Analysis of Funding Microlending Campaigns

Nellie Spektor

Abstract

In the last few years we have seen remarkable growth in crowdfunding websites and with this growth, many researchers are gaining interest in what makes a campaign successful on websites such as Kickstarter and Indiegogo. There has also been growth in another variety of crowdfunding, charitable microlending on websites such as Kiva. Kiva is a website where small businesses or community groups working with charitable foundations request loans and if they are approved, the borrowers pay the lenders back. The incentives for lenders on charitable microlending sites are very different from those on Kickstarter, therefore I have analyzed loan solicitation texts on Kiva in an effort to understand what these incentives are and create a program that can predict whether a campaign was funded based on its linguistic features.

1 Introduction

All over the world, there exist motivated people who have an idea for a project or a business and just need some funding to get off the ground. These ideas range a wide gamut, from opening a clothing store in Nicaragua to launching an app in Silicon Valley. A magnitude of research has been done fields such as psychology regarding persuasion; and with advent of the internet and with more and more people turning to online crowdfunding, there is a new direction that this research has taken.

I am exploring a specific kind of crowdfunding website, Kiva.com. Kiva is different from most other crowdfunding sites because it is not reward-based; it is based on charity. This leads most people to hypothesize that incentives of people loaning money on Kiva are similar to the incentives that drive people to donate to charity. However, unlike charity, lenders on Kiva get their money back, introducing financial sense into their incentives. A quick glance on which loan solicitations are funded lead me to believe that the degree of trustworthiness the borrower is able to portray plays a very important role in which loans get funded. My research into which phrases are likely to lead to success on Kiva can be used to analyze these various incentives. I posit some preliminary analysis at the end of this paper.

An example of a funded loan solicitation text on Kiva:

“Leonora is 43 years old, with three children. With two children in school, Leonora works very hard to provide for them.

“Leonora runs a general store in the Philippines and requested a PHP 6000 loan through NWTF to buy items to sell like beverages, canned goods, junk foods, home product care, and other groceries.

“Leonora has been in this business for 4 years and sells a variety of items such as rice, sugar, soap, and other groceries.

“In the future, Leonora would like to save enough money so she could afford to send her children to college.”

In the above passage, you can see that Leonora is hardworking, and has a sustainable source of income as well as a concrete goal. She is asking for a loan to buy things to sell in her store so that she can earn some money that can go towards a better life for her three children. This story not only tugs at our heartstrings and has a noble cause with which many of us sympathize, but it also conveys a sense of credibility that the loan will be paid back.

In contrast, here is an example text that was not funded:

"Hector is 63 years old. He lives in his own home with his life partner and daughter. He has made a living in agriculture for 20 years, and he sells his produce in neighboring villages. With his earnings, he manages to take care of his household expenses. He is requesting a loan to buy sulphate, seeds, insecticide, herbicide, formulas, etc. to improve his land which is shown in the photo. This will enable him to continue growing his crops, and earn more income."

Although Hector's story has some characteristics that would make it seem like a likely candidate for funding, most people would clearly rather loan their money to Leonora. The general impression is that Hector appears like he wants the loan for personal gain and it does not seem like Hector's income is stable enough to ensure that he will repay the loan.

I have written a system that can put this analysis into concrete numbers and predict whether a text will be funded based on machine learning and various features.

2 Related Work

Similar work has been done by Tanushee Mitra and Eric Gilbert of the Georgia Institute of Technology. A few years ago, they published a paper about phrases that predict success on another crowdfunding platform, Kickstarter (Mitra and Gilbert, 2014). They used a bag of words model and penalized logistic regression model to come up with a dataset of 20K phrases that predict success on Kickstarter and maybe other crowdfunding platforms. They used the Linguistic Inquiry Word Count Dictionary as well as other resources and techniques to analyze these phrases. My first step was to download these phrases and see how they were at predicting success on Kiva. As I predicted, however, the incentives are very different on Kiva and the phrases were only mildly predictive.

I drew my inspiration for this paper from a sociology paper called, *Charity or Investment: Linguistic Features of Identifiable Victim Effect in Microlending* by Natalia Levina and Semi Min of NYU, (2017). This paper focuses on using that same Linguistic Inquiry Word Count (LIWC) Dictionary to analyze whether the "identifiable victim effect" is an effective indicator of what gets funded on Kiva.

This "identifiable victim effect" is a theory that postulates that an identifiable victim solicits greater compassion, which leads to greater charitable giving. This has been thoroughly studied and multiple researchers have proven this phenomenon empirically. What Levina and Min are studying is whether this effect can be linguistically identified in funded Kiva texts. They drew on research which split identifiable victim effect into various psychological categories. The LIWC dictionary has a mapping of words to psychological categories, so in analyzing the Kiva text, Levina and Min looked for words that fit the categories of identifiable victim effect.

They found that funded texts were more likely to have words that indicated that there was an identifiable victim, however not every category was supported in their conclusions and there seemed to be something missing. The researchers posited that missing piece was an added factor that potential lenders seek out those people who they trust to repay their loan, and this factor does not have much to do with the existence of a victim.

In my other research, I understood that another serious flaw of this research on Kiva was focusing exclusively on individual words. A single word can carry many different meanings and if one fails to look at the features of the rest of the sentence, it is impossible to know what it means. This means that the results from the LIWC cannot be taken for granted. In a paper by Andrew Schwartz and others, I learned that one can achieve great results by ignoring highly ambiguous words, (Schwartz, 2013). Gustavo Paetzold and Lucia Specia developed a bootstrapping algorithm that uses word embedding to infer certain psycholinguistic properties better than the LIWC, (Paetzold, 2016). These are both very promising directions that I plan to explore in further research.

3 Method

Inspired by these papers, I decided to go about the question of why people loan on Kiva a slightly different way then Levina and Min did. I combined the approach of the Kickstarter researchers with the data of the Kiva researchers. As I mentioned, my first step was a naïve combination of these two research projects. I wrote a program to check if the phrases found by the Mitra paper to be predictive of Kickstarter success were also predictive for the Kiva data. The short answer is that they were not.

I decided to compile my own list of words and phrases (words, bigrams and trigrams to be precise) which would be indicative of success on Kiva. I did this by scraping the Kiva website for 150,000 loans from 2015-2016 with Beautiful Soup. The range of my data is restricted to loans that were funded in at most three days and is controlled for loan amount and repayment terms, among other things.

My next step was to clean the data and compile it into a development, training, and test set. My development and training set were an 80-20 split of 80% of the data and the test data was the remaining 20%. Cleaning the data included removing the html and unwanted newlines, punctuation, and numbers.

Following this, I looped through each funded and not-funded text in the training set and made a set of tokens, bigrams, and trigrams of each. I removed those phrases solely consisting of stop words, as defined by the Natural Language Toolkit (NLTK). I then added these to a master dictionary of phrases and their frequencies.

After looping through each text, I generated a list of the most common phrases. I have included this list in the Results section of this document.

My final step step was to use scikit learn's Random Forest Classifier to create a system that is given an un-classified set of Kiva loan solicitation texts and predicts whether it was funded or not. My first step in this part of the process was deciding which machine learning program to use. I read about various systems including Maximum Entropy and Naïve Bayes. I came to the conclusion that scikit learn was very good for my intentions and not too difficult to learn. I chose not to use Maximum Entropy because it was part of one of our homework assignments and I wanted to take the opportunity to learn about a new system. Perhaps I will implement a Maximum Entropy System later and see the different results and compare these methodologies. The program I built using Scikit learn's Random Forest Classifier achieves an 86% accuracy score on the test set.

4 Results

Here are the top 50 words and phrases seen in funded texts on Kiva:

this loan	33263
the organization	31276
access to	25953
low income	25890
to help	18717
to buy	17879
in rural	17455
its clients	14775
rural communities	14713
kiva's	14685
financial services	14126
part of	14016
to expand	13594

to provide	13581
her business	13479
her family	13273
a loan	13165
rural areas	13034
kiva lenders	12954
loan is	12881
years old	12376
services to	12013
the loan	11387
loan to	11323
this loan is	11297
is years	10362
women for	10249
for tomorrow	10248
negros women	10247
tomorrow foundation	10247
negros women for	10247
women for tomorrow	10247
for tomorrow foundation	10247
is years old	10215
due to	10006
able to	9858
used to	9774
their families	9476
to support	9262
to improve	9031
women in	8983
variety of	8912
organization's	8833
one of	8817
micro entrepreneurs	8692
to low	8584
be used	8381
to low income	8250
lenders funds	8156
and rural	8151

In Comparing this list to the list generated in the Kickstarter study, it is clear that the incentives are very different. Common Kickstarter success phrases include 'pledgers will receive', indicating that reciprocity is very important. On the other hand, Kiva lenders care about communities and organizations.

There have not been other systems attempting to predict success on Kiva, so I lack a perfect baseline system with which I can compare. The papers I have mentioned about Kickstarter and Kiva did not include a prediction system and its results. When looking at the results of other document classification systems it is clear that my program has room for improvement. In particular, the winner of the Kaggle movie review classification challenge achieves an accuracy of 99%. I believe this has to do with what machine learning algorithm these systems use and more analysis of more sophisticated features.

5 Future Work

For this project, I have made the baseline system that I set out to create. I have many ideas for how I want to proceed with this research. As I mentioned before, I would like to take these predictive phrases that I found and analyze them with the LIWC and other resources to explore whether they backup Levina and Min's findings about the identifiable victim effect. My hypothesis is that the results will be largely similar, but with more definitive data given the more complex approach. Next, I will analyze what other psycholinguistic properties are present in funded texts that are not present in texts that were not funded. The motivation behind this research is to help discover if lenders are motivated only by charity or also for personal gain, or at minimally avoidance of personal loss.

In this analysis I would like to go beyond the LIWC and explore word sense disambiguation techniques as well as the bootstrapping algorithm created by Paetzold (2016).

In addition to my psycholinguistic analysis goals, I plan on exploring various machine learning techniques and discovering how my system yields results implementing a range of approaches.

References

Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*. O'Reilly Media Inc.

Gustavo Henrique Paetzold and Lucia Specia (June 17, 2016). Inferring Psycholinguistic Properties of Words. *Proceedings of NAACL-HLT 2016*, pages 435–440. http://www.aclweb.org/anthology/N16-1050

H. Andrew Schwartz, Johannes Eichstaedt, Lukasz Dziurzynski, Eduardo Blanco, Margaret L. Kern, Stephanie Ramones, Martin Seligman, and Lyle Ungar (June 13-14 2013). Choosing the Right Words: Characterizing and Reducing Error of the Word Count Approach. *Association for Computational Linguistics,* pages 296–305. http://www.aclweb.org/anthology/S13-1042

Natalia Levina and Semi Min (2017). Charity or Investment: Linguistic Features of Identifiable Victim Effect in Microlending. *Thirty Eighth International Conference on Information Systems*

Tanushree Mitra and Eric Gilbert (February 15–19, 2014). The Language that Gets People to Give: Phrases that Predict Success on Kickstarter. *Association for Computational Linguistics.* http://comp.social.gatech.edu/papers/cscw14.crowdfunding.mitra.pdf