

**Predicting magnitude of success and failure on the Kiva micro-lending platform through linguistic analysis**

Christina Kronser

MSc Computer Science

University College London

Submission Date: XXXX

Supervised by

Dr. Licia Capra

Dr. Giacomo Livan

*This report is submitted as part requirement for the MSc Computer Science degree at UCL. It is substantially the result of my own work except where explicitly indicated in the text. The report may be freely copied and distributed provided the source is explicitly acknowledged.*

**Abstract**

XXX

**Table of Contents**

List of tables

List of figures

1. **Introduction**
   1. Motivation
   2. Thesis structure
2. **Related Work/ Background**
   1. Research conducted on Kiva
   2. Linguistic analysis of online sources (?)
3. **Project Scope**
4. **Data Set**
   1. Data source
   2. Data sets chosen
   3. Data cleaning
   4. Data engineering
      1. Explanatory variables
      2. Dependent variables
5. **Exploratory Analysis**
6. **Modelling Analysis**
   1. Methodology
   2. Results
      1. Modelling funding time
      2. Modelling funding gap
7. **Conclusion & Future Work**
   1. Conclusion
   2. Limitations
   3. Future work
8. **Bibliography**

**List of Tables**

**List of Figures**

**1 Introduction**

**1.1 Motivation**

Micro financing has become a vital component in today’s emerging economies. In 2006, the awareness for this rather new form of social entrepreneurship was substantially lifted when the Bangladeshi Muhammad Yunus was awarded the Nobel peace prize for his engagement related to micro financing and social entrepreneurship [REFERENCE]. During his Nobel lecture, Dr. Yunus put a significant emphasis on the fact that poverty and missing access to banking is a major threat to world peace with sixty percent of people living on only six percent of the world income. He was convinced that these tiny loans could make a disproportionate difference to a poor person’s live and were therefore part of the solution [REFERENCE: https://www.kiva.org/microfinance]. Inspired by Dr. Yunus’ years of work, the micro-funding organisation Kiva was established in 2005.

As pioneer in crowdfunding and nowadays largest micro-financing website [REFERENCE], the non-profit organization allows low-income entrepreneurs with limited or no banking options in the developing world to borrow money via the Internet from people in the developed world. Thereby, the lender might support the borrower in many different ways such as helping a mother to send her children to school, a farmer to raise new livestock or a women to expand her sewing business. The online lending platform works with a large network of microfinance institutions that distribute the funds in more than 30 countries. Since its beginnings Kiva enabled over 1.7M people to grant 2.9M borrowers loans valued at $1.18B in 82 countries. Repayment rate 97%.

The website asks borrowers to provide several attributes of their undertakings, resulting in the project profile that is ultimately viewed by the lender. It is reasonable to infer that borrowers try to create an interesting, promising profile that is perceived as trustworthy in order to convince lenders to invest in this project. Information such as the thematic sector (e.g. Health, agriculture, education), the geographic location and the borrower’s gender is fixed, leaving few leeway to shape the lender’s perception. Other characteristics such as the amount requested, the repayment duration in months and the project description can however be used to influence the lender’s impression.

One factor distinguishing Kiva from other peer-to-peer lending platforms is its mission to not only alleviate poverty through lending *in an environment of trust and transparency* but to also create lasting connections between borrowers and lenders [Matt article - REFERENCE]. It aims at establishing a new awareness and understanding what live in poverty and the difficult process to escape looks like. This is achieved through the unique stories told by Kiva’s borrowers, a few examples of which are a tomato farmer in India, a hotdog stand man in Cambodia or a carpenter in Gaza. These stories are not only featured on Kiva’s website but particularly, borrowers tell them through their projects’ textual descriptions. **This newly raised awareness helps build the human connection between borrower and lender which in turn might improve the borrower’s funding prospects.**

Due to rather loose guidelines provided by Kiva [TODO: REFERENCE], borrowers are offered lots of freedom in how to present their projects through their descriptions, i.e. which emotional tone to use, how much information to provide and which topics to address. Given all this freedom and the variety of factors to consider, how can borrowers leverage their project description to the fullest, thereby increasing the project’s attractiveness towards lenders?

This thesis therefore seeks to examine linguistic features in the project description that might affect a project’s success prospects. With a 95% funding rate, however, we aim to investigate factors that provide more granularity to the analysis, instead of the binary funding outcome. We therefore analyse the significance of linguistic parameters on the time it took for a successful project to get funded. Similarly, we investigate the effects of these parameters in the case of unsuccessful, i.e. not funded or expired, projects in terms of the amount by how much the loan missed its targeted amount. The findings may allow us to **XYZ**

awareness what live in poverty + escape process look like → through unique stories → help build human connection btw. borrower + lender → better funding prospects

This newly raised awareness fosters a connection between lenders and borrowers.

linguistic characteristics → help build connection → connection leads to better funding prospects in terms of speed

-----

Thereby, it seems reasonable to infer a certain relationship between linguistic characteristics such as the emotional tone of the description and the project’s funding prospects.

-----

where lenders and borrowers share a connection over a project. It seems reasonable to infer that borrowers seek to provide descriptions that adopt an emotional tone to attract funds.

-----

Some lenders are seeking to fund projects fulfilling certain criteria and relating to an area of their professional or personal interest. In order to keep the platform relevant

It aims at creating an environment of trust and transparency where lenders and borrowers create a long-term relationship.

----

is **constantly innovating to meet people’s diverse lending needs**. Whether it’s reinventing microfinance with more flexible terms, supporting community-wide projects or lowering costs to borrowers, we are always testing and learning.

----

Copied from BSC: Kiva’s aim is to foster an environment where lenders and borrowers share a connection over a project. It seems reasonable to infer that borrowers seek to provide descriptions that adopt an emotional tone to attract funds. The work would involve using sentiment analysis to decipher how emotional tones of project descriptions affect lending patterns.

the correlations between features analysed in this project and the results of emotional tone could give an indication what emotional tones inspire trust and persuasion to lend. The work would be carried out using computerised text analysers such as the Linguistic Inquiry and Word Count (LIWC) [21].

**1.2 Thesis structure**

This thesis is structured as follows:

* The section “Related Work” reviews research conducted on the micro-lending platform Kiva and helps to determine the nature of of this project’s research. It further describes work done related to linguistic analysis within online sources.

* The section “Project Scope” determines the main focus of this project and delimitates clearly between research questions within and outside of the scope of this project.

* The section “Data Set” explains where the data used in this project originates and justifies choices made to clean and manipulate the data. It further elaborates on the hypotheses-driven determination of the explanatory variables and their composition.

* The section “Exploratory Analysis” aims at providing a deeper understanding of the explanatory variables and those to predict. It is further used to discover areas of interest to explore deeper.

* The section “Modelling Analysis” provides an overview of the methodology used by explaining the descriptive and predictive models used for this project. The results are presented and finding interpreted within the context of the micro-lending landscape.

* Lastly, the section “Conclusion and Future Work” summarises the main findings of this project and identifies its limitations. It then hints to ideas on how this project could be extended and which further directions might be worth exploring.

**2 Related Work**

**2.1 XXX**

Research conducted on Kiva:

* Genevskz and Knutson (3):
  + neuropsychological mechanisms influence funding success → pos. features of photos promoted success (e.g. happier facial expressions)
  + surprisingly powerful, even more than loan amount
  + photograph identifiability positively associated with fundraising success, gender as well (female)

Lenders:

* Yang Liu, Roy Chen, Qiaozhu Mei, Yan Chen and Suzy Salib (4):
  + “I lend because “ → classify lenders’ self-stated motivations into 10 categories with ML classifiers
  + Findings: lenders with strong interests in one particular sector and that justify their loans with precise motivations → more frequent lenders
  + robust team activity → also more frequent lenders
* Jonathan Meer and Oren Rigbi (5)
  + effect of financial and social aspects of the loan (transaction costs, how relatable the project is (social distance))
  + Social distance measured in terms of same mother tongue or foreign language
  + Findings: transaction costs as a major influencer of lending behaviour relative to social distance

Lending teams:

* Yand and Kraut (7)
* Roy Chen, Wei Ai, Yan Chen, Qiaozhu Mei and Webb Phillips (8)

Research conducted on other platforms: (study behavior of lenders)



**3 Project Scope**

The scope of this thesis is to capture an underlying relationship between linguistic characteristics and the magnitude of a project’s success or failure. In a previous study, Bourdeau de Fontenay (2018) investigated factors of success and failure of Kiva projects and discovered that more than 95% of all projects are successfully funded. Therefore, this thesis will not investigate the binary outcome of funding success, but rather focus on a more granular level - how fast successful projects are funded, i.e. funding speed, and by which amount unsuccessful projects miss their target, i.e. funding gap. Bourdeau de Fontenay’s findings further revealed significant influences on funding success such as the loan amount asked for, the 15 thematic sectors that group loans according to which area they belong to, the borrower’s gender as well as the geographic location that divides the loans along the five continents. To specifically analyse the effects of linguistic characteristics we controlled for these already discovered and examined influences by narrowing down the provided data set as described in section 4. Thus, this project’s aim is to understand how linguistic features affect the magnitude of a project’s success or failure while controlling for other significant influences through a limited data set.

Similar to other micro-lending, or more generally crowdfunding platforms, Kiva requires borrowers to constitute a project profile where they are given the opportunity to describe the project with the aim of convincing lenders of their initiative. Next to the aforementioned features that may influence the lender’s attractiveness assessment of the borrower’s undertaking, the project description is a heretofore neglected but informative and crucial aspect. Contrary to other features, the project description grants a great extent of creative leeway for individual ideas, encouraging the entrepreneurial spirit of borrowers. Having said that, it is, however, often unclear from a borrower’s perspective what lenders expect to hear, how many information shall be disclosed, whether to adopt formal language based on facts and numbers, or an emotional approach and in which way business knowhow can or should be shown. Kiva offers lenders a unique diversity of choice in which thematic sector, geographic location and purpose to invest in. It seems reasonable to assume that the textual project description plays a decisive role and borrowers with an appealing description are given a leading edge, particularly when comparing similar thematic projects. This raises questions regarding the factors making a borrower’s project description more or less appealing to lenders.

We will therefore be guided by three main hypotheses in our analysis of linguistic features that affect the magnitude of a project’s success or failure. Our first hypothesis deals with the length of the description. It is assumed that the more detailed the description is, the more likely the project is to be funded quickly when successful, or conversely, the more likely the project is to have a smaller funding gap when unsuccessful. This notion is captured and measured through the length of the description, thereby equating description length with the level of detail provided. The second hypothesis predicts a connection between the level of emotion involved in a description and the funding speed as well as gap. It is hypothesised that a more emotional description as opposed to an emotionless one is related to a quicker funding success for successful loans or, conversely, smaller funding gap for unsuccessful loans. This assumption is tested by means of a sentiment score and a sentiment magnitude of each description. The third and last hypothesis proposes a connection between certain topics covered in a description, such as family, health and business-related information, and an increased likelihood of being funded fast when successful, or ending up with a smaller funding gap for unsuccessful projects, respectively.

The hypotheses are explained in more detail in section 4.4.

We build predictive models to capture these hypotheses, allowing us to identify underlying linguistic differentiators of successful, quickly funded loans as well as unsuccessful loans with a large funding gap. Within the scope of this analysis we are also interested to understand how the dynamics have changed over time on the microfinance platform. Is it possible to predict funding speed and gap of current projects equally well when using data from the early years of Kiva to train our model? In other words, is the behavior of Kiva’s first users and early adopters representative of Kiva’s current users? Similarly, we are interested in investigating the lenders’ behavior by examining the speed our model learns at. Do Kiva’s lenders exhibit stable behaviour, enabling us to retrieve an equally well model fit with a significantly smaller portion of training data? In other words, is it possible to predict lenders’ behaviour to the same extend when training our model on significantly less data? The findings may allow us to draw conclusions about

**4 Data Set**

**4.1 Data Source**

All data used in this project was obtained from the Kiva API. To avoid making thousands of requests to the API, archived data was downloaded on the 7th of June, 2018. The website is updated nightly and the downloadable data snapshots consist of three data files in csv format: ‘lenders.csv’, ‘loans.csv’ and ‘loans\_lenders.csv’.

1. Lenders

This data set provides a detailed overview of all registered lenders on Kiva. Information such as the lender’s location, occupation and lending motive are optional whereas the lender’s username, the date on which the lender became a member as well as the number of loans the user has contributed to are mandatory or collected automatically by the system and therefore exhaustive. This data set contains 2,349,176 entries.

1. Loans

This data set contains information about all projects registered on the platform. It details in total 35 characteristics about funded, expired and currently fundraising projects. Specifics include the loan amount asked for, the project description, any payments made to the loan so far, the list of lenders on the loan and information about the borrower. The earliest project posted in this data set was on the (TODO: ) XXX, whereas the last one was on the (TODO: )XXX. An exemplary observation can be seen in Table (TODO: ) XXX. In total, this data set consists of 1,419,607 entries.

TODO: Table XXX

1. Loans-Lenders

The third data set contains one list of lenders that are financially supporting a project for each funded loan on Kiva. The lenders are made available as arrays of distinct usernames which is linked to the project through its loan ID. This data set encompasses 1,387,434 rows.

**4.2 Data Sets Chosen**

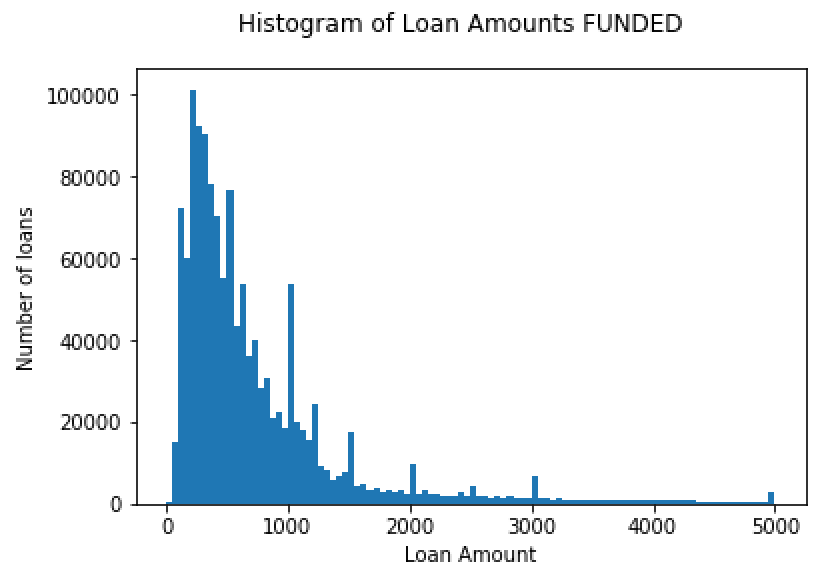
The chosen data set and sole source for this project was the ‘loans’ data set. By providing detailed information about the borrower and the project itself, it is the most suitable to analyse the research question and test the hypotheses established throughout the course of this project. It contains all loan characteristics necessary for linguistic analysis as well as information required to draw conclusions about financial and temporal aspects of the loans, such as the time it took for a successful loan to get funded (i.e. funding time) and by how much an unsuccessful loan missed its target amount (i.e. funding gap).

In order to control for already observed influences on funding speed and gap, the data set was further limited. According to **Bourdeau de Fontenay (2018),** loan amount, the borrower’s gender, the thematic sector of the project as well as the region the borrower is located in significantly affect the funding duration of a project. A female borrower looking to start a manufacturing project in Europe by raising a small amount such as $300 is very likely to receive funding faster than the average project does. Thus, it was decided to use only data of projects in the agricultural sector, from borrowers that are exclusively male and located outside of Europe to investigate funding speed. When choosing the data set, two things were paid special attention to: All restrictions needed to result in a sufficiently large data set with a reasonably balanced distribution of funding duration (Figure XXX & XXX - here or below?).

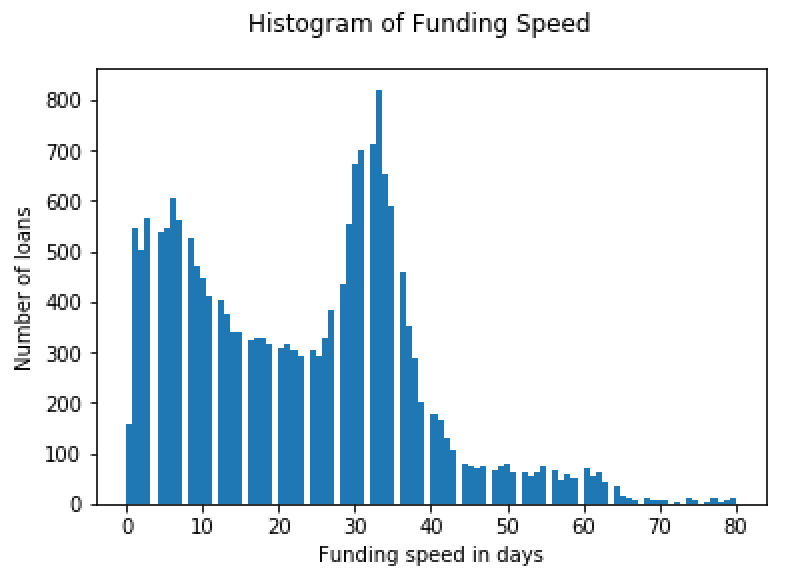
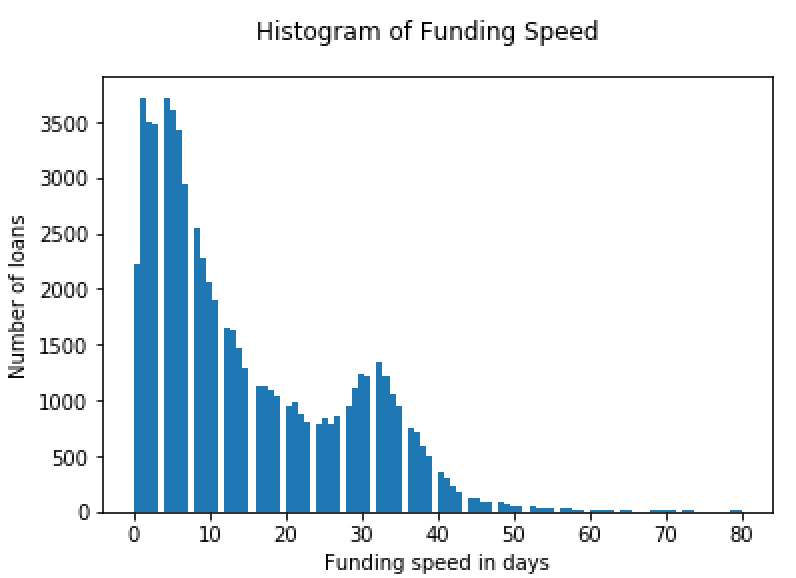
To investigate whether language has a greater impact on larger loans than on smaller ones **(first hypothesis)**, the data set established to analyse funding speed was further divided into data with a loan amount below $1,000 and above $1,000. Figure XXX shows that the distribution of loan amounts is at an inflection point at around $1,000 after which it flattens quickly, leaving comparatively few funded projects with a high loan amount. Thus, the two established data sets to analyse the impact of linguistic features on funding speed are the following:

Data set A: < 1000 & Agriculture & excl. male & not Europe

Data set B: > 1000 & Agriculture & excl. male & not Europe



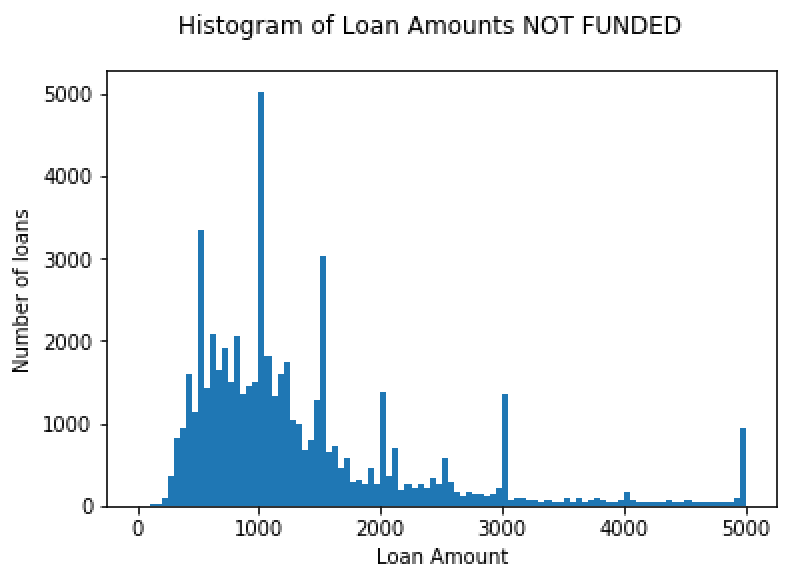
The distribution of funding speed for both sets can be seen in Figure XXX and XXX below. Data set A contains 66,411 entries, while B consists of 19,587 data rows. These numbers are after the data cleaning process described in the section.



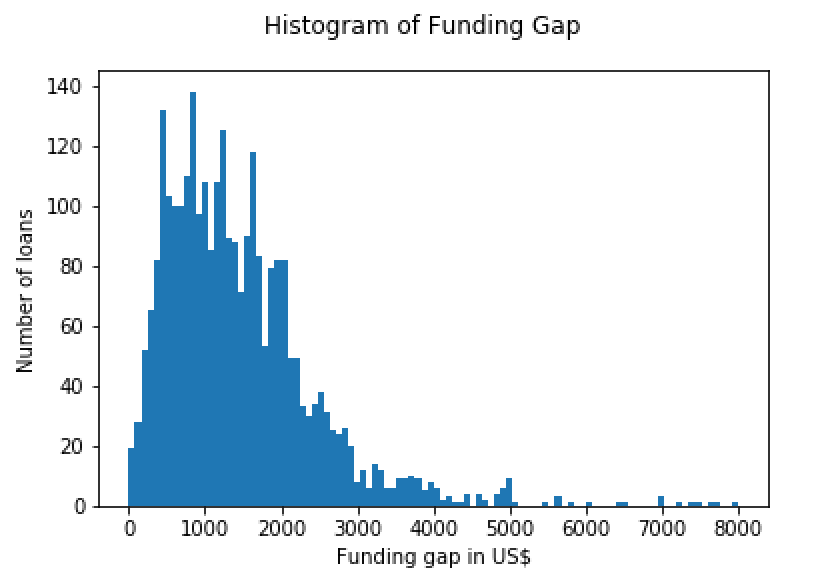
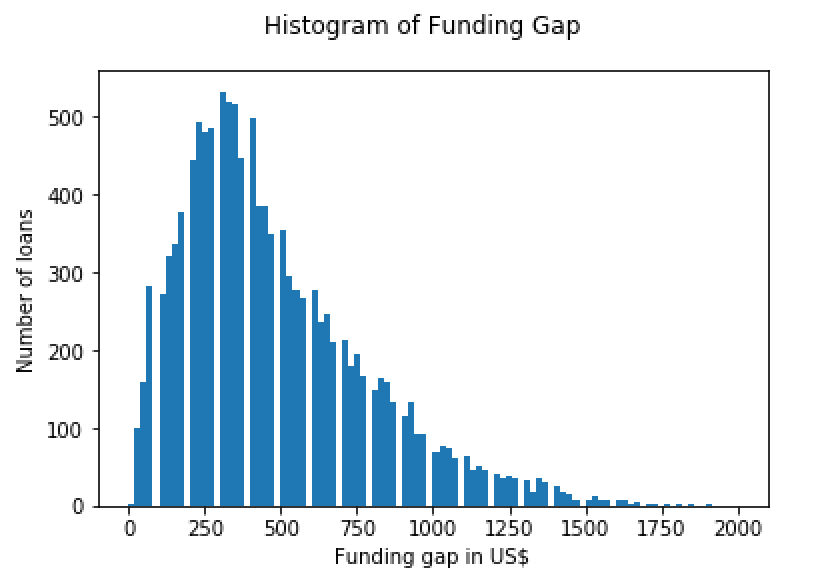
When analysing funding gap, the amount by which an unsuccessful loan missed its target amount, **Bourdeau de Fontenay’s** findings suggest that only loan amount and thematic sector play a large enough role to be considered significant. With the size of the data set and a reasonably smooth distribution in mind, the agricultural sector as the category with the largest number of projects, is chosen. To investigate further relationships between different loan amounts and linguistic features, the resulting data set is split further into projects with loan amounts below $1,700 and above this point. A cutoff point of $1,700 is chosen due to its location at an inflection point after which the amount of loans of expired projects drops significantly (Figure XXX). Hence, the following two data sets have been formed to explore the funding gap:

Data set C: < 1700 & Agriculture

Data set D: > 1700 & Agriculture



The distribution of funding speed for both sets can be seen in Figure XXX and XXX below. Data set A contains 11,472 entries, while B consists of 3,642 data rows. These numbers are after the data cleaning process described in the section.



**4.3 Data Cleaning**

All data used in this project was imported directly into a SQLite database. Thereafter, inconsistencies that stood out during exploratory analysis were removed. First, all project duplicates, in total 22 entries, were eliminated. Second, all projects for which the date of raising the whole amount demanded lies before the date when the project was created and uploaded by the borrower were deleted, amounting to 204 occurrences. Third, all projects without a description text were removed as the major part of the analysis is dependent on it. Descriptions below 10 words would also be impede a proper analysis and were deleted. Similarly, in the limited scope of this project it was not possible to evaluate descriptions written in another language than English and with missing translation correctly. In total, 37203 projects were affected through an unsatisfactory description. Out of the 1,419,607 data entries in the “loans” data set were thus 1,382,178 entries left.

**4.4 Data Engineering**

Prior to fitting models to our data, the relevant metrics required to analyse the aforementioned hypotheses need to be retrieved and/ or calculated. The explanatory variables all constitute continuous variables and fall in one of three categories: description length, sentiment analysis, LIWC topics count. The two dependent variables we seek to predict were computed as continuous and ordinal variables, respectively.

**4.4.1 Explanatory variables**

1. Description length

Our first hypothesis proposes a mutually reinforcing connection between a successful project’s description length and its funding speed. The rationale behind this hypothesis essentially consists of the association between a longer description with a deeper level of detail. It seems reasonable to infer that a longer description offers the possibility for the borrower to explain the project to a greater extent, providing more information, for example, about the desired impact, the plan how to reach his target, himself and his daily life and family. On the one hand, possible lenders are able to identify and empathise more easily with the borrower while on the other hand, the process of understanding the rationale behind his project is facilitated through more information. As XYZ [TODO: REFERENCE] discovered and explored to a great extent, being able to identify oneself with another person vastly improves the chances to receive support. This supposition is captured through the length of the project description, i.e. the count of the words. Conversely, for an unsuccessful project, it is hypothesised that a project’s description length is negatively associated with a project’s funding gap. More specifically, this means that a longer project description only shortly misses the loan amount it asked for.

b) Sentiment Score and Magnitude

Our second hypothesis is concerned with the emotional tone within a project’s description. It is theorised that a description that conveys a large amount of emotion, whether positive or negative, is positively associated with funding speed and negatively with funding gap. Two possible reasons worth considering are the following. Borrowers filling their description with more positive, emotionally charged words often convey a feeling of hope and enthusiasm. According to XYZ [TODO: REFERENCE] feeling hopeful inspires support blablabla (TODO: or similar quote), making it more likely that descriptions filled with positive emotions convince other individuals to support them (TODO: bad sentence - rephrase). A second reason might be that descriptions that are perceived rather as negative often arouse more sympathy and compassion from lenders, leading to faster funding. An example would be a spinach farmer telling about an insect infestation owed to the missing money to buy insect protection that leaves him with only 1/10 of his usual harvest. As this is by far not enough to support his family, he is trying to avoid this for his next harvest and asks for a loan to buy insect pesticides. It is reasonable to assume that lending individuals sympathise with the farmer and his starving family. [TODO: look for REFERENCE/ article “why do ppl from developed world micro-lend to ppl in developing world?” → sympathy]

In order to measure the effects of sentiment on the two continuous variables to predict, two scores were retrieved from the Google Cloud Natural Language API and used as independent variables. **The choice of the Google Cloud API as sentiment analyser is explained in more detail in section 6.1.** First, a sentiment score was calculated, indicating the overall emotion of a description. It ranges between -1.0 (negative) and 1.0 (positive) and consists of the average of the individual sentences’ sentiment scores in a description. Second, a value for the sentiment magnitude is retrieved. It specifies the emotional content present within the description and is comprised of the absolute sum of the individual sentences’ magnitude values in a description. While a sentiment score of zero could stand either for an unemotional description or one involving both negative and positive emotion that cancel each other out, the sentiment magnitude provides greater clarity, thereby preventing ambiguity.

Sentiment score is calculated as the average of the individual sentences’ sentiment scores. A description has an average number of eight sentences with the majority being rather neutral with regard to the emotional level. This often high amount of neutral sentences in a description leads to a sentiment score balanced around 0 for the whole description (TODO: Figure xyz). Intuitively, one would however assume that despite many neutral sentences, the emotions conveyed by a few sentences stand out and are formative to the entire description. Therefore, in order to put more emphasis on the sentences conveying emotions, three alternative calculations of the sentiment score were tested:

First, instead of using the average, the median of all sentences’ scores was taken to calculate the overall description’s sentiment score with the aim of being a better representation of the individual scores. A second method to calculate the sentiment score is to exclude all neutral scores from the average calculation. This stresses the emotional scores to the largest extent possible. The third alternative used assumed that the impressions conveyed at the beginning of the text as well as at the end of the text are the deepest and should therefore be stressed more. The description’s sentiment score is calculated by assigning different weights based on the position of the sentence in the description. Sentences in the first and last quarter of the description are assigned double the weight of the sentences in the middle. [TODO: Is there a reference for it somewhere?] These three alternatives to the basic sentiment score of a description provides more granularity to the analysis.

c) LIWC

The last independent variables are retrieved using the computerised Linguistic Inquiry and Word Count (LIWC) text analyser. It counts words in psychologically meaningful categories [TODO: Reference- <http://journals.sagepub.com/doi/abs/10.1177/0261927X09351676>? journalCode=jlsa] and is used throughout our analysis to provide information on the extent certain thematic areas are represented in descriptions. The third hypothesis refers to these themes and assumes that descriptions proving a large amount of attentional focus to one or more of the six chosen categories are likelier to be funded fast or, conversely, to have a smaller funding gap when unsuccessful. The word counts per category are normalised with the total description length.

Categories chosen are family, health, work, numbers, pronouns and insights. In our analysis, the category “family” is strengthened by adding the category “humans” to it as most words belonging to the latter category such as children would intuitively have been assigned to the “family” category. Therefore, both categories are merged into one with the title “family”. The idea behind choosing this category is the sub-hypothesis that borrowers involving their families are often perceived as more likeable and sympathetic persons leading to faster funding for successful and smaller funding gaps for unsuccessful projects [TODO: REFERENCE]. The category “health” was chosen under the sub-hypothesis that borrowers mentioning a sick child or a disability will trigger sympathy and speed up the funding process. “Work” is chosen as category as the usage of many work-related words creates the impression of a hard-working and responsible borrower. It is amplified by adding the category “achieve” to it, symbolising [TODO: change word] ambitious borrowers. Together the category might also prove a business-oriented mind and a focus on achieving one’s goals, thereby returning the money. In an extreme case, it could signal to a lender that the borrower likely to return the invested money.

The fourth category “numbers” is also composed of two LIWC classes: “number” and “quant”. This category was included under the assumption that return focussed lenders are likelier to invest faster when the textual description is underlined by quantitative measures. Furthermore, the use of pronouns in a description was analysed by including the category “pronouns”. It is reasonable to assume that a larger amount of pronouns used usually equates to a description written in a rather personal tone, thereby making it easier for the lender to emphasise with the borrower than when writing in a neutral manner. The last category to be investigated is “insight”. Through words such as knowing, acknowledge, logic, understand, prove, rational and reasonable it conveys the impression of a thorough, reflected argumentation and understanding of the task at hand when used in a description [TODO: rephrase/ better argument]. The following table includes the LIWC category used and demonstrates exemplary words for each [TODO: add table name].

[TODO: add table with each LIWC category + words mentioned]

[TODO: add table summarising independent variables]

**4.4.1 Dependent variables**

Successful and unsuccessful loans are analysed using two different metrics. In the data set retrieved from the Kiva API projects may have one of four different **statuses**. The table below **aufweisen** and helps to differentiate between them by providing a short explanation for each. Projects with the status “funded” and “refunded” have been grouped into successful loans where projects with the status “expired” **ausmachen** the unsuccessful loans. Projects that are currently fundraising have been **ausgelassen** from our analysis.

Funding duration, the time it takes for the project to reach the targeted amount, is used as **predictive** variable for successful loans. It is calculated by deducting the time the project was posted from the time the targeted loan amount was reached:

Raised Time - Posted Time = Funding Duration

We extract the number of days from the resulting timestamp and obtain an ordinal variable measuring the time each project is fundraising in days.

Unsuccessful loans are examined through the amount by which the project misses its target, the so-called funding gap. It is calculated by subtracting the amount the project raised till its expiration date from the targeted loan amount. The result is a continuous, **predictive** variable measuring the funding gap for each observation that ranges between X and X (TODO: see figure XYZ).

Loan Amount - Raised Amount = Funding Gap

The table below summarises the dependent variables used in our analysis and their types.

[TODO: add table]

**5 Exploratory Analysis**

In this section, we explore the hypotheses described in section 4.4 in more detail. Visualisations of the data through graphs and figures help us understand the data sets, its particularities and differences to a greater extent. Specifically, we will get a first feeling whether the hypotheses our intuition led us to are leading into the right direction in the search for a relationship between linguistic features of the project’s description and funding speed as well as funding gap. We start by examining the funding speed and the independent variables for successful loans. Next, we do the same for unsuccessful loans where we inspect funding gap next to the independent variables. For both categories we seek to further analyse the difference between small and large loans.

**5.1 Successful Loans**

1. Funding Speed

The variable to predict when examining linguistic factors of successful loans is funding speed, i.e. the time it takes for a project to be funded with the amount asked for. All our hypotheses for successful loans deal with the relationship between some linguistic features and funding speed. Before delving into the independent variables, however, we will first take a closer look at the dependent, ordinal variable itself in order to gain a better understanding of how long successful projects need to reach their targeted amount. Data sets A and B, established in section 4.2, represent the limited data sets for small and large successful loans, respectively, we use throughout our analysis. Figure XXX (a) shows the distribution of loans with a requested amount below $1,000 (“small loans”) while (b) represents the loans with a loan amount above $1,000 (“large loans”).

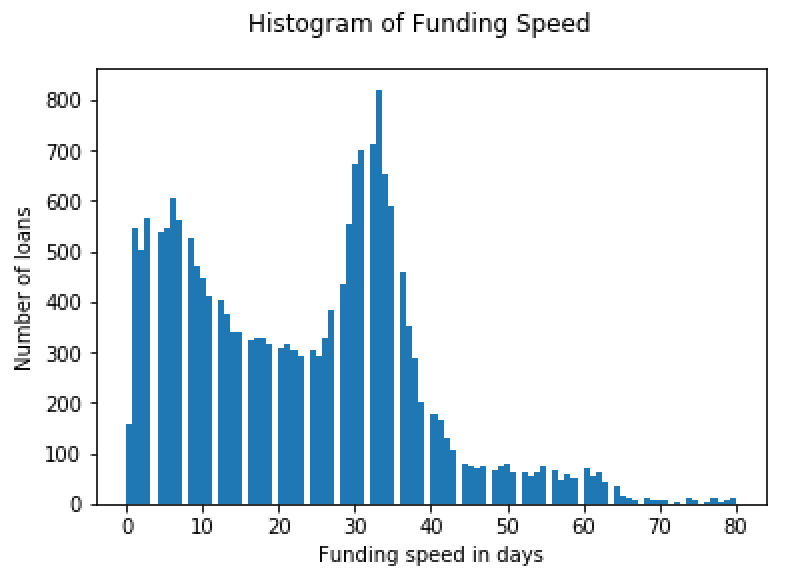
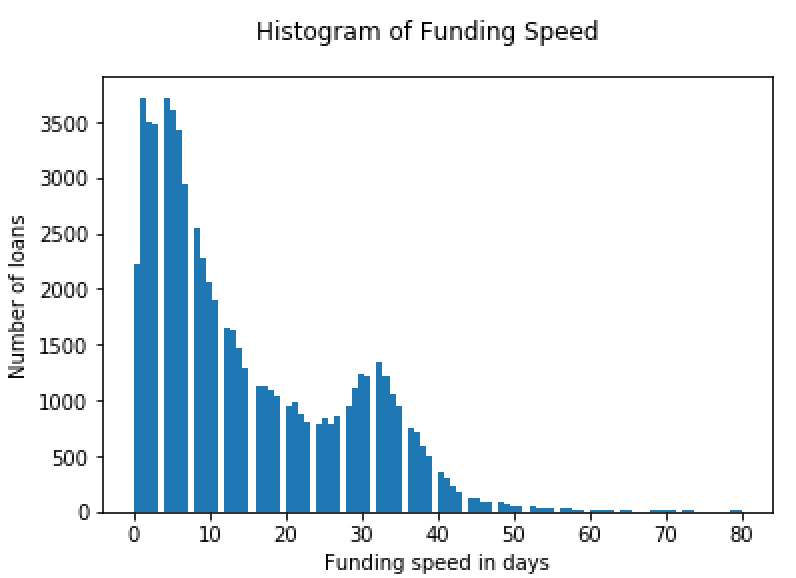


Figure XXX: (a) …. (b) ….

As is clearly depicted, the majority of small loans are funded very fast, usually within 1 to 6 days. From day 7 onwards, the distribution decreases steadily until day 28. We notice a last bump in the data between the 28th and the 35th day which is likely to be caused by a website mechanism that pushes projects back to the top of the page after some time. This bump is also clearly visible and even more pronounced for large loans. Despite the same general tendency of being funded rather early as for small loans, it even seems as if more loans are funded within this short period between day 28 and 35 than in any other time period.

b) Description Length

This leads us to describe the independent variables for successful loans, starting with description length. In order to gain a better overview of how long the project descriptions to be analysed are, figure XXX (a) and (b) depict the distribution of description length for small and large loans, respectively. Small loans have a minimum of 15-word-descriptions and a maximum of 1140 words whereas large loans have a minimum of 20 and maximum of 1124 words. Both distributions clearly show that the majority of loans have descriptions of a length between 60 and 200 words with a peak around 100 words. Moreover, large loans have another comparatively small bump around 320 to 400 words. It seems reasonable to expect larger loans to have a longer, more detailed project description as the more money a borrower asks for the more convincing the project profile should ideally be, often in form of a more extensive description. Nevertheless, the distribution of large loans does not seem to be significantly more skewed towards a higher word count per descriptions as we intuitively would have anticipated.

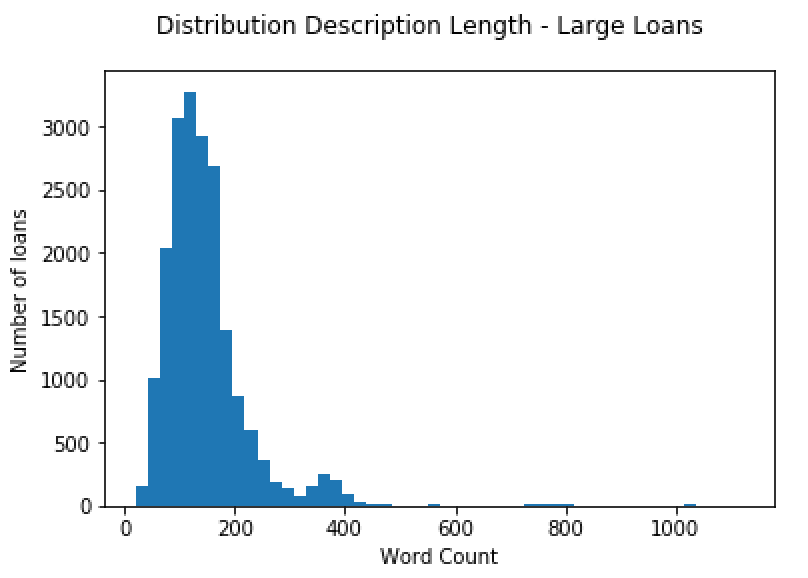
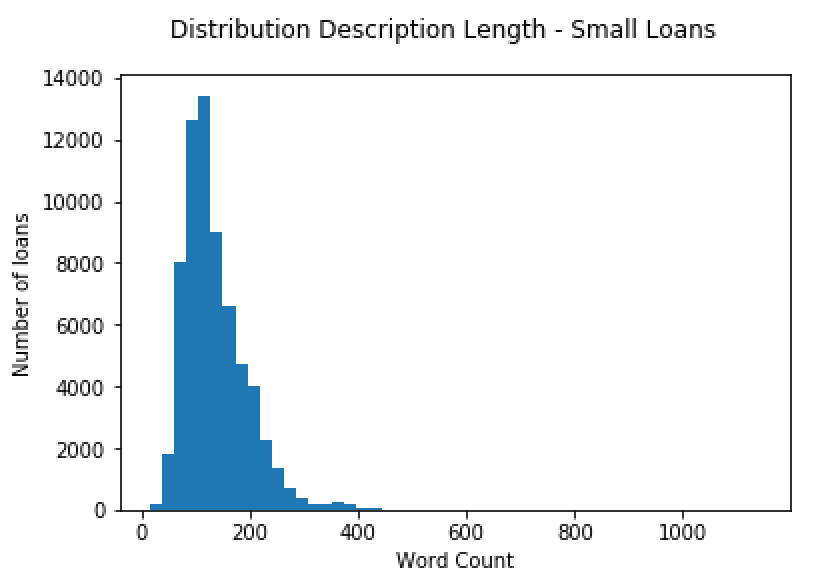


Figure XXX: (a) …. (b) ….

The first hypothesis established proposes a positive relationship between the length of a project’s description and its funding speed. A scatter plot shown in figure XXX depicts the correlation between these two variables. The red line presents a least squares fitted line and depicts a general downward trend, meaning a description with a high amount of words is likely to be funded faster than one with a short description. Judging by this first examination, this result supports our first hypothesis.

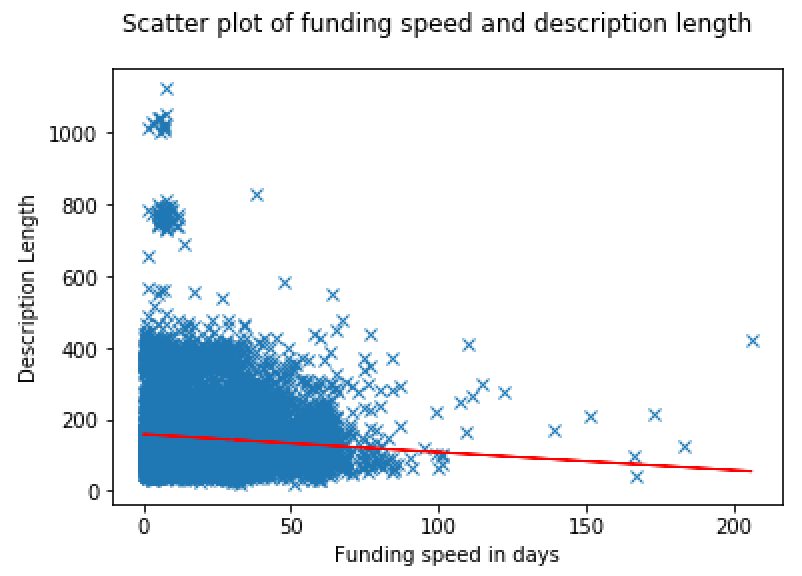
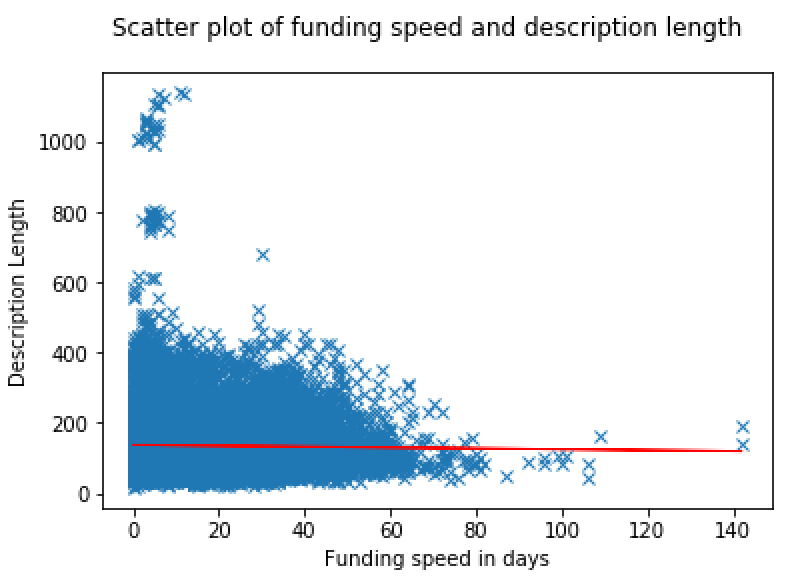


Figure XXX: ...

c) Sentiment Score & Magnitude

The second hypothesis deals with the sentiment involved within a project description. In order to gain a better understanding of the sentiment score for each observation, figure XXX (a) and (b) show the distribution for both small and large loans, respectively. -0.7 and -0.5 constitute the minimum sentiment scores for small and large loans. The maximum value for both kind of projects is 0.9 with the majority of loans representing sentiment scores between 0 and 0.4. Both distributions are also visible skewed towards the right, i.e. towards positive emotions, with over 90% of all observations having scores above or equal to 0. This suggests that loans have a much greater tendency to be filled with positive rather than negative emotions. The high maximum values as opposed to the leveled minimum further support this deduction.

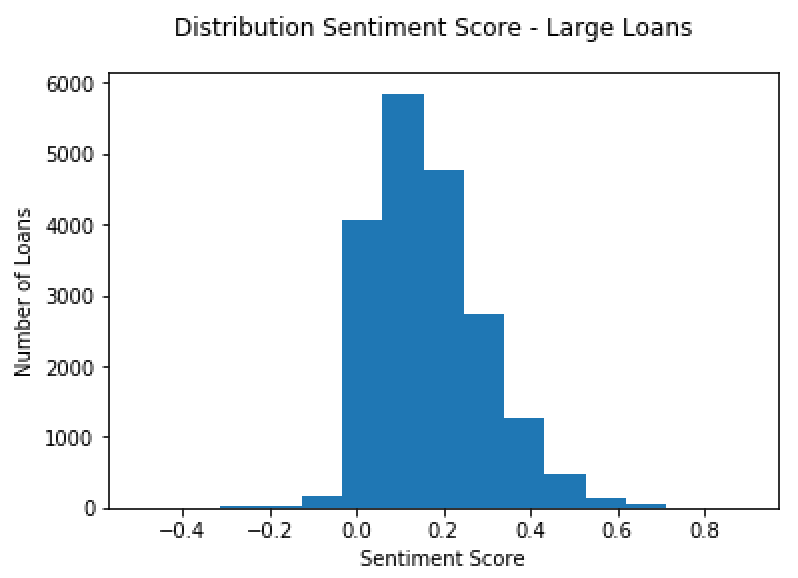
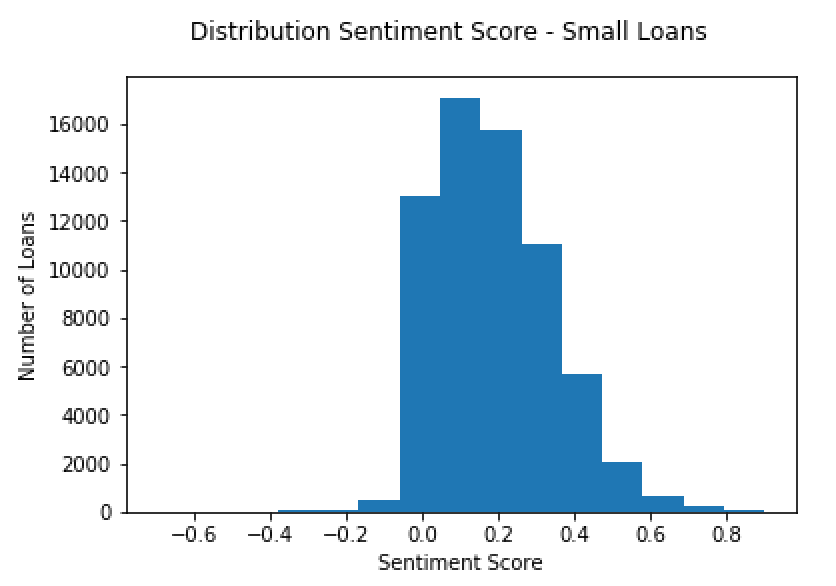


Figure XXX: (a) …. (b) ….

In order to depict the relationship of the sentiment score and funding speed for small and for large loans, two scatter plots are shown in figure XXX (a) and (b). It seems reasonable to assume a non-linear relationship as suggested in our second hypothesis which states that any emotion involved - whether positive or negative - is related to a higher funding speed. In accordance with our expectations, both scatter plots show a curved relationship, there supporting the hypothesis. Another outstanding feature is that funding speed decreases significantly with more negative emotions in a description. This is the case for both small and large loans, is however particularly intense for smaller loans.

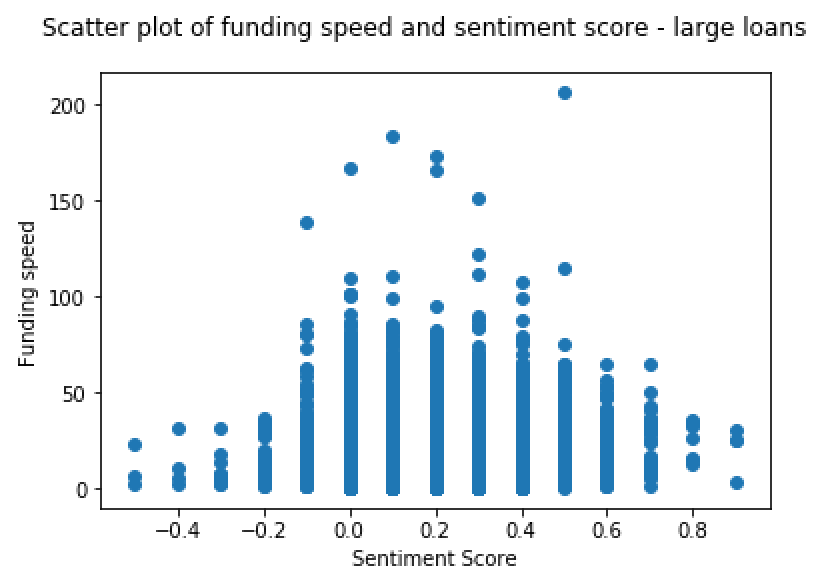
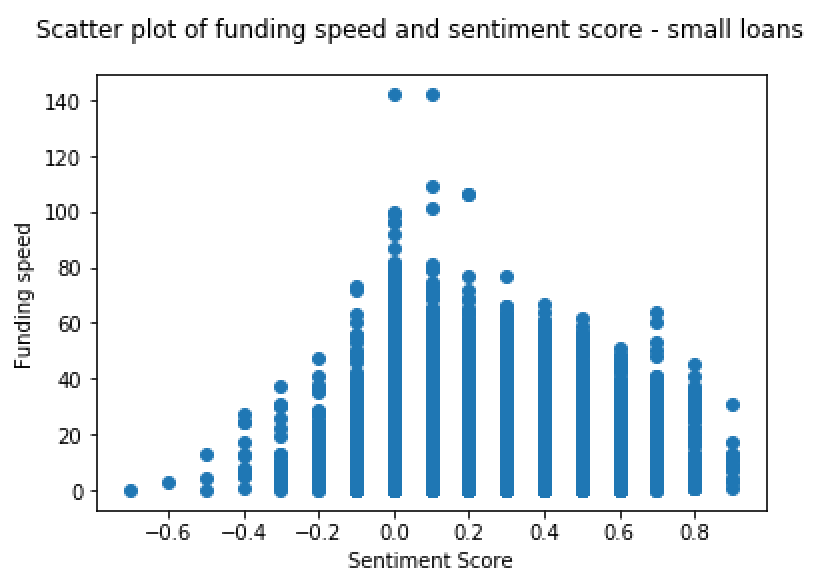


Figure XXX: (a) …. (b) ….

Another attribute classifying the emotional tone of the descriptions is sentiment magnitude which indicates the overall strength of emotion within a given text. Similar to the previous variable, we plotted the distributions for small and large loans in figure XXX (a) and (b), respectively. Unlike the sentiment score, magnitude is not normalised and each expression within the description contributes to its magnitude. Both achieve a similar maximum value of around 18.0 and have a minimum of 0.0. The figures clearly show a tendency towards low magnitude values, with the majority having a score between 0 and 3. This means that most projects have a description with a relatively few to zero emotions, suggesting that those address lenders with a rational and objective tone.

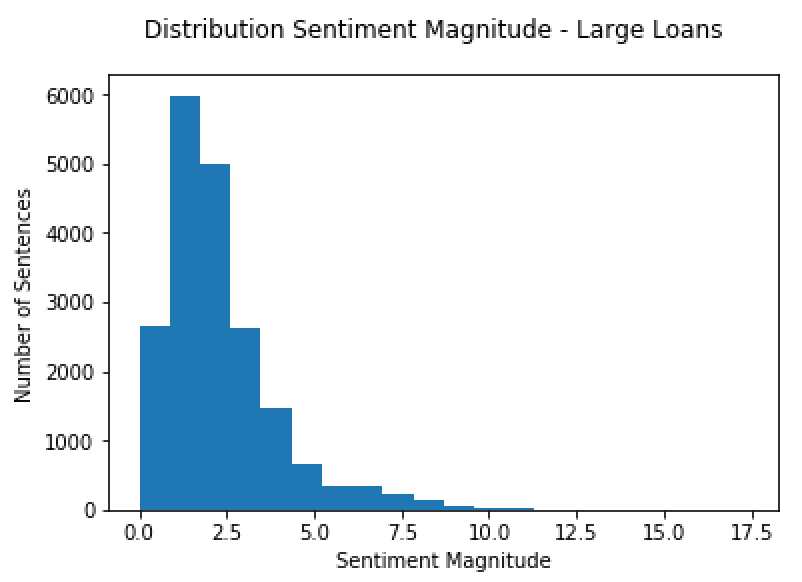
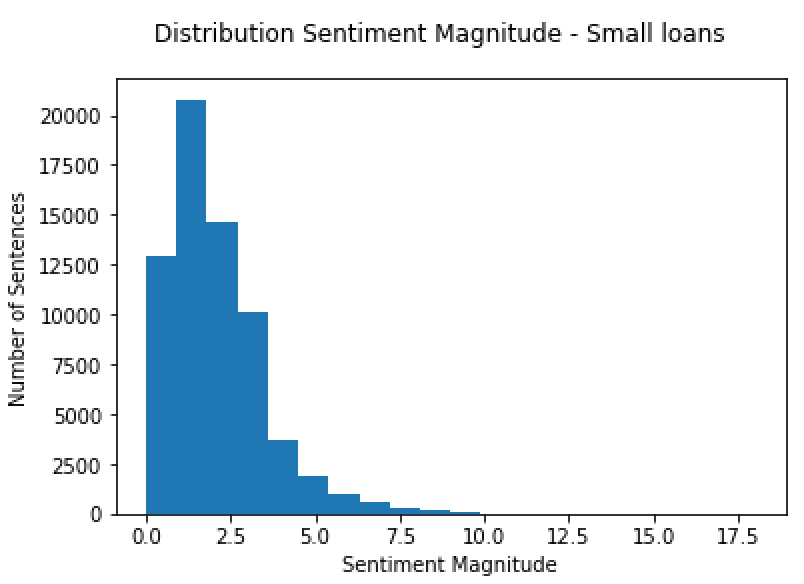


Figure XXX: (a) …. (b) ….

According to the second hypothesis established a higher amount of emotional content relates to a short funding duration. In order to gain a better understanding of this relation, figure XXX (a) and (b) show scatter plots between these two variables. Judging from a first examination, the relationship shown in the plots is linear and supports our initial hypothesis. The higher the sentiment magnitude within a description, the shorter the time it takes to receive full funding for the project.

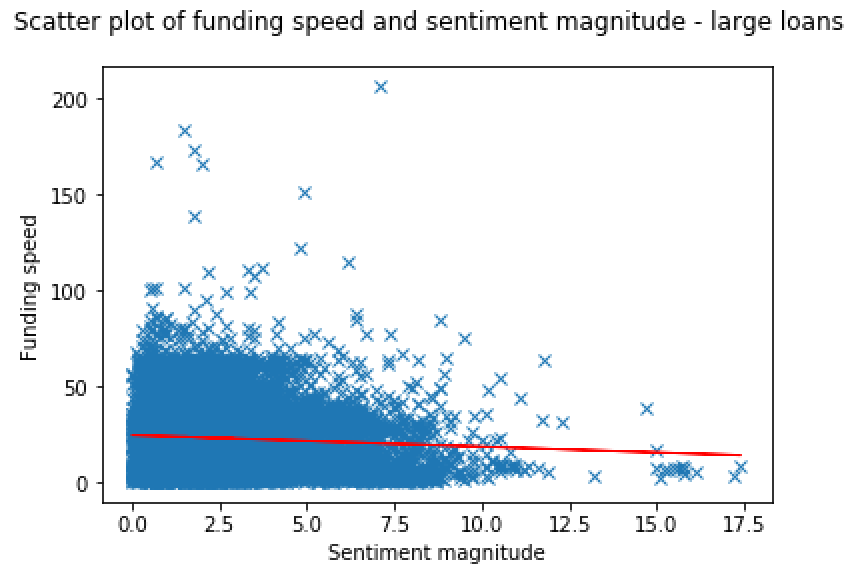
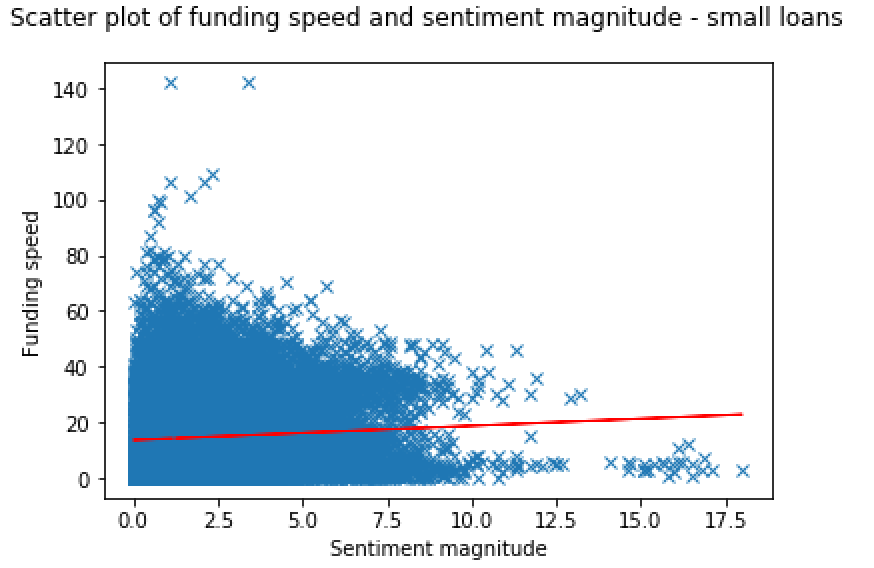


Figure XXX: (a) …. (b) ….

d) LIWC

The next linguistic features under examination are the selectively chosen LIWC categories: family, health, work, numbers, pronouns and insights. We are interested in evaluating the frequency of occurrence within the average description with the aim of gaining a deeper understanding of how these categories compare **in relative terms** with regard to the borrowers’ attentional focus. Besides the explanation of his undertaking, is the borrower talking mostly about his family, his health or rather stating his work proposal in detail? And how do these occurrences relate to the duration of fundraising? As the data set we are using for the analysis is limited to data from the agricultural sector, health-related information within the description cannot be confused with project-related information as would be the case for projects in the health sector. The figure below shows two pie charts for small and large loans, respectively.

The following figure further describes the distribution of these categories. As the relative averages shown in the pie chart already suggest, the shape of the distributions point out that work-related information is the category mentioned the most by far and thus where the attentional focus lies the most. This suggests that while borrowers describe their undertaking thoroughly, they usually do not delve deeply into other, often more personal topics. The amount of work-related information, however, also varies a lot as it ranges from 0 to 30 which indicates that some borrowers invest more time into communicating their business vision than others, thereby often proving a business-oriented mind.

As the significant correlation between the LIWC category “work” and funding speed depicted in figure XXX proposes, borrowers have a great chance to reduce their funding duration through this.

In summary, ….

**5.2 Unsuccessful Loans**

1. Funding Gap

After examining the independent and dependent variables we will use in our analysis for successful loans, we will now explore them for loans that have expired before being able to raise the targeted amount. The variable to predict here is funding gap, i.e. the amount by which the project missed its target. Again, we start by analysing the dependent variable first, before delving into the independent ones. Data sets C and D, established in section 4.2, represent the limited data sets for small and large unsuccessful loans, respectively. Figure XXX (a) shows the distribution of expired loans with a requested amount below $1,700 (“small loans”) while (b) represents the loans with a loan amount above $1,700 (“large loans”).

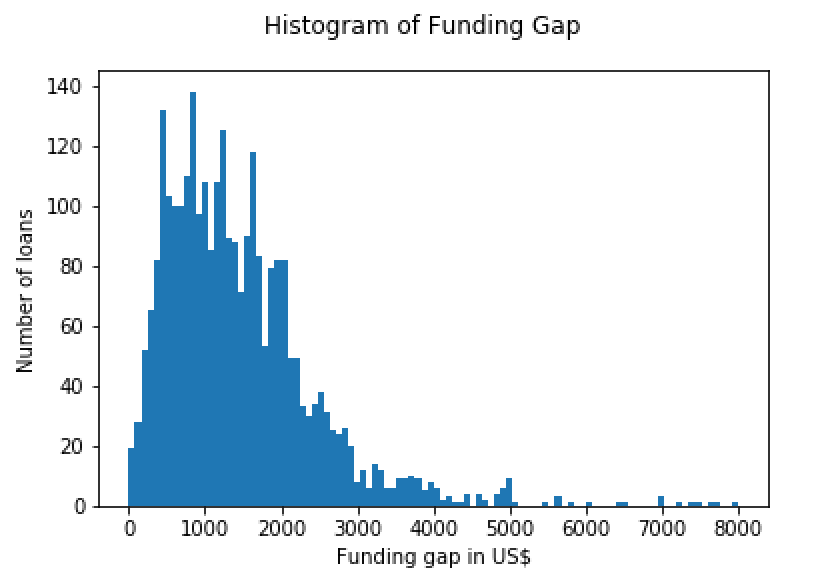
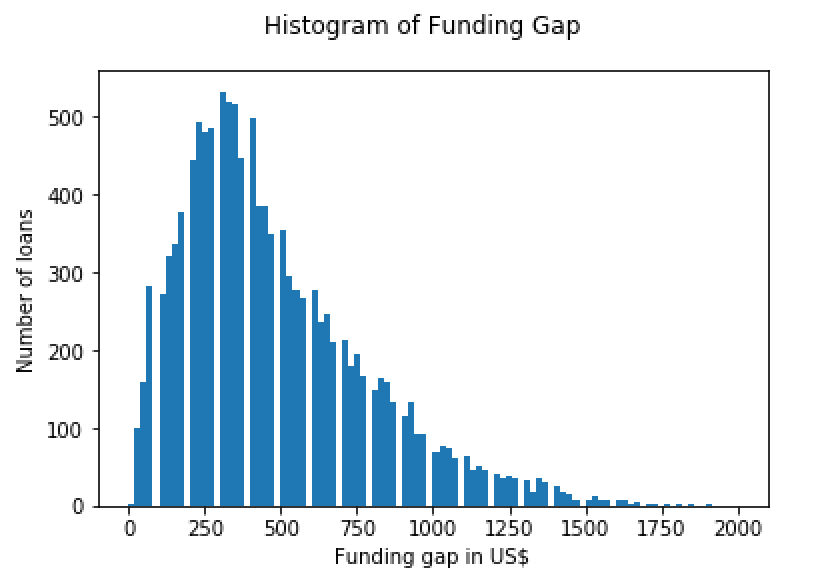


Figure XXX: (a) …. (b) ….

Through the high amount of funded projects, one would intuitively expect the majority of unsuccessful projects to be expired rather close to the targeted amount. As can be seen through the distributions, it indeed seems as if most of the small and large loans are left with comparatively small gaps with the average being XXX for small and XXX for large loans. This difference can be explained through the much larger loans that are included within data set D. Especially the distribution for large loans, however, shows a long tail towards larger funding gaps with a maximum gap of XXX, suggesting an inevitable amount of projects that are far from being “almost funded”.

b) Description length

The first explanatory variable we will examine for unsuccessful loans is the length of a project’s description. To obtain a quick overview of how long the descriptions for expired loans are, figure XXX depict the distribution of description length. Small loans as shown in XXX (a), have a minimum word count of 17 and a maximum of 465, while larger loans start at 22 words and end at 537 words. Compared to the maximum of successful loans which centered around 1130 for both small and large loans, the description length is reduced by more than 50%. This striking fact might support our second hypothesis of a positive relation between description length and funding speed as the unsuccessful loans have not been funded fast enough to raise the targeted amount before their expiration date. Both distributions peak between 100 and 200 words, similar to successful loans. The fact that larger loans have a higher maximum, are a bit more skewed towards larger word counts and register a small bump between 400- and 500-word-descriptions suggest that borrowers asking for larger amounts have longer descriptions, possibly because feeling obliged to justify their project to a larger extent.

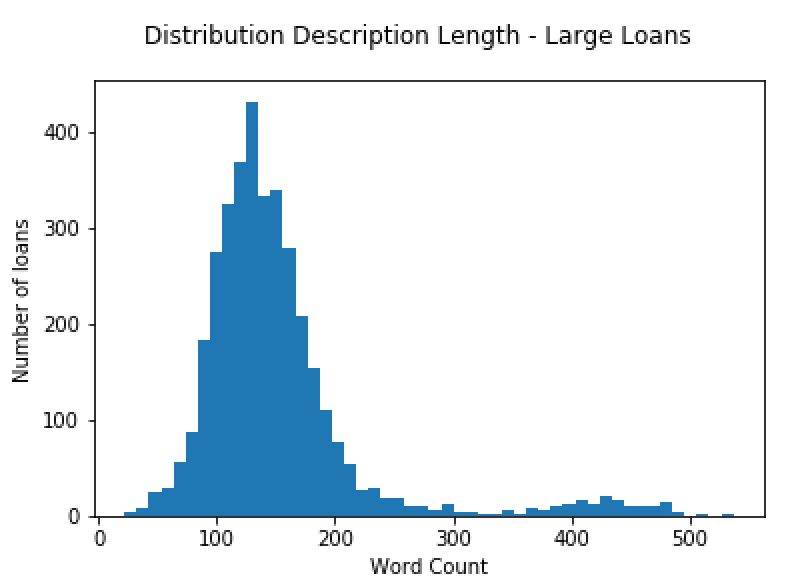
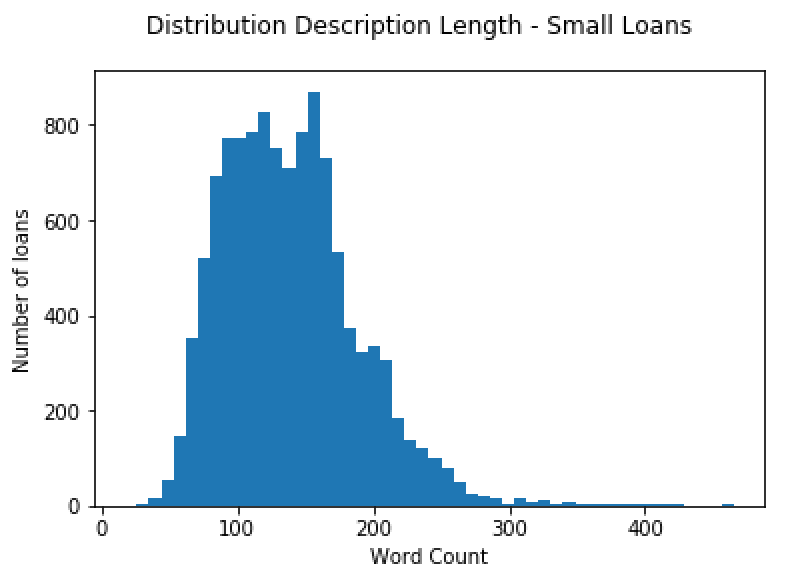


Figure XXX: (a) …. (b) ….

Looking at the scatter plots depicting the relationship between funding gap and description length in figure XXX, some surprising notions are revealed. While the relation for small loans seems to be at least to some extent as predicted by our hypothesis, the scatter plot for larger loans seems to react in the opposite way. This would suggest that a larger description length corresponds to a larger funding gap. A rather steep least squares fitted line reinforces this relationship. A possible explanation might be that lenders whose attention is not captured at the beginning of a description quickly lose their interest for projects with long descriptions.

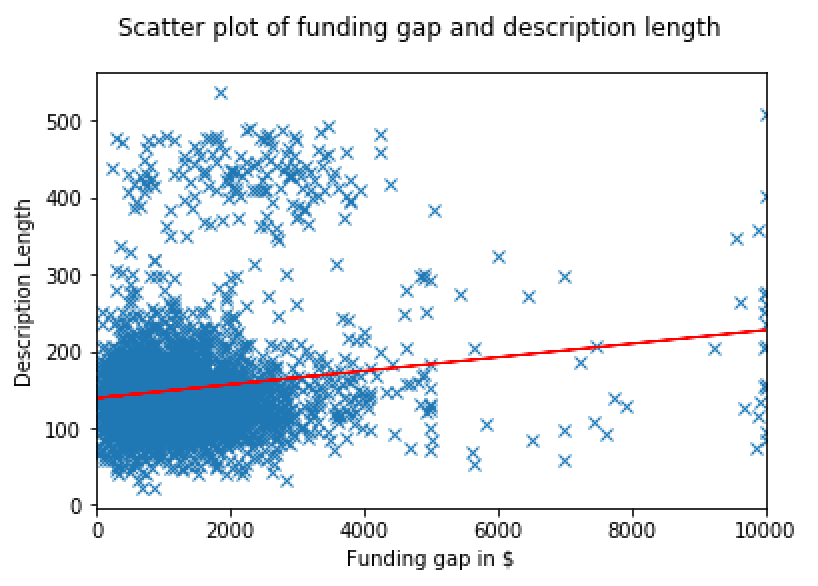
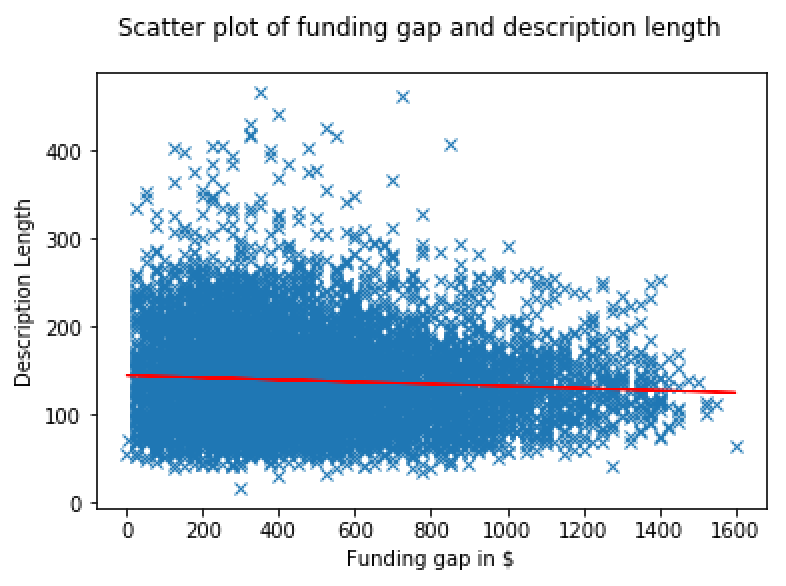


Figure XXX: (a) …. (b) ….

c) Sentiment Score & Magnitude

The second hypothesis deals with the level of emotion within a description and its negative relation to the funding gap of an unsuccessful project. As previously explained, two values are investigated to test this hypothesis: sentiment score and sentiment magnitude. The sentiment score for small and large expired projects is very similar distributed to the ones of successful projects shown in figure XXX (a) and (b), respectively. Their maxima are equally high values of 0.9 for small and 0.8 for large loans. An outstanding feature for both kind of loans is however the minimum value of -0.3 as opposed to -0.7 and -0.5 for funded loans. Again, it is noticeable that more than 90% of all observations have neutral or positive scores, suggesting that descriptions have a greater tendency to contain positive rather than negative emotions.

In order to gain some more insights on the relationship between funding gap and sentiment score, two scatter plots are shown in figure XXX. Similar as for successful loans, it seems reasonable to expect a non-linear relationship as according to our second hypothesis both negative and positive emotions are related to a smaller funding gap. In line with this assumption, both scatter plots depict slightly curved relationships.

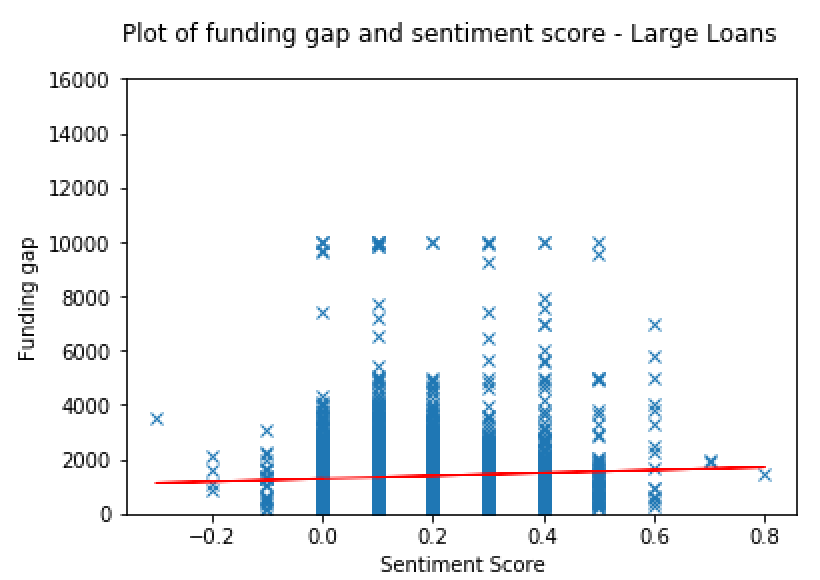
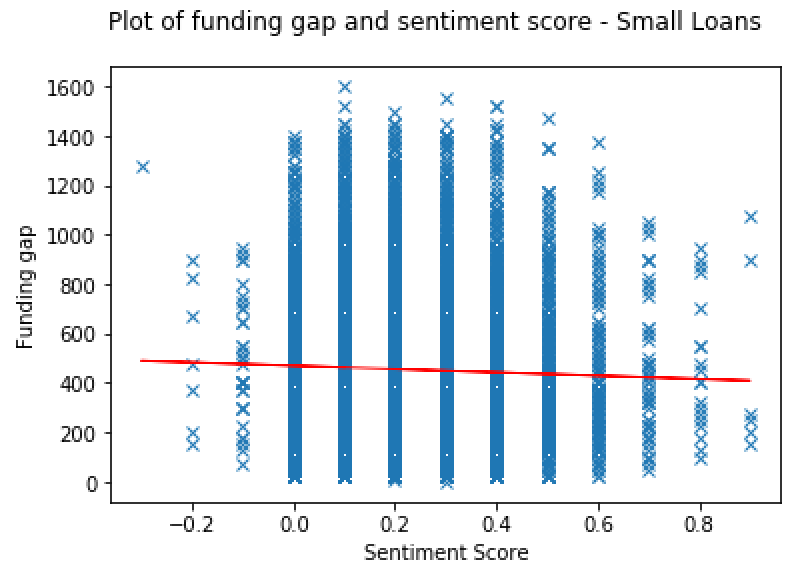


Figure XXX: (a) …. (b) ….

Similar to the sentiment score, the distribution of sentiment magnitude of unsuccessful loans does not differ much from the one for successful loans in figure XXX. The majority of loans fall within the area between the neutral score of 0 and 4 with a peak around a score of 2. Unlike successful loans with a maximum of 18.0, expired projects only reach a maximum magnitude value of 13.0 for small and 10.0 for large loans. A noteworthy insight is that we have seen lower maximum values for each of the independent variables so far compared to the ones for successful loans.

Figure XXX (a) and (b) illustrate the relationships between funding gap and sentiment magnitude for small and larger loans, respectively. Consistent with the second hypothesis that a higher sentiment magnitude relates to a smaller funding gap, scatter plot (a) demonstrates a downward relationship. Equally surprising as for description length, however, scatter plot (b) reveals a surprising finding for larger loans. Here, a higher sentiment magnitude relates to a larger funding gap, thus contradicting the hypothesis.

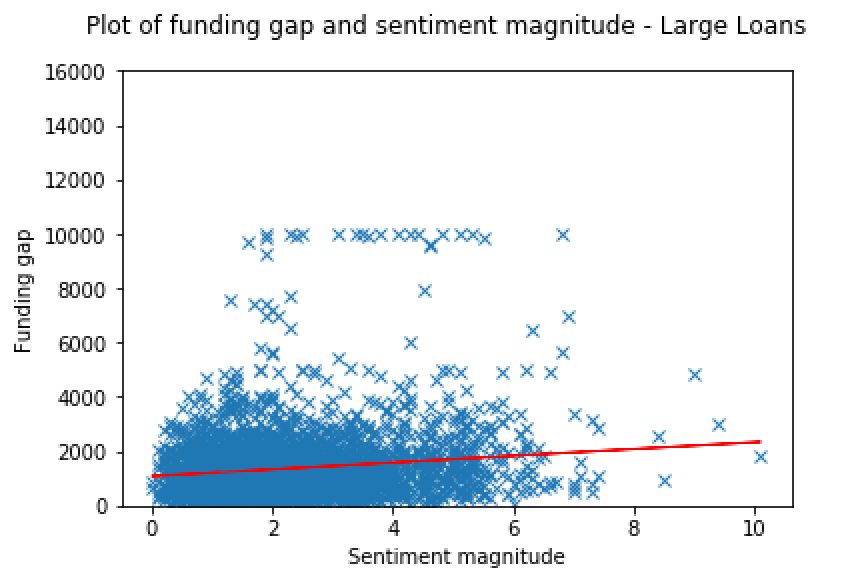
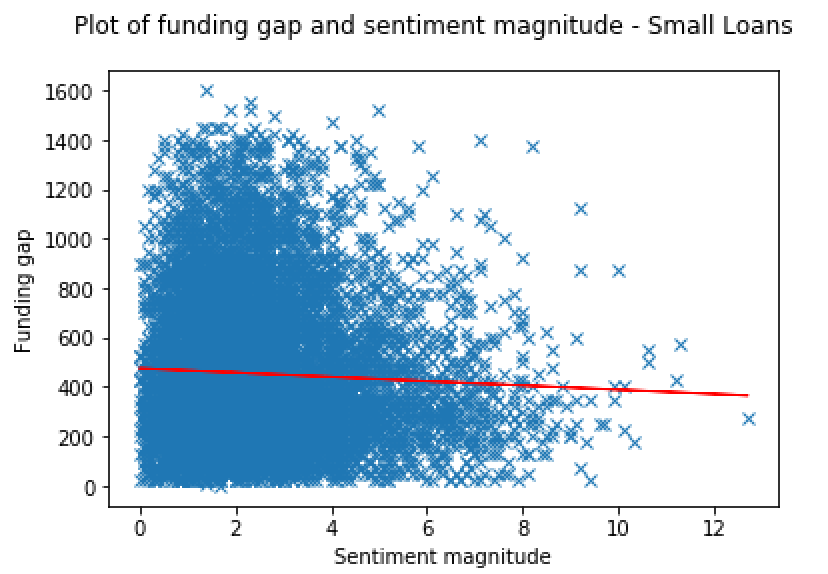


Figure XXX: (a) …. (b) ….

d) LIWC

In order to gain a broader awareness for the attentional focus in project descriptions for expired loans, figure XXX shows pie charts for the LIWC categories we selected for our analysis. As can be seen XYZ

**6 Modelling Linguistic Analysis**

**6.1 Methodology**

We seek to investigate the relationship of certain linguistic characteristics of a Kiva project and its magnitude of funding success or failure. Our first approach is to model the dependent variables funding speed and funding gap with linear regression models, using the ordinary least squares method (OLS). Due to a rather bad model fit, we continue with a second approach. This time we use a predictive model, the random forest algorithm. The table below summarises all predictor variables used in this analysis along with their types and ranges as they appear in out data sets.

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Type** | **Range** |
| 1. Description length | Integer | 10 to infinity |
| 2. Sentiment score | Float | -1 to 1 |
| 3. Sentiment magnitude | Float | 0 to infinity |
| 4. Family count | Integer | 0 to infinity |
| 5. Health count | Integer | 0 to infinity |
| 6. Work count | Integer | 0 to infinity |
| 7. Numbers count | Integer | 0 to infinity |
| 8. Pronouns count | Integer | 0 to infinity |
| 9. Insights count | Integer | 0 to infinity |
| 10. Loan amount | Integer | 0 to infinity |

Table XXX: Predictor variables, types and ranges of model data set

1. Linear Regression

For our first approach of modelling the relationship between the project’s linguistic factors and its funding speed or gap, respectively, we use a linear regression with the ordinary least squares method, the most commonly applied form. It provides a solution to the problem of finding the best fitting straight line through a set of data points and works by minimising the squared residuals between the predicted dependent variable and the one actually observed to estimate the model’s parameters [TODO Reference: Lawson, C. and Hanson, R. [*Solving Least Squares Problems.*](http://www.amazon.com/exec/obidos/ASIN/0898713560/ref=nosim/ericstreasuretro) Englewood Cliffs, NJ: Prentice-Hall, 1974.].

The size of the coefficients obtained for each independent variable allows us to deduce their direct impact on the outcome of the dependent variable. Specifically, this helps us gain awareness of which parameters influence the magnitude of a project’s success to which extent. In order to make the coefficients easier to interpret, we normalised the independent variables about their mean and standard variation - that is, we subtract the mean from each observed value and divide it by its standard deviation. Now that the variables are all measured on the same scale, the coefficients are easily comparable.

In order for a linear regression model to be fully applicable, several assumptions are to be fulfilled. First, the analysis requires all independent variables to have a linear relation to the variable to predict. Second, linear regression needs all variables to be multivariate normal. Third, no or only multicollinearity between variables is allowed. Lastly, the regression assumes homoscedasticity, meaning that the residuals are equally distributed across the regression line.

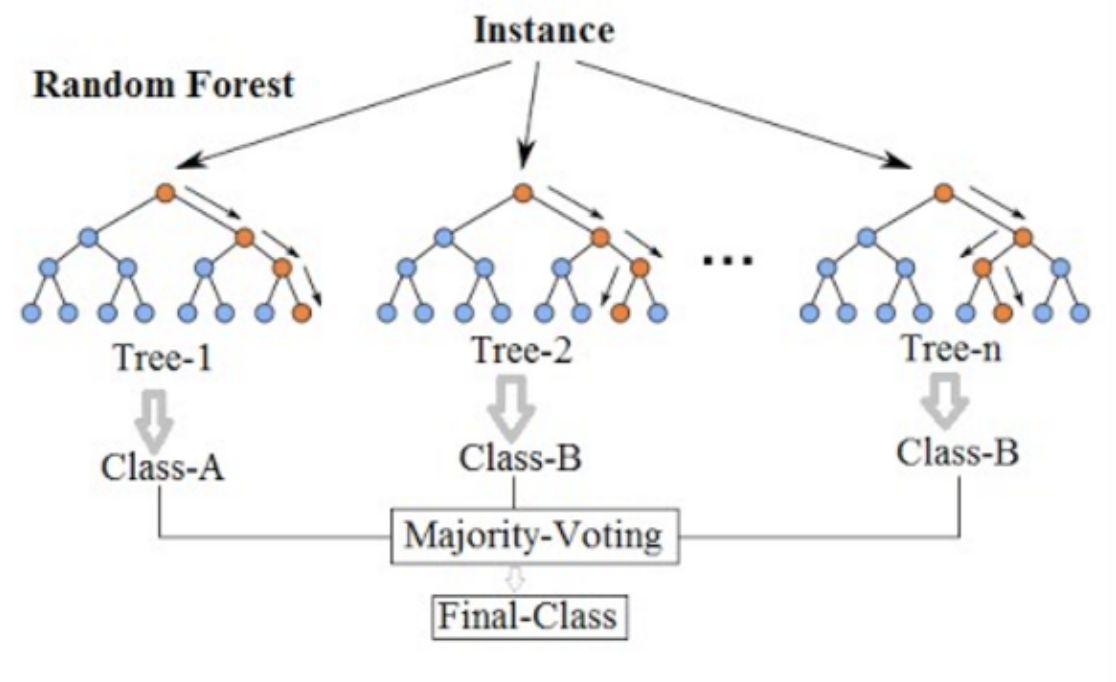
Through the explanatory analysis in section 5 we know that some of these assumptions are not fulfilled, however. As intuitively expected, Figure XXX shows a non-linear relationship between funding speed and sentiment score for successful loans. Additionally, when interpreting figure XXX and XXX, we notice that neither sentiment score for both small and large loans nor magnitude for larger loans show a linear relationship with funding gap. Next to the linearity assumption, the normality assumption is also not violated for one variable. Figure XXX depicts the distribution of the dependent variable funding speed with two significant peaks, thus being not normally distributed.

The linear regression models for both dependent variables, funding speed and funding gap, were constructed with means of the Statsmodels’ ordinary least squares model [TODO: REFERENCE].

1. Random Forest

As second approach we used the random forest model to predict the magnitude of a project’s success or failure through the use of linguistic factors. It is an ensemble supervised learning method which can be used for both classification and regression. According to the main principle behind ensemble methods which is to combine a group of machine learning techniques in order to form a strong model with increased predictive capabilities, the random forest algorithm operates by constructing an ensemble of decision trees at training time. It merges these trees to gain a more accurate and stable prediction and outputs the mean prediction for a regression or the class that is the mode of the classes of the individual trees for a classification. To build our model we used the random forest regressor from scikit-learn’s library of ensemble-based methods [TODO: Reference: http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html].

Throughout the project we experimented extensively with a number of configurations of the random forest model. First, it was applied on the complete set of independent variables for each of our four data sets, while experimenting with the number of trees the algorithm builds before taking the average of the predictions. In general, a higher number of trees not only corrects for the decision tree’s habit of overfitting to training data but it also increases the performance and makes the predictions more stable. However, at the same time, it also slows down the computation. This tradeoff between speed and predictive power must always be considered when choosing hyperparameters of the model like the number of trees used or the maximum number of features allowed to try in an individual tree.



We then applied the model on various combinations of the independent variables by using alternative compositions of the sentiment score. The train test split function [TODO: REFERENCE] was used every time the model was fit to randomise the training and testing data. Besides the predictive power of the model, we are also interested to understand how fast our model learns. Varying test/train splits allowed us to investigate the model’s behavior. In order to gain a better awareness of how the dynamics of the Kiva website changed since its beginnings in 2015, we further tested our model’s performance to predict the magnitude of success and failure of recent projects while training it on early data.

Another great quality of the algorithm is the importance score the scikit-learn’s library provides of each feature. As explained by Breiman & Jerome (20XX) [TODO: Reference 12] it measures the features importance by examining how much the tree nodes using a feature reduce impurity across all trees in the forest. After training, the score is automatically computed and scaled for each feature so that the scores sum to 1. It describes how informative and useful the feature is in predicting the dependent variable. In the context of this thesis, this helps us gain a better understanding of which variable contributes to predicting the magnitude of the project’s success or failure, respectively. Feature importances were obtained using the features importances attribute of scikit-learn’s library [TODO: REFERENCE: http://scikit-learn.org/stable/auto\_examples/ensemble/plot\_forest\_ importances.html#sphx-glr-auto-examples-ensemble-plot-forest-importances-py].

* confusion matrix

**6.2 Results**

In this section, the results of the models used are presented. We start by demonstrating and explaining the results of the linear regression followed by the random forest algorithm aimed at predicting our first dependent variable funding speed. Thereafter, we examine both models used for the second dependent variable funding gap. Finally, we interpret our findings in the context of Kiva project’s magnitude of success and failure.

**6.2.1 Modelling funding time**

1. Linear Regression

In order to compare the coefficient values and thus the individual variable’s impact on the dependent variable, the predictor variables were brought to the same scale by applying a z-score normalisation. We first applied a linear regression to successful, small loans as provided by data set A. The table below provides an overview over the regression results, including the regression coefficients, their significance values and the overall model fit. With only 9.8% of the variation within funding speed being explained by the independent variables used in our model, the linear regression delivered unsatisfactory results.

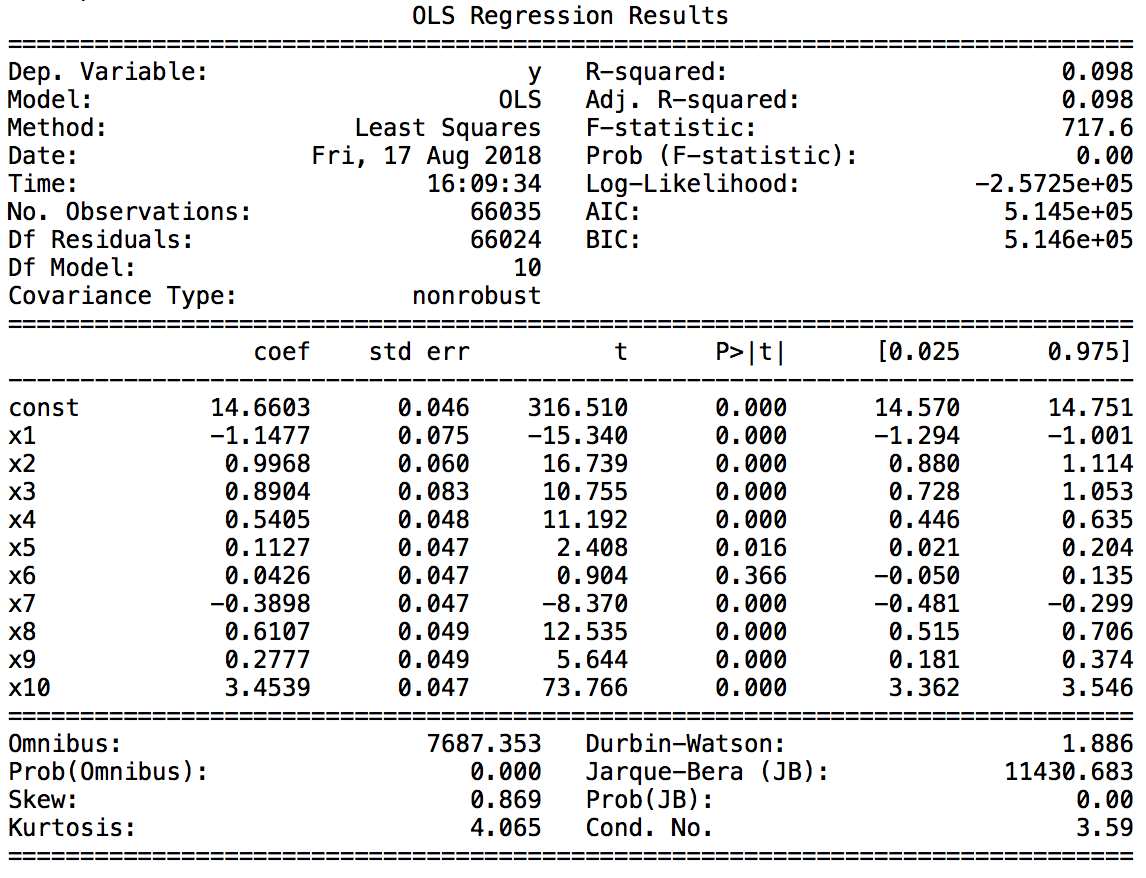


Table XXX

Table XXX summarises the results of the linear regression applied to successful projects with loan amounts above $1,000, as represented through data set B. Similar to the ones for small loans, the findings of this analysis can be regarded as unsatisfactory with r-squared values consistently below 0.1.

One reason for the bad fit of both models might, among many other possible causes, be the two violated assumptions for two of the ten predictor variables of the model as described in section 6.1. In some cases, it’s also possible that additional predictors can increase the true explanatory power of the model. More probable, however, is that the data contain an inherently higher amount of unexplainable variability. After all, we are dealing with linguistic characteristics as predictors that are very likely to be perceived only unconsciously by potential lenders.

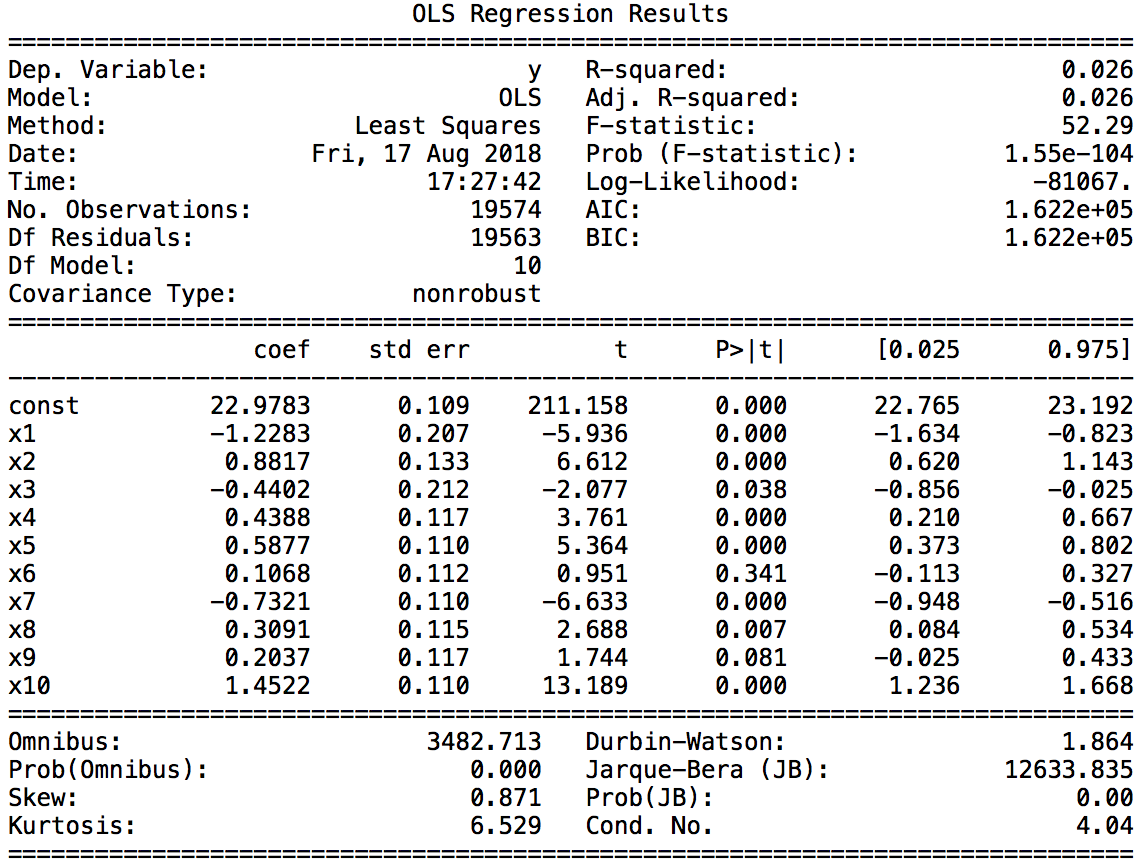


Table XXX

Despite a low effect size, the findings may, however, still be worth being interpreted as nine out of ten variables are significant according to their p-values. For both kinds of loans - small and large - the loan amount bears the most significance compared to the other predictors. The relatively high beta values allow a clear interpretation that increasing the amount requested is likely to increase the funding gap and particularly the time it takes to raise the full amount. This difference between the two coefficients indicates that while the loan amount still matters for large loans, it is of more importance for small loans. These findings are in accordance with the findings of Bourdeau de Fontenay [TODO: Reference] whose research we used as basis to control for other significant influences.

The next biggest beta score suggests that sentiment score has an influence on funding speed, albeit small, which is equally large for both small and large loans. Strangely, the positive coefficients of sentiment score disagree with the hypothesis we established initially. According to the model, a description including a large amount of positive emotion, boosts the time needed to receive funding. However, we hypothesised that the more emotion involved - independent of being positive or negative - the faster the borrower will raise the targeted amount. As this does not depict a linear relationship between sentiment score and each of the response variables, the linear regression model likely encountered problem predicting this variable properly.

Surprisingly, however, sentiment magnitude reveals a positive beta value as well, contradicting the aforementioned hypothesis by suggesting that a higher level of emotional content increases funding time for small loans. Although this result does not seem intuitive, this connection was already signaled by the scatter plot depicting the relation between funding speed and sentiment magnitude in figure XXX.

Two more relevant findings can be deduced by the results of the linear regression. First, the amount of work-related words within a description seems to be not significant when predicting the magnitude of a project’s success. Second, the Durbin-Watson score is close to 2 for small as well as large loans, indicating no significant signs of autocorrelation. This is not only relevant as it is one of the assumptions of linear regression models but also affects the random forest algorithm which will be presented next.

Due to the bad model fit of the linear regressions for both small and large loans, we did not experiment with alternative compositions of the sentiment score predictor variable with this model.

1. Random Forest

Compared to the linear regression, our second approach, the random forest model, left us with a slightly better model fit. As this is a predictive model which needs to be trained on separate data than on which it is tested, we used 75% of the data to train and 25% to test the predictive power of the random forest approach.

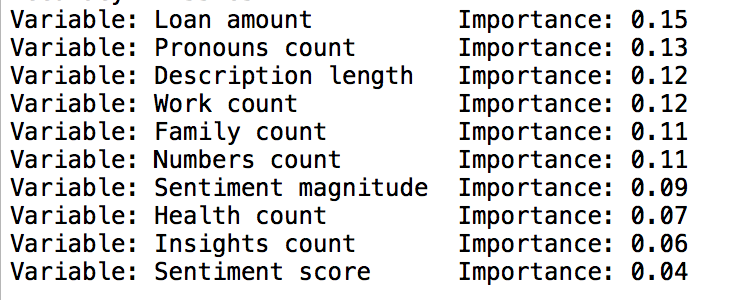
When fitting the original model as described in section 6.1, we obtain a r-squared value of 0.17 for small and 0.1 for large loans, indicating that 17%, or 10% respectively, of the variability in funding speed is explained by our predictors. By changing the amount of trees to be generated by the algorithm we only achieve minor improvements of this value. Even when testing a rather extreme amount of 1,000 trees to be built we do not realise improvements of a size that justify the big increase in computational power and time the algorithm needs. For this reason, we decided to use the algorithm with an amount of 100 trees throughout this project. This is the case for small as well as for large loans. The table below depicts the results generated for a varying number of estimators for small loans.

|  |  |  |
| --- | --- | --- |
| **Number of Estimators** | **R-Squared** | **Mean Absolute Error (MAE)** |
| 10 | 0.08 | 226.34h |
| 30 | 0.14 | 221.06h |
| 100 | 0.17 | 218.78h / 9.14d |
| 1,000 | 0.18 | 217.93h |

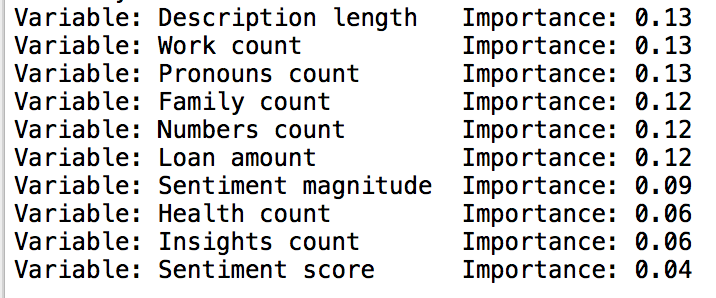
Table XXX: R-squared values for different number of estimators for data set A

In order to compare the results and gain a deeper understanding of them, we established two baselines. The first baseline is calculated by using the mean of the response variable as prediction for each observation and thereby computing the mean absolute error (MAE) between the predictions and the actually observed values. The second baseline is computed similarly. Instead of using the mean as prediction we use the median of the response variable to generate the MAE. This allows us to test whether our model beats the simple average or median, respectively, of the predictor variable for all observations. Applying these two baselines, we receive a MAE of 252.12 hours (10.52 days) and 240.26 hours (10.03 days) for small loans, respectively. For large loans, the baseline results are 279.1 hours (11.63 days) for mean and 305.31 hours (12.71 days) when using median predictions. Comparing these results to the errors shown in table XXX **(see above)**, we can conclude that our model beats both baseline models for all sizes of loans.

The random forest model indicates which predictors are most useful for predicting the independent variable through importance scores. The table below depicts the importances of features in our model for small and large loans. The higher the importance score, the larger the impact this variable has on the response variable. Comparing the feature importances for small to the ones for large loans, we notice that both models consider most predictors as equally important. In general, the top positions are scored by loan amount, the description length, the work, pronouns, family and numbers count. The only discrepancy which is particularly apparent is the importance of loan amount which with 15% is regarded as the most relevant feature for predicting funding speed of small loans. Surprisingly, both sentiment indicators are ranked around the lower positions.



Dataset large loans:



In order to visualise the performance of our model, we made use of a confusion matrix. For this purpose, we classified all observations into four quartiles, depending on the magnitude of their response variable. For example, the first quartile contains loans that are funded very fast, specifically with values that fall in the range between 0 and 4 days. The cutoff point for the next quartile is the 50th percentile, namely 10 days, thereby reflecting projects which are funded with medium speed, and so on. A confusion matrix allows insights not only into the errors being made by our classifier but more importantly into the types of errors that are being made. The figure below depicts the confusion matrix for small loans, as represented by data set A. The percentages within the matrix indicate how many projects are predicted to fall within the category shown on the x-axis while actually being in the one shown on the y-axis. For example, 4% of all projects that are funded fast, i.e. belong to quartile 1, are actually predicted to be funded rather slow, i.e. in quartile 3, by our model. The colored range to the right of the matrix sets the percentages in the context of absolute numbers.

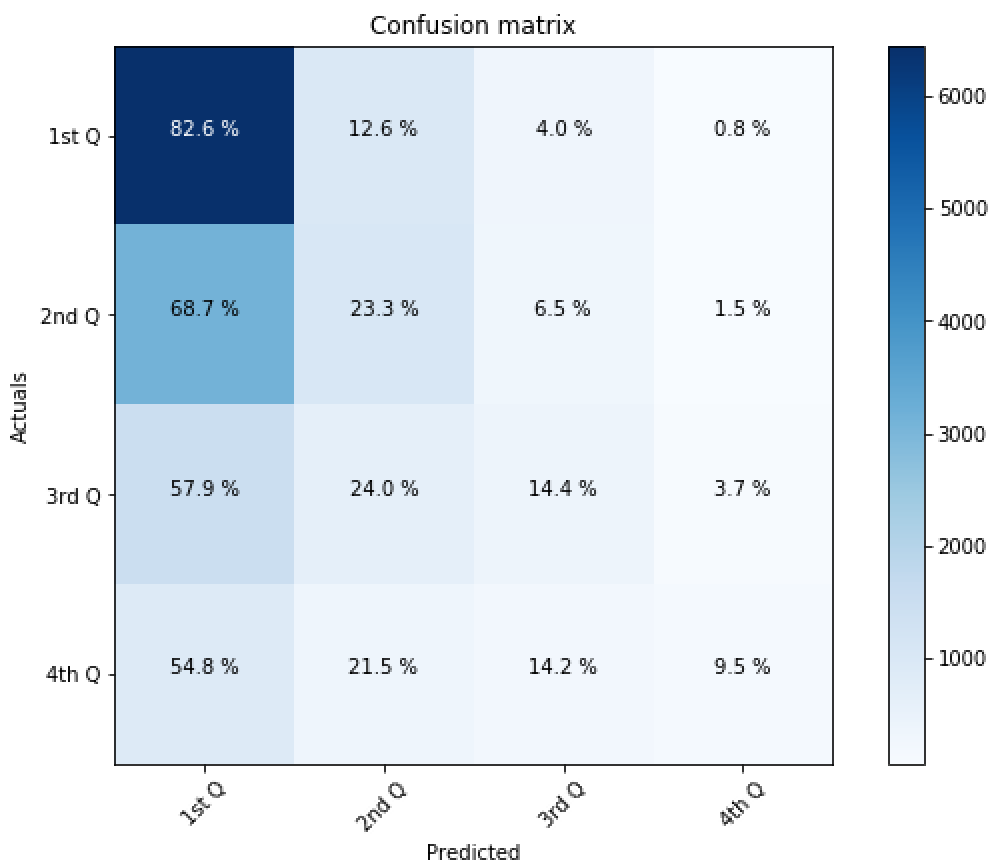


Figure XXX: Confusion matrix for data set A

When interpreting the percentages detailed in the matrix, we notice a clear tendency to predict an observation to be in the first quartile, i.e. be funded very fast. This is probably due to the large amount of projects that fall within this first cohort. Thereby, our model predicts 82.6% of observations that are funded fast to also fall into this category. With 24%, 14.2% and 9.5%, respectively, the other three cohorts have enormously lower success rates. The overall accuracy lies at 48.5%, meaning that 51.5% are misclassified by our model. This misclassification rate is especially noticeable for loans that are funded very slowly.

The next figure shows the confusion matrix for large loans. In contrast to the previous figure, this matrix seems to be much more balanced. Although projects with a high funding speed again have the highest accuracy of 43.5%, the matrix shows no tendency to predict classes with a higher funding speed more accurately than for example the slowest quartile of projects. The numbers in the diagonal reflect the success rates of each class and suggest through consistent percentages above 25% that our model performs better than simple guessing. One exception is the second quartile, i.e. the projects that are funded with medium speed, as here the accuracy lies at 25%. Overall, the accuracy of the matrix is at 34.9%, leaving a misclassification rate of 65.1%. These results suggest that our model performs better for predicting smaller than larger loans and thereby support the higher model fit of the random forest regressor for small loans as opposed to large ones.

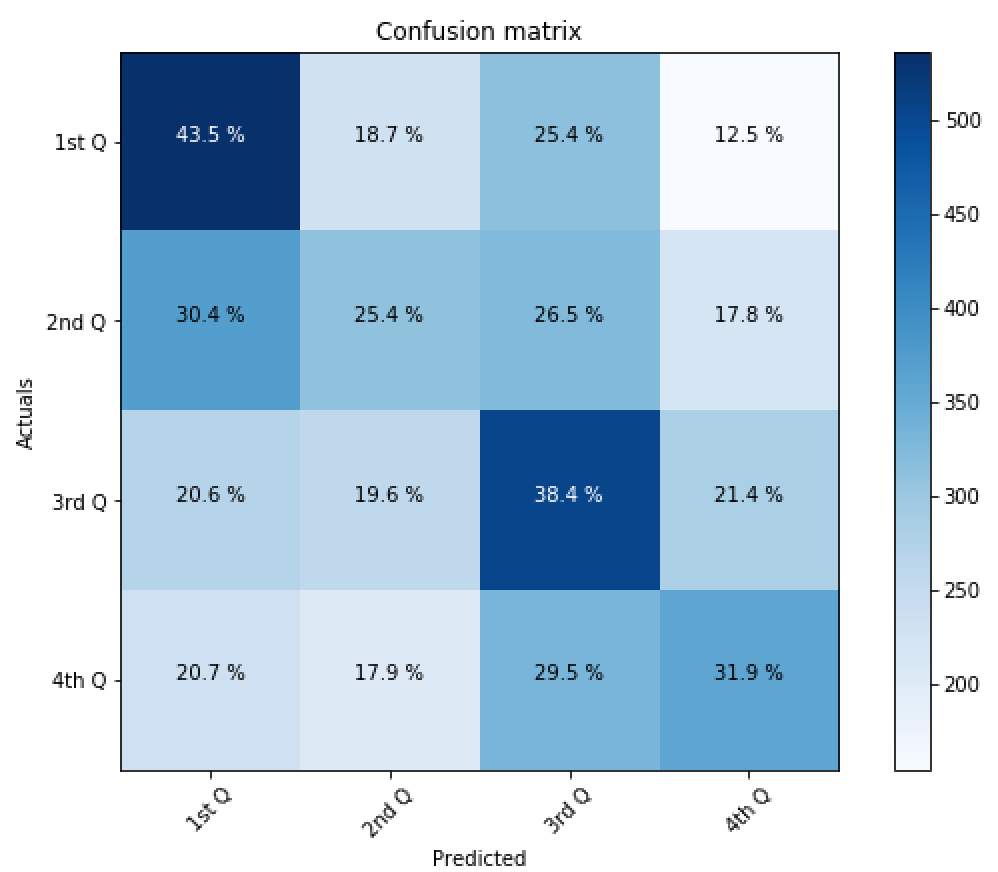


Figure XXX: Confusion matrix for data set B

Next, we experimented with different perceptions of sentiment within a given text. We ran the random forest model with three alternative compositions of the variable sentiment score. First, instead of using the average of of all individual sentences’ scores, we made use of the median. Second, we excluded all scores of individual sentences that were 0 to avoid the mitigating effect of neutral sentences in a description. Last, we assigned sentences in the first and last quartile of a description double the weight than the ones in the middle. The resulting model fit is compared to the one achieved with the basic sentiment score and shown in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Basic** | **Median** | **Excluding zeros** | **Differing weights based on sentence position** |
| **Small loans** | 0.167 | 0.171 | 0.169 | 0.172 |
| **Large loans** | 0.098 | 0.098 | 0.098 | 0.104 |

Table XXX: R-Squared of random forest model with alternative sentiment scores

None of the alternative compositions of the variable seem to be a deterioration to the model fit and most of them even achieve improvements, albeit minor. The one enhancing the r-squared value of our model the most seems to suggest that our idea of humans being influenced by the emotions contained in the first and last sentences of a text might be a step in the right direction. By changing to this composition of sentiment score, we also enhanced its importance scores, rising from a mere 4% to 11% for both kind of loans.

A further examination of our model was carried out to compare various test/ train splits. First, we aim at gaining a deeper awareness of the dynamics of the micro-lending website over time. Instead of using randomised training and testing data, the model is trained on data from Kiva’s earliest years while being tested on 25% of the most recent data. From r-squared values of -0.134 and -0.099 for small and large loans, respectively, it can be deduced that the behavior of Kiva’s early adopters is not representative of the current lenders’ behavior, neither for small nor for large loans.

Second, we seek to understand the model’s learning aptitude. The table below reflects r-squared values when testing further variations of a randomised train/ test ratio. Comparing these results to the r-squared of our model when being trained on a standard 75% of the data, we can conclude that our model learns rather quickly, suggesting a relatively stable lending behavior of lenders.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **25% Train / 75% Test** | **50% Train / 50% Test** | **75% Train / 25% Test** |
| **Small loans** | 0.154 | 0.159 | 0.167 |
| **Large loans** | 0.054 | 0.075 | 0.098 |

Table XXX: Comparing r-squared values of differing train/ test ratios

**6.2.2 Modelling funding gap**

1. Linear Regression

The linear regression carried out to model the continuous dependent variable funding gap for small loans based on linguistic factors delivered satisfying results. 32.4% of its variability is explained by the ten predictor variables. Through a z-score normalisation, all variables were brought to the same scale in order to compare and interpret their coefficients more easily. The table below shows the results of the linear regression model for small loans as represented by data set C. Important features are the regression coefficients for each predictor, the p-values, here expressed by the denotation “P> |t|”, the model fit as well as its overall significance level.

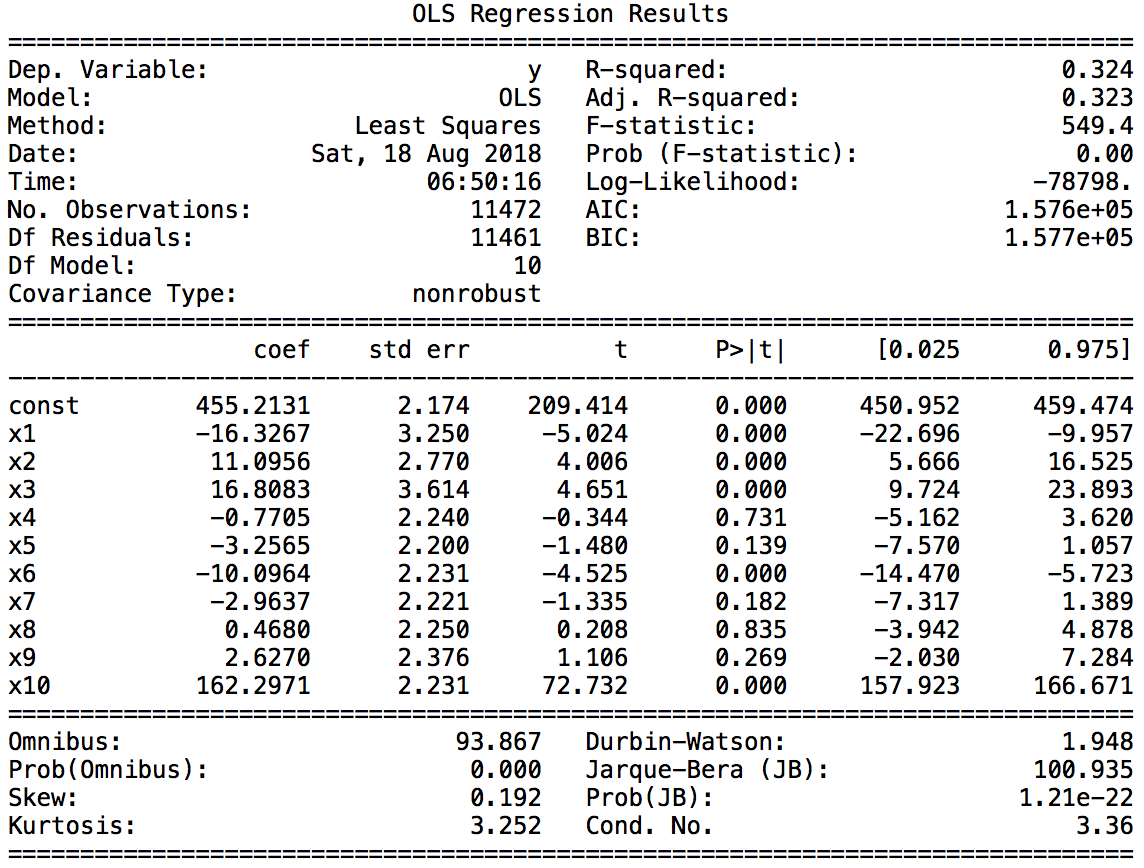


Table XXX: Linear regression results on data set C

A second linear regression was used to model the relationship between funding gap and the linguistic factors for larger loans as represented through data set D. Table XXX depicts its results. The r-squared value of 0.626 and a high significance level as represented through the F-test of overall significance suggest a strong relationship between our model and the response variable.

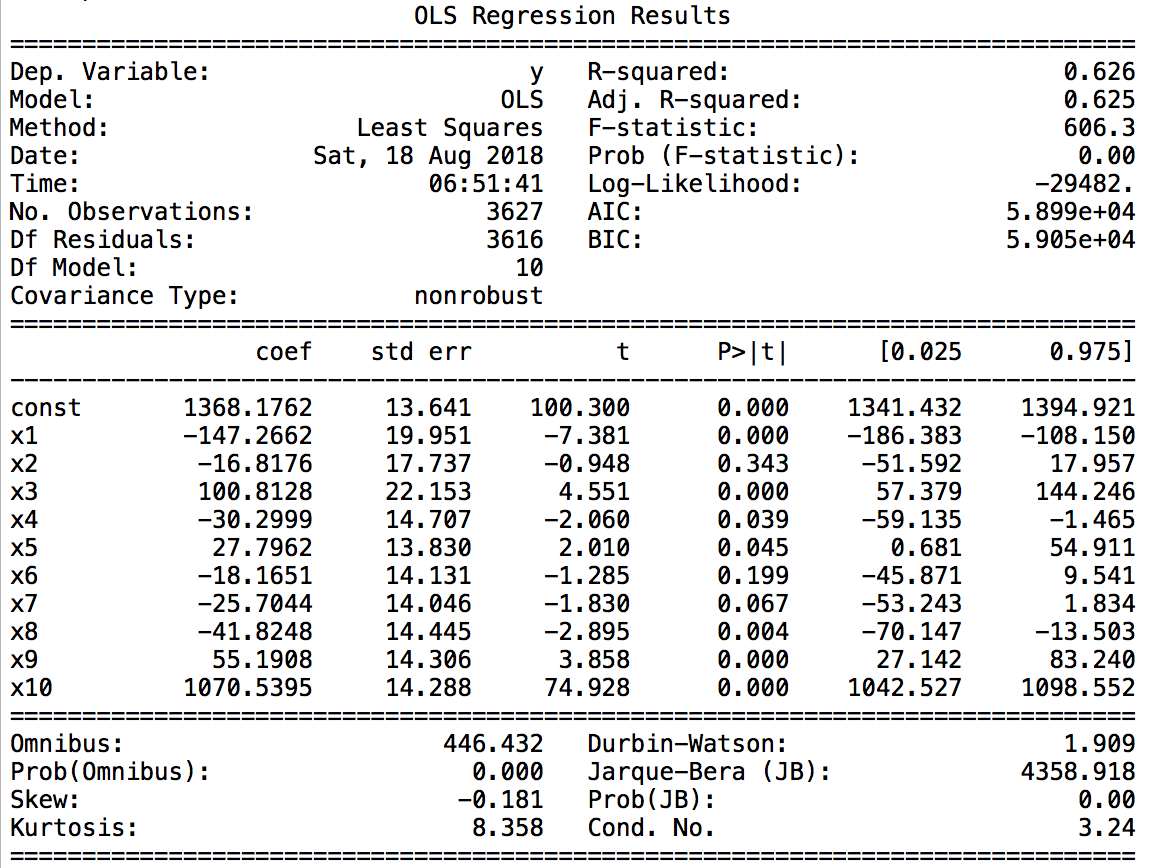


Table XXX: Linear regression results on data set D

When interpreting the regression coefficients obtained, loan amount stands out for both small and large loans by being the highest of all beta values. It has the highest impact on funding gap relative to the other examined predictors, suggesting that even a small raise in the targeted loan amount, changes the borrower’s prospects with regard to the size of a funding gap. The positive sign of the coefficient indicates that, as intuitively expected, the best chances of ending up with a small funding gap are simply to ask for less.

The next highest beta coefficients for both loan sizes belong to the two predictors description length and sentiment magnitude. Whereas the high, negative coefficient of description length supports our hypothesis that a more detailed description relates to a smaller funding gap, the positive value for sentiment magnitude, however, contradicts our initial belief that a more emotionally charged project description relates to a smaller funding gap.

Apart from these three predictor variables, the funding gap of small loans is largely attributable to different factors than the one of large loans. While the sentiment score still plays a large role for small loans with a more positive emotional score relating to a larger funding gap, the remaining variability that is explained by the model is mainly predicted by the amount of words belonging to the selected LIWC categories. Surprisingly, insights- and health-related information can be associated with higher funding gaps, whereas a higher amount of pronouns, family- and numbers-related words predict smaller funding gaps. These findings only partly support our hypothesis.

The difference in variables that are not statistically relevant between small and large loans further highlight the fact that small loans are influenced by different features than large loans are. Additionally, the higher model fit for large loans suggests that the language used within a loan description plays a larger role for projects asking for higher amounts of money. As our exploratory analysis proved, however, two assumptions of the linear model were not fulfilled for every predictor variable. Coupled with a few surprising findings that were revealed throughout this section, we decided to further conduct a non-linear model, the random forest.

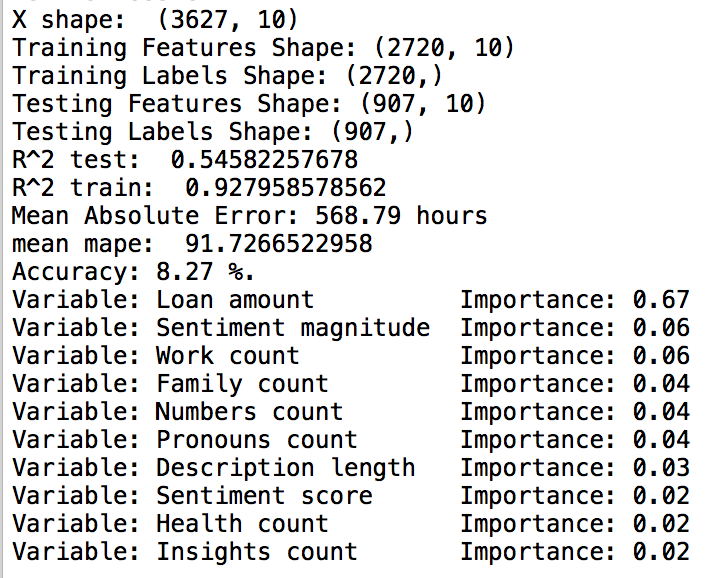
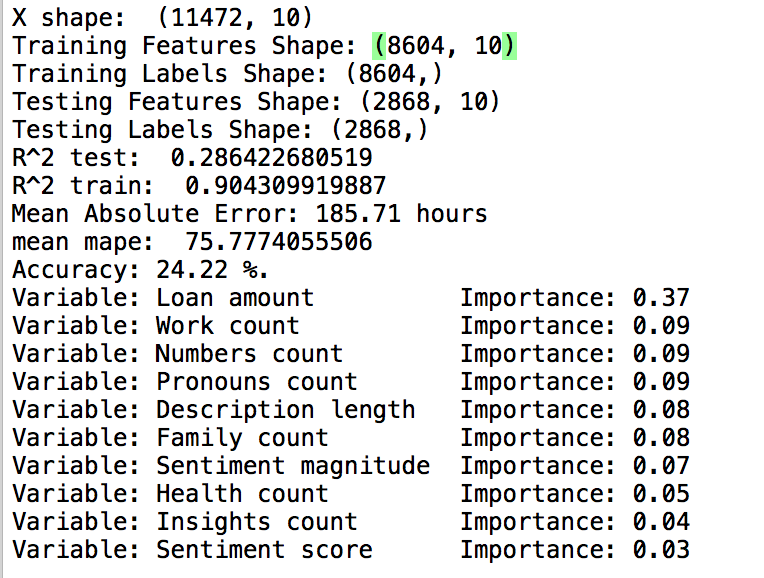
1. Random Forest

The random forest algorithm resulted in a reasonably good model fit with r-squared values of 0.286 for small and 0.546 for larger loans. In order to assess the quality of our model, we compare the mean absolute error to the ones computed by using the two baselines established in section 6.2.1. The first baseline uses the mean funding gap as prediction for each observation whereas the second baseline uses the median of the funding gap as predictions. The table below compares the values and we conclude that our model beats both baseline model for all sizes of loans.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Model results** | **Mean baseline** | **Median baseline** |
| **Small loans** | $185.71 | $225.13 | $219.21 |
| **Large loans** | $568.79 | $751.09 | $724.25 |

Table XXX: Comparison of mean absolute errors of model vs. baseline

The importance scores outputted by the algorithm indicate which predictors are most useful for predicting the response variable. The figure below compares the importance values for smaller projects to the ones for larger projects. We clearly observe that the most important factor for predicting the funding gap for any kind of loan size is the loan amount. We notice a significant gap between loan amount and the next highest variable after which the remaining 9 variables line up closely together. Out of all features considered, changing the amount requested by the borrower influences the magnitude of a project’s failure the most. This result suggests that linguistic factors do not play a predominant role when predicting the size of funding gap for unsuccessful projects.



To gain a new perspective of our results, we visualised them using a confusion matrix by dividing our observations into four different quartiles based on the magnitude of funding gap. The first quartile contains unsuccessful projects that only narrowly failed, while the last three contain projects that failed with a medium gap left to their targeted amount, a large gap and a very large gap, respectively. Figure XXX and XXX show the confusion matrices constructed for small and large loans as represented through data set C and D, respectively.

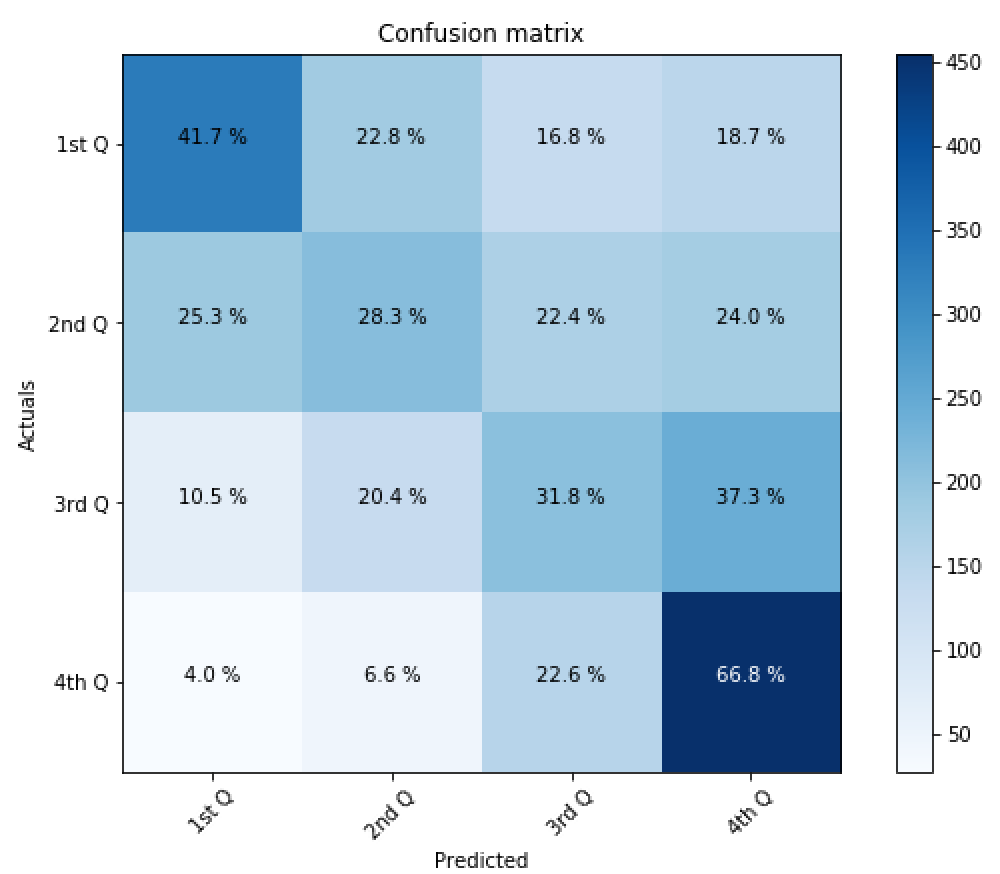


Figure XXX: Confusion matrix for data set C

Although no clear tendency is recognizable, we notice higher success rates for the first and the last quartile in both matrices. This suggests that projects that only failed by a narrow margin, i.e. those who have a small funding gap, as well as unsuccessful projects with a very large funding gap, are likely to have some distinguishing characteristics. With both accuracy values being around 40%, our model predicts small and large loans equally well.

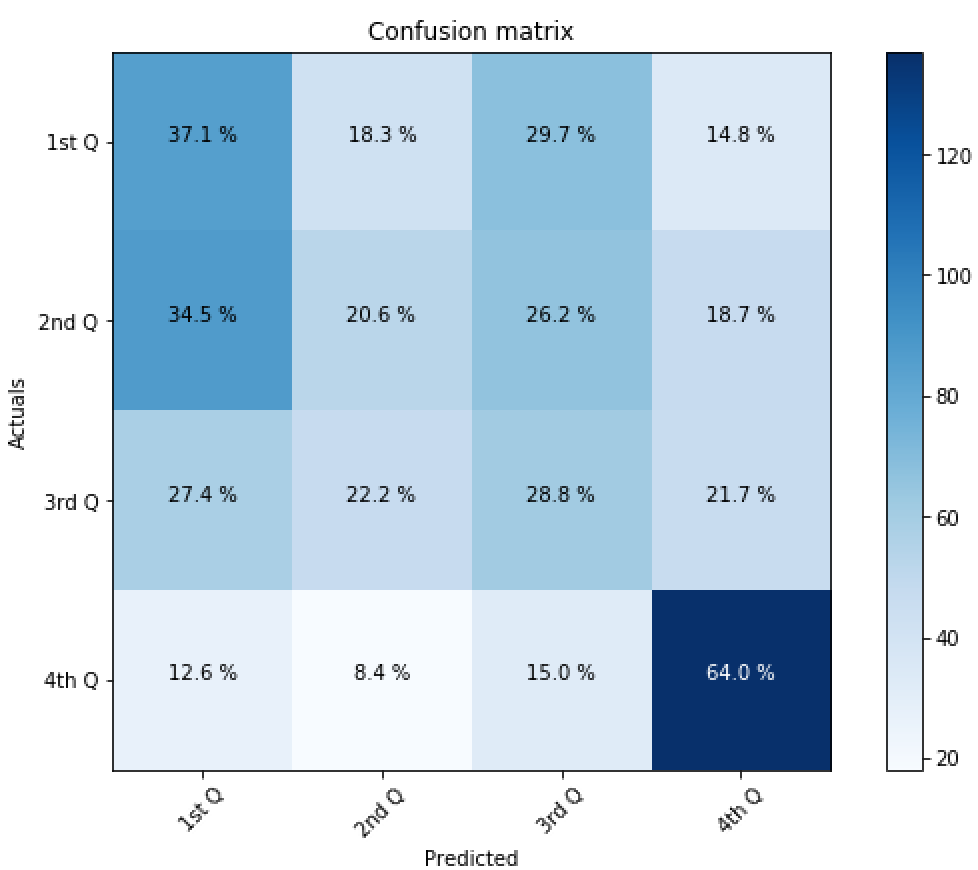


Figure XXX: Confusion matrix for data set D

In order to test various hypotheses about how humans perceive emotions, we ran our model with three alternative compositions of sentiment score. First, we used the median of the individual sentences’ sentiment score instead of the mean as done by the Google Cloud API. Second, we excluded all neutral sentences from the overall description’s score. Lastly, we lay double the emphasis on all sentences in the first and last quartile of the description. The r-squared of the resulting models are shown in the table below. None of the compositions changed the model fit significantly for both kind of loans. Interpreting these findings, we deduce that emotional factors do not play a predominant role when predicting the magnitude of a project’s failure, reinforcing the findings of our basic model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Basic** | **Median** | **Excluding zeros** | **Differing weights based on sentence position** |
| **Small loans** | 0.286 | 0.284 | 0.277 | 0.282 |
| **Large loans** | 0.546 | 0.547 | 0.543 | 0.537 |

Table XXX: R-Squared of random forest model with alternative sentiment scores

Next, we investigated whether the behavior of lenders at Kiva changed over time by training our model on data from Kiva’s early years. When testing it on recent projects, some interesting findings are revealed. While the r-squared value for smaller loans does not change significantly, the value for large loans drops from 0.546 to 0.321. This implies that the behavior of lenders investing in small loans has remained relatively stable over time. The behavior of current lenders interested in large loans, however, has changed significantly from the behavior of Kiva’s early adopters. This is also reflected when experimenting with various test/ train splits. While the model’s fit does not vary noticeably for small loans, it does for large loans. With only 25% of training data, the model performs much worse, thereby implying that the lenders’ behavior is not rather unstable.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **25% Train / 75% Test** | **50% Train / 50% Test** | **75% Train / 25% Test** |
| **Small loans** | 0.280 | 0.284 | 0.286 |
| **Large loans** | 0.363 | 0.572 | 0.546 |

Table XXX: Comparing r-squared values of differing train/ test ratios

Variables we have are better in capturing behavior for large loans

**7 Conclusion & Future Work**

**7.1 Conclusion**

The aim of this thesis was to understand, characterise and predict the relationship between certain linguistic features and the magnitude of a project’s success or failure on the micro-lending platform Kiva. After retrieving relevant data through the Kiva API, we controlled for already observed influences such as borrower’s gender, thematic field and geographic location by limiting the data sets accordingly. The exploratory analysis helped us gain a deeper understanding of the predictor variables and to discover interesting areas to explore more into detail such as the various possible compositions of sentiment score. While the linear regression model delivered rather unsatisfactory results for predicting funding speed, the random forest algorithm allowed us to explain

To test the broad hypothesis

**7.2 Limitations**

**7.3 Future Work**

**8 Bibliography**

Ly, Pierre & Mason, Geri, 2012. "[Competition Between Microfinance NGOs: Evidence from Kiva](https://ideas.repec.org/a/eee/wdevel/v40y2012i3p643-655.html)," [World Development](https://ideas.repec.org/s/eee/wdevel.html), Elsevier, vol. 40(3), pages 643-655.