Understanding pro-social landing: How lenders respond to loan descriptions on Kiva

Zi Yin
Department of Electrical Engineering
Stanford University
zyin@stanford.edu

Yuanyuan Shen
Graduate School of Business
Stanford University
yyshen@stanford.edu

1 Introduction

Kiva is one of the world's first online philanthropic microcredit crowdsourcing platform. Its mission is to connect people through lending to alleviate poverty. Founded in 2005, Kiva has attracted more than 2 million lenders and raised more than 700 million dollars with a 98.6% repayment rate.

On the demand side, Kiva partnered with local microcredit lending agencies in the third world to screen the borrowers and post their requests online. On the supply side, Kiva connects lenders online to fund the loans in increments of $25 dollars. The lenders are philanthropists who do not charge any interest for the loan. They browse through Kiva's online platform and select a fundraising loan to contribute. Each loan is listed on the website for up to a month (with some exceptions). Loans that are not fulfilled within a month will be expired. Depending on its amount, purpose (e.g. for education, for business) and description, each loan is funded at a different speed.

To remain competitive in the market, Kiva wants its loans to be funded as fast as possible. In this project, we will study how the loan descriptions will affect the funding speed of the loans. This is crucial for Kiva to provide guidance to its partners who write the descriptions for the loans.

In particular, we will investigate how loan descriptions affect the loan fulfillment days, and provide suggestions on how to write a good loan description to have one's loan quickly fulfilled.

2 Problem statement

We obtained Kiva's loan data from its API website. A data snapshot as of Dec 18, 2015 contains the information cumulative through that date. Table 1 summarizes the features of the loans we are able to obtain from the data set. We know whether each loan was fully funded and how fast it was funded. The funding time, ranging from 0 to 80 days, is the output for our model. The loan descriptions are paragraphs ranging from 50 to 300 words in length. We will apply models of deep learning in NLP to predict the funding speed and see which words and phrases are linked to higher fulfilling rate. We will also use the features in Table 1 as inputs to the hidden layers. A relevant study by Liu et al. [1] evaluates the impact of lenders' motivations on their funding enthusiasm applying standard Natural Language Processing models. They found that certain phrases in the lenders' motivations (filled online upon registration) are associated with higher activity levels. Here we will study the online platform from a demand side. Our work is also related to word classification of user-generated content and social media e.g., [2, 3]).

Table 1: Features for each loan

Variable	Description	Type
Status	Status of the loan (funded, repaid, etc)	Categorical
Sector	Purpose of the loan defined by Kiva	Categorical
Loan description	Story that explains the purpose of the loan	String
Loan amount	Amount that the borrower(s) requests	Numerical
Funded amount	Amount that has been raised on Kiva	Numerical
Delinquent	Whether the repayment was delayed	Binary
Partner ID	Id of the partner that manages the loan	Categorical
Posted date	Date that the loan was posted on Kiva	Time
Funded date	Date that the loan was fully funded on Kiva	Time
Planned expiration date	Planned date of fundraising expiration	Time
Disbursal date	Date that the loan is actually disbursed to the borrower(s)	Time
Borrower(s)'s gender	The gender of the borrower(s)	List of Binary

Our goal is to incorporate the loan descriptions as an important predictor of the loan fulfillment speed. Each description is a paragraph with an example given below:

> Tuipulotu T., 19, is single with no children. She has many years of experience in the Elei (traditional fabric printing) business. She sells to the general public 6 days per week. She has 2 previous loans with SPBD. She expects her weekly net cash flow to be 600 Tala (≈250 USD). SPBD loans are Tuipulotu 2019s only access to capital because she was never able to qualify for a loan with the traditional banks.

We use the square loss, i.e. $||\hat{y} - y||^2$ where y is the actual fulfilled days. There are several experiments we will run and compare:

1. The mean square error using the benchmark regression model and features excluding the loan description;
2. The mean square error using only the loan description and the language model;
3. The mean square error when both the loan description and other features are used.

3 Technical Approach and Models

TWe first tokenize the loan descriptions. We will apply several different networks (RNNs, LSTMs and Recursive neural networks) and evaluate their performance in terms of mean square errors.

The Kiva dataset is of size 5.5GB on disk, in its raw form. Since some loan descriptions are written in languages other than English (and hence do not fit in our word embeddings), and some loans have missing entries, we performed a data cleaning procedure before feeding the data into the models. There are 860,000 loans in total after data cleaning.

To train the model, we feed the loan descriptions into the recurrent neural network. At each unfolded time point, the output $\hat{y}_i^{(t)}$ is the predictor of the loan fulfillment days, and is compared to the true fulfillment day y_i. Hence the loss function corresponding to the ith data point, with loan description length T_i, is

$$\sum_{t=1}^{T_i} (\hat{y}_i^{(t)} - y_i)^2$$

We then sum over the training set and get the loss function we try to minimize:

$$\textbf{Loss} = \sum_i \sum_{t=1}^{T_i} (\hat{y}_i^{(t)} - y_i)^2$$

In the end, we will try incorporating the other features in Table 1 to the final hidden layer.

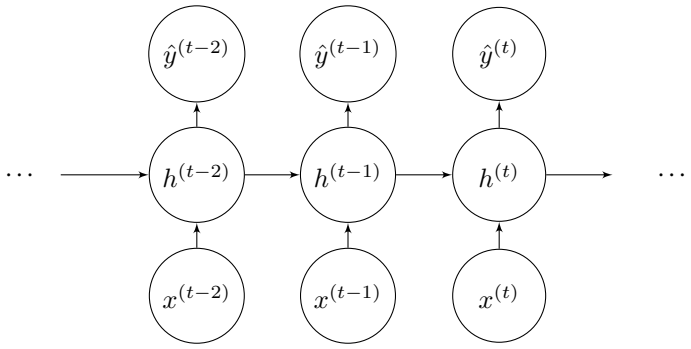

Figure 1: Network structure

4 Intermediate/Preliminary Experiments & Results

As a benchmark, we regress the funding time for the funded loans on sector, loan amount and partner Id. We get a RMSE of 8.25.

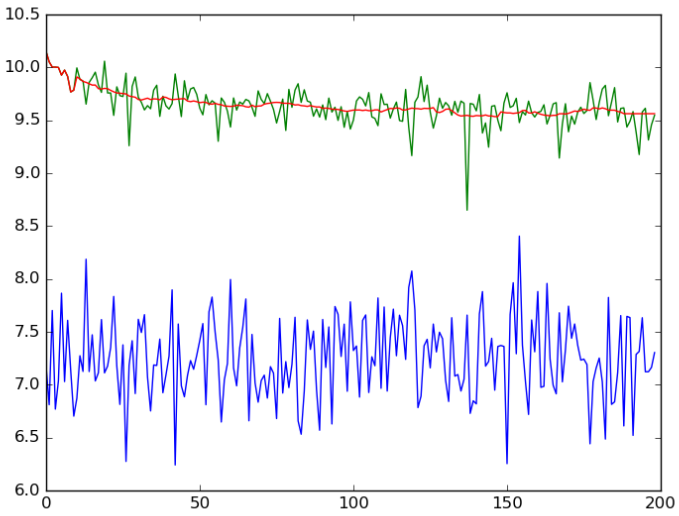

Figure 2: vanilla RNN training and validation error

We implemented a vanilla recurrent neural network, with one hidden layer of 144 neurons. Above is a figure of the training and validation errors during the training process. The blue curve is the training error; the green one represents the validation error (on a test data set), while the red line indicates the moving average. We are able to achieve a RMSE of 6.46 using this network on the final test dataset.

References

[1] Liu et al. (2012) "I Loan Because...": Understanding Motivations for Pro-Social Lending. *WSDM'12*.

[2] Agichtein et al. (2008) Finding high-quality content in social media. In *Proceedings of the international conference on Web search and web data mining*, pages 183-194. ACM, 2008.

[3] Sriram et al. (2010) Short text classification in twitter to improve information filtering. In *Proceeding of the 33rd international ACM SIGIR conference on research and development in information retrieval*, pages 841-842. ACM, 2010.