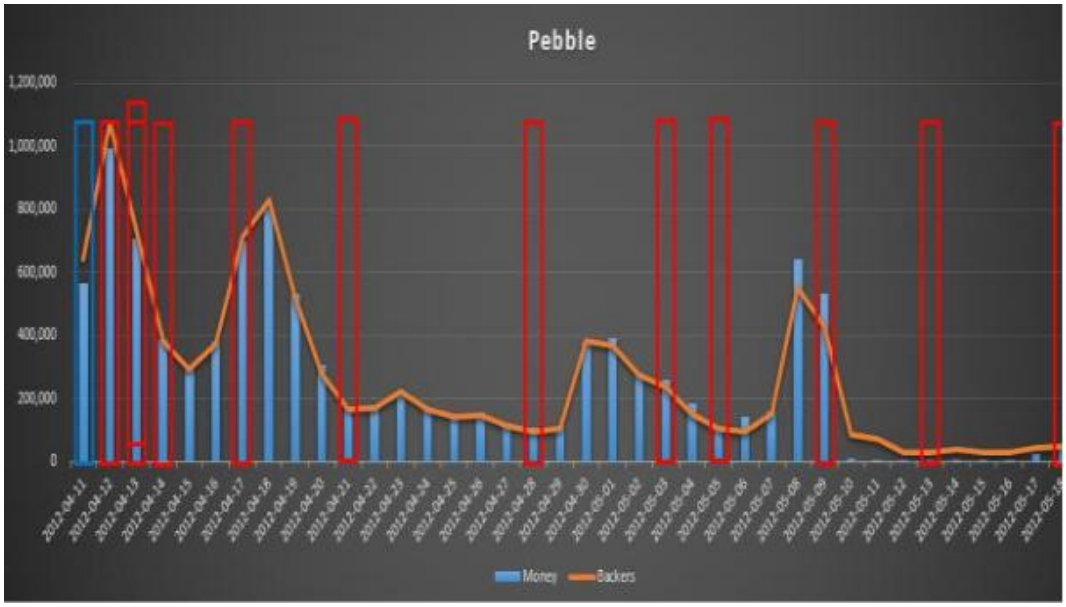

Figure 5. Time series Analysis of a Non Scam campaign based on Pledging amount.

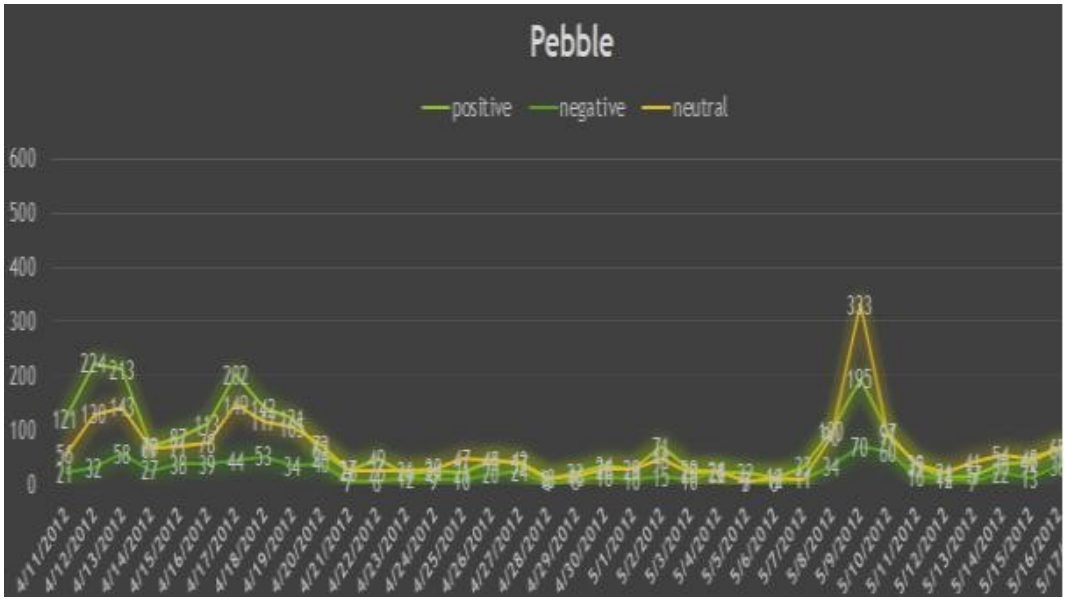

Figure 6. Comments categorization based on sentiments.

In Figure 7, bars in red indicate negative comments and green and yellow lines show positive and neutral comments, respectively. This graph is for an example scam case.

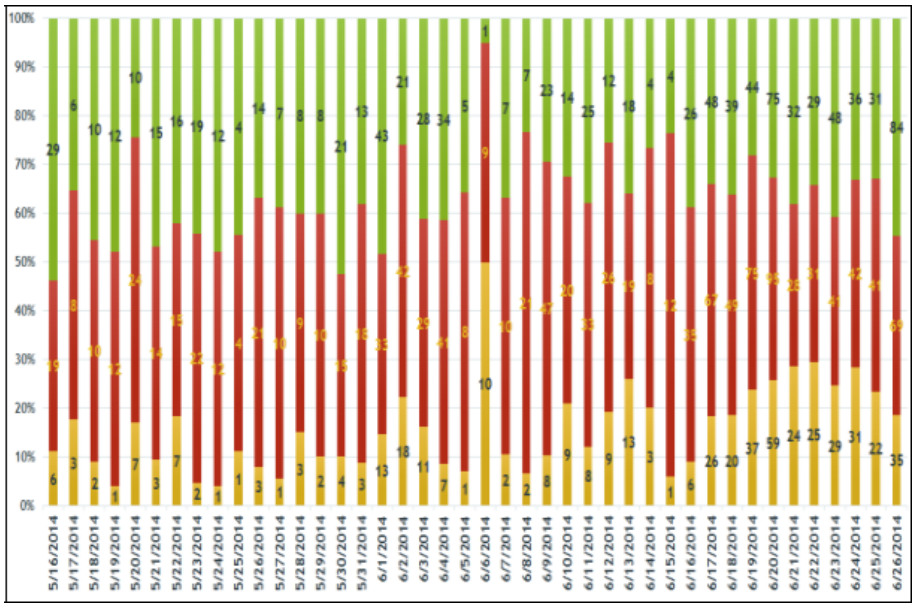

Figure 7. Sentiment Analysis of a Scam Campaign Comments over time.

For the purpose of training our model, we used 70% of our data set as training data and rest of 30% was used for testing. Our results show that sentiment features and other project features can play an influential role in classifying a project into scam or non-scam categories. Though, including more features and more levels of sentiment analysis, the results can be improved.

The data set was too large as we had a sentiment score for each comment in a single project along with generic project features. We aggregated the sentiment scores for comments based on their publishing dates, i.e., for each project, we can divide its period into three stages; funding period, after funding period, till expected delivery date, and after the expected delivery till current date. This way, it became easier to understand and analyze the commenting behavior and attached sentiments of backers.

Following Table 2 shows the summary of sentiment results for test data.

Table 2. Sentiment Analysis Results.

Class	Positive	Negative	Neutral
Scam	2026	3216	6696
Non Scam	5923	1606	1186

6. Conclusion

Though crowdfunding is getting popular with each passing day, it is also facing challenging risks and threats of scamming and fraudulent behaviors at the very same time. This study uses sentiment analysis to analyze the comments on a crowdfunding project to classify it into a category of scam or non-scam campaign. Our results and analysis give us an idea that using this information can help us find different patterns of emotions or sentiments over time. We aim to use more features and analysis tools to improve our results and find other dynamics.

References

1. Go, A.; Bhayani, R.; Huang, L. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford* **2009**, *1*, 2009.
2. Heminway, J.M. Securities crowdfunding and investor protection. *CESifo DICE Report* **2016**, *14*, 11–15.

3. Pang, B.; Lee, L.; Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques. *arXiv preprint cs/0205070* **2002**.
4. Liu, B.; Zhang, L. A survey of opinion mining and sentiment analysis. In *Mining text data*; Springer, 2012; pp. 415–463.
5. Department, S.R. Percentage of successfully funded Kickstarter projects as of November 2020, 2020. [Available online], https://www.statista.com/statistics/235405/kickstarter-project-funding-success-rate/.