|  |
| --- |
| **KIVA.ORG** |
| **Project Kiva - Analytics 3.0: Examine** |
|  |

|  |
| --- |
| **GROUP – TEAM EAGLES**   1. Chethan Kumar Thimmaiah - 11510880 2. Sai Deekshith Ratna - 11525390 3. Vinay Vuppala - 11538317 4. Omkar Vamsi Krishna Chinta - 11515782 5. Leela Anjani Dwarampudi - 11524768 |
|  |

Table of Contents

[1. Abstract: 2](#_Toc1931315803)

[2. Introduction 3](#_Toc2057293883)

[2.1. Kiva.org 3](#_Toc1926443791)

[2.2. What is Microlending through Crowdfunding? 3](#_Toc1585139373)

[2.3. Loan process 4](#_Toc1481300534)

[3. Methodology 4](#_Toc76251160)

[3.1. Capturing Data 5](#_Toc154915277)

[3.1.1. The Lenders 5](#_Toc1336741516)

[3.1.2. The Field/lending Partners 5](#_Toc1029772446)

[3.1.3. The Loans 6](#_Toc619257338)

[3.2. Processing Data 10](#_Toc896885204)

[3.3. Methods of Analysis 11](#_Toc1046342384)

[1.1.1 Theoretical Overview 12](#_Toc755313227)

[3.4. Data Exploration: 13](#_Toc1277741372)

[3.5. Model Building: 13](#_Toc471590135)

[3.6. Testing & Validation: 14](#_Toc620283887)

[3.7. Model Evaluation: 14](#_Toc297489622)

[3.8. Model Comparison & Model Evaluation: 14](#_Toc120531631)

[4. Conclusions & Further Study 14](#_Toc778491391)

[6. References 15](#_Toc480806393)

## Abstract:

Non-profit Micro-finance organizations offer lending opportunities to alleviate poverty by financially supporting underprivileged, yet competent entrepreneurs who are in dire need of an establishment that provides loans to those who have need. Micro lending through crowdfunding is the technique of raising investment to fund a venture or loan through several depositors using an innovative digital platform. *Kiva.org*, a widely used crowd-funded micro-financial service, offers students & researchers an extensive volume of publicly available data comprising a rich set of varied information about micro-financial operations. The aim of this project report is to study the available data of borrowers & lenders with Kivs.org using CRISP DM process in Enterprise Miner to build a most appropriate machine learning model to predict whether loan application gets funded or not. Also, to examine and conduct descriptive and diagnostic analysis using data visualization tool – Tableau to identify factors that are behind Kiva’s success. The report emphases on Kiva.org, one of the most successful crowdfunding platforms in the world.

**Key Words: Loan, Crowdfunding, Field-Partner, Repayment, Disbursement, Descriptive analytics, Machine learning, Accuracy, Sensitivity, Specificity,**

## Introduction

## Kiva.org

Founded in 2005, Kiva.org is one of the world’s largest online crowdfunding platforms where people can lend money to underserved entrepreneurs across the globe. With the mission of “connecting people through lending to alleviate poverty,” Kiva endeavors to provide safe and inexpensive access to investment in areas where traditional banking services is ineffective to meet the requirement, thereby empowering people to create opportunities to become entrepreneurs. In fact, Kiva acts like *UBER*, an online connection between borrowers and lenders. The profiles of borrowers from different parts of world who need affordable loans are posted on Kiva’s platform. Lenders who volunteered to support needy people can browse the different profiles and provide loan to chosen projects, based on loan application story, through a video or characteristics of the loan request and the borrower. Based on data collected in Nov 2021, Kiva is operating in more than 70 countries and works in partnership with 304 field partners. This partnership with field partners is a unique feature of Kiva, compared to other lending platforms. The field partners’ role is to enable the process by creating a regional presence and advancing the capital to the borrowers prior to posting the loan request on the Kiva platform, to give the borrowers a head start in their business endeavor. The focus areas of Kiva are a) agriculture b) refugees and c) gender. To date, the organization has crowdfunded more than 4.4 million loans, totaling more than $1.77 billion, at a repayment rate of 96.3%.

A picture containing logo

Description automatically generatedLogo, company name

Description automatically generated

## What is Microlending through Crowdfunding?

Crowdfunding is the method of raising money to fund a project or business venture through numerous investors and via an Internet platform. Online crowdfunding is a relatively new phenomenon that has increased the number of ways in which consumers, entrepreneurs and organizations can access capital. In principle, crowdfunding platforms are designed to put individuals who are willing to lend or invest their money in contact with other individuals, projects or businesses that need financial support. Crowdfunding as a concept is evolving in Internet space since 1997 when fans of the British rock band Marillion raised US$60,000 in donations by means of an Internet campaign to underwrite an entire U.S. tour. As per publicly available sources, there are 1,478 crowdfunding organizations in the US (Crunchbase, 2021). Currently the three largest crowdfunding platforms are Kickstarter, Indiegogo, and Crowd Supply. As of January 2021, Kickstarter has raised more than $5.6 billion spread over 197,425 projects.

**2.2.1 Microfinance credit theory.**

The idea of group lending is frequently recognized as the key development in microfinance and asserts to offer a solution to the drawbacks of inaccurate credit markets, particularly the difficulty of overcoming information asymmetries. Adverse selection and moral hazard are two separate phenomena that can result from information asymmetries. When there has been an adverse selection, the lender is unaware of how risky the borrowers are. Riskier borrowers should be charged higher interest rates to make up for the higher risk of default because they are more likely to do so than safer borrowers. Assortative matching or screening to address adverse selection and peer monitoring to address moral hazard are two techniques that the conventional lending model typically includes.

**2.2.2 Women empowerment theory:**

Many women have access to loans for investments in their own self-managed businesses. Despite the difficulties they encounter on a daily basis, the vast majority of them have excellent payment histories. Contrary to popular belief, they have demonstrated that lending to the underprivileged and women is a wise decision. The gender rights movement claims that targeting women is justified due to the high women's repayment rate, the assumption that giving more women access to microfinance services will result in economic, social, and political empowerment for each individual, and the fact that women are responsible for the wellbeing of their households. The idea is that giving more women access to microfinance will, by itself, raise household income, improve well-being, and empower women to affect larger changes in gender inequity. According to this view, women can be empowered by having fair and equal access to resources, particularly those in the microfinance and financial sectors. The independent variables are validated by the theory.

## Loan process

The process begins when a borrower approaches a field partner to ask for a loan, or vice versa. The partner then assesses the profile of the barrower, after initial screen to meet requirements of kiva guidelines, partner uploads the borrower’s profile onto the Kiva platform, with a photograph of the borrower or along with a short video consisting of need for the loan and the amount requested along with intended use. Loan requests are posted on the portal for a span of 30 days, during which time lenders may respond to the request for funding. Finally, those applications that do not get full loan amount are announced as expired loans.

On the other hand, lenders who wish to loan can browse and choose a borrower they desire to fund. lenders can start with a minimum funding amount of $25 to any amount and a loan is funded by more than a few lenders. The lenders then transfer their funds to Kiva through a special service on PayPal, which waives its transaction fee and thereby saving costs other than funding. After receiving lenders' money, Kiva aggregates loan capital from the individual lenders and transfers it to the appropriate Field Partners, which disburse the loan to the borrower. Kiva does not charge interest on the capital sent to Field Partners, but often Field Partners do charge some level of interest to borrowers to cover administration costs. Interest is typically higher on loans from microfinance institutions in developing countries than interest rates on larger loans in developed countries because of the administrative costs of overseeing many tiny loans, and the increased risk. As the entrepreneurs repay their loans with interest, the Field Partners remit funds back to Kiva. As the loan is repaid, the Kiva lenders can withdraw their principal or re-lend it to another entrepreneur.

## Methodology

## Capturing Data

The Kiva data set is composed of various entities, a set of transparent data including unstructured data (e.g., text, image, and video) as well as structured data (e.g., geo-spatial, numerical, categorical, and ordinal data). Lender datasets contain key fields of name, image, registration timestamp, city, loan count, and other fields. The data also provides information about the number of loans funded and to any number of lender teams with which s/he is associated.

## The Lenders

Lenders has 157879 listings with 14 variables and third dataset loan lenders has 213078 listings with 2 variables. Lenders on Kiva may opt to disclose their info publicly. A lot of missing data in lenders profile may be due to their choice of disclosing. Every lender is recognized by the attribute lender\_id, which is a name, such as chethan, plus a number, as in chethan2749, to avoid repetition. The ID is missing in many cases. Other attributes we examine in this data set are country\_code (the two-digit ISO code), member\_since, occupation, loan\_count and invitee\_count. The data are quite sparse, with a high proportion of missing values in some attributes, as discussed below.

Table 3.1.1 Field names and descriptions of lender & their profile data.

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| PERMANENT\_NAME | Name of the lender ID |
| DISPLAY\_NAME | name of the lender |
| MAIN\_PIC\_ID | Image ID of the lender |
| CITY | City Name |
| STATE | State name of the lender |
| COUNTRY | Name of the Country |
| MEMBER\_SINCE | Lender membership date |
| PERSONAL\_URL  OCCUPATION  LOAN\_BECAUSE  OTHER\_INFO  LOAN\_PURCHASE\_NUM  INVITED\_BY  NUM\_INVITED | URL of the lender website if any  Occupation of lender  Intent of loan offerings  Any other information  No. of loans where she/he offered  Name of the person who referred to Kiva  No. of references s/he has given |

## The Field/lending Partners

Kiva’s lending partners are a broad network of groups spread across 70 countries on 5 continents. The majority of lenders are small to medium microfinance organizations, including schools, NGOs, social enterprises, and more. They engage with the communities, evaluate borrowers, help borrowers to connect with Kiva, post their loan application on the portal & support internal teams with loan disbursement & debt recollection. In addition to these services, they also provide entrepreneurial training, literacy skills, lending quality seed and farming inputs, and providing access to savings accounts and insurance. The motto of Kiva is with these lending partners to reach needy people across the globe by enhancing their lives through safe and fair access to credit.

Graphical user interface, application, Word

Description automatically generated

**Map

Description automatically generated**

## The Loans

The loan is the most significant data object in the Kiva database as other related loan characteristics are in some way related to a loan. Each loan has a record where Kiva or a partner can update any changes into the status. Each record has details of the loan, such as the sector of the intended use of the loan, the borrower’s gender and country of origin, and the status of the loan. Status is key in determining whether a loan has been *funded or not*. At a particular point in time, the loan status may be fundraising, funded, in repayment, paid, expired, defaulted, or refunded. If the loan has expired, defaulted, or been refunded, then it has not managed to complete the entire crowdfunding process successfully. The success rate may be impacted by several aspects, such as the gender and region of the borrower, credit terms or of a broad picture of the intended use of the loan. There are a few objects in the Kiva world that are important: a loan, borrower, lender, and partner. A loan is requested by a borrower and supplied by either an individual lender or a partner, a microfinance institution that partners with Kiva. The dataset shows the following table of loan status numbers wherein 95% of loan applications are approved & funded.

|  |  |  |
| --- | --- | --- |
| **Status** | **No of Loans** | **% Of Loans** |
| **Expired** | 93,435 | 4% |
| **Funded** | 2,078,544 | 95% |
| **Fund rising** | 5,810 | 0% |
| **Refunded** | 9,329 | 0% |
| **Total** | **2,187,118** | **100%** |

Table 3.1.3 Field names and descriptions of borrower and loan characteristics

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| LOAN\_ID | Unique ID for loan |
| LOAN\_NAME | name of the loan |
| ORIGINAL\_LANGUAGE | Language of loan description with levels English (en), Spanish (es), French (fr), Indonesian (id) |
| DESCRIPTION\_TRANSLATED | Loan description translated to English |
| FUNDED\_AMOUNT | Loan amount funded |
| LOAN\_AMOUNT | Loan Amount requested |
| STATUS | Loan Status |
| IMAGE\_ID | Image-ID of the borrower picture |
| IMAGE-BI | Image-ID Binary level |
| ACTIVITY\_NAME | Activity for which loan is requested |
| SECTOR\_NAME | Sector in which loan is sought |
| ACTIVITY\_NAME\_INT | Activity name categorized into numbers |
| SECTOR\_NAME-INT | Sector Name categorized into numbers |
| LOAN\_USE | Description about use of loan |
| COUNTRY\_CODE | Country Code |
| COUNTRY\_NAME | Country Name |
| COUNTRY\_CODE\_INT | Country code categorized into Numbers |
| TOWN\_NAME | Town Name |
| CURRENCY\_POLICY | currency policy whether forex fluctuation risk is shared or standard |
| CURRENCY\_EXCHANGE\_COVERAGERATE | % Of Forex fluctuation risk shared between field partner & lenders |
| PARTNER\_ID | Field partner ID |
| POSTED\_TIME | Loan posted time in Kiva portal for fund raising |
| PLANNED\_EXPIRATION\_TIME | expiration time of the loan |
| DISBURSE\_TIME | Disbursed time of the loan |
| RAISED\_TIME | Loan application raised time |
| LENDER\_TERM | Lender terms |
| NUM\_LENDERS\_TOTAL | Total number of lenders for a given loan |
| NUM\_JOURNAL\_ENTRIES | Number of journal entries for a loan |
| NUM\_BULK\_ENTRIES | No. of bulk entries for a loan |
| BORROWER\_GENDERS | Gender of the borrower |
| REPAYMENT\_INTERVAL | Repayment schedule |
| DISTRIBUTION\_MODEL | Loan distribution model whether it is Direct or through field partner |
| BORROWER\_NAMES | names of borrowers |
| VIdeoID | If a video was posted, its ID. |

**Loans by region**:

Asia has the maximum borrowers (more than one third) & rest are in African countries & other nations. Philippines has been the top 1 in loan barrowing so far with In African countries falling in top 10 countries. Most loans are sought for the food, retail, and agriculture sectors. Agricultural loans are the most popular in Asia, with food and retail scoring highly in that region as well. Agriculture is less popular in the Middle East and North America, where the records show a number closer to average. In the Middle East, services have the highest number of loan requests, followed by food and retail.

Map

Description automatically generated

Graphical user interface, text

Description automatically generated

Word cloud by country

Chart, waterfall chart

Description automatically generated

**Loans by Sector:**

The data set includes 2,187,118 loan submissions, grouped into 15 sectors, which are further segmented into 149 activities. Loan applications come mostly from the food, agriculture, and retail sectors. Activities related to the food sector include, for instance, fish selling, bakery, cereals, and dairy. At the other extreme, loan requests from the entertainment, wholesale and health sectors are the least frequent loans.

We used Tableau desktop applications to analyze and visualize loan data sets & created lot of descriptive charts to which tells us the patterns of loan application, majority of activity names, key regions where more applications are submitted etc.

Chart, treemap chart

Description automatically generated Chart, bubble chart

Description automatically generated

Chart, bar chart

Description automatically generated Text

Description automatically generated

**Loans by timeline:**

The number of loans registered on Kiva increased considerably in 2012 because of the credit limits plan that was introduced in 2011. The plan is designed to give partners flexibility by relaxing monthly fundraising limits and encouraging the quantity and range of loans on Kiva. Partners reacted vigorously to the plan by recruiting significantly more borrowers. As a result, the supply of loan requests significantly improved. Figure 6.1.1 depicts the annual posting of loan requests on the Kiva website since its inception

Chart, histogram

Description automatically generated

To conclude the description of the loan data, we note that the amount of the loan varies greatly, from $25 to $500,000, with a mean of $2773. In 2021, Kiva had issued loans of an average of $348,219 each day or $14,509 every hour or 242 per minute which amounts to $ 223 million in over 300000 borrowers in 65 countries.

**3.2. Stakeholders**

A Stakeholder is someone who plays a vital role in a company’s long-term success. Now, the potential stakeholders of KIVA.ORG are lenders, borrowers and MFI’S (Micro financial institutions) and the key roles of stakeholders are providing financial support to the organization and helping with business initiates.

**Lenders** (The people who are raising the money), kiva is a crowdfunding organization as the individuals pool money for a specific project or need. Lenders are called the external stakeholders because they have a financial interest in the success of the project for which it has lent money. Kiva has lenders from over 70 nations covering 5 continents and so is the important aspect in regards with the rate of the organization’s success which is 96%.

**Borrowers** (The people who are pleading money), The ultimate gain or loss effected are those who receives the loan called borrowers. The primary stakeholder for the success of Kiva.org as they are involved in the repayment of the borrowed loans. The borrowers most found are from food, agriculture sectors. The percentage of loans repaid to the lenders is 96.3 from the year 2005 and the one of the major reasons is 0% interest offered and where KIVA funded over $1.68 billion worldwide.

**Microfinance institutions** (The groups who provide the platform), Kiva works with more than 330 local organizations worldwide to distribute loans. These are the lending partners on the ground, meeting borrowers and delivering loans. They allow entrepreneurs and small business owners in poor or rural regions to obtain small amounts of financing that would be difficult to obtain otherwise. Their primary target is less developed countries to promote economic growth, financial inclusion, and prosperity.

**3.2.1. Research objective**

Our main objective is to predict the loan status by determining economic levels of borrowers using kiva loan data set in addition to other economic as well demographic data which includes nation’s multidimensional poverty index (MPI ranging from 0 to 1, with higher values implying greater poverty), percentage contribution of deprivations of each working sector to overall poverty of a country. This helps the field partners to filter the entrepreneurs and post the qualified one’s profile on kiva’s website. It also gains trust in the lenders for their returns and could increase the participation of lenders in lending money for multiple borrowers, eventually leading to the Kiva organization’s growth. This also gives an idea for a borrower by looking into qualified loan profiles so that they can predict their application status before applying.

**3.3 Similar Organization problems:**

**3.3.1 Zidisha :**

One of the microfinance organizations is Zidisha, similar to Kiva whose CEO is Julia Kurnia as mentioned, over 90% of loans in the last month were used to start or expand enterprises, 9% for college tuition or other educational investments, and less than 1% for consumer costs like health care or home improvement.

Zidisha required borrowers to describe to lenders the source of income they will use to repay loans when they are not used for activities that generate money. For instance, part-time employment can help university students pay back their student loans while they are enrolled in classes.

In conventional microfinance programs, loan administrators frequently meet with applicants in person to assess their ability to repay the loan and their cash flow. In the countries of borrowers, Zidisha has no employees or loan officers.

Instead, Zidisha worked to prevent excessive debt by forcing everyone to start with a modest loan and progressively raising credit limits as those modest early loans are successfully repaid.

**3.3.2 Building Resources Across Communities:**

Other organization which is similar to kiva is Building Resources Across Communities (BRAC). It is the biggest non-legislative improvement association on the planet regarding the number of workers. Hasan established BRAC in 1972 and it utilizes in excess of 120,000 individuals in 11 nations. BRAC has a microfinance program, essentially in Bangladesh, which has lent to 5.6 million borrowers, 87 percent of whom are ladies. Dissimilar to Kiva and Zidisha, which work one individual to another loaning administrations, BRAC disperses advances to moneylenders on its own utilizing gifts and different assets. BRAC likewise takes care of business irrelevant to microfinance, putting resources into schools and in water, cleanliness and disinfection administrations.

1.**1. Research Questions**

* + 1. To determine the best fit machine learning model to predict loan funding based on model efficiency and model evaluation results (confusion matrix)
    2. To determine meaningful insights from loan descriptions by using NLP models / text analytics?
    3. What are the top features in loan descriptions that attract funding

1. **1.1.** **Processing Data**

We initially processed all the metadata available in all 3 datasets & related to our objective of project report. It was understandable that loan data has key variables required to build predictive models to determine whether a loan application gets funded or not. With this presumption we started with basic data cleansing, data transformation & data explorations tasks using Python programming on Jupyter notebook as well as visualizing data in Tableau application.

## 1.2. Methods of Analysis

To carry out Machine Learning projects efficiently, it is important to define the tasks to be completed and the roles involved. This defines a structured process that drives the project team towards a well-defined goal and ensures a common understanding of the business needs. Even though many process models can be used for Data Mining projects, these models cannot be effectively applied to Machine Learning projects without adapting and adding tasks. We adapted the well-known and widely accepted Cross-Industry Standard Process for Data Mining (CRISP-DM) to reflect Machine Learning specifics and ensure a successful execution of Machine Learning projects. Also, the participating roles and their responsibilities within the project are defined which is an important element that is missing in CRISP-DM. Since Machine Learning includes a broad spectrum of methods, this article will focus on adapting CRISP-DM to the specifics and requirements of unsupervised and supervised learning, encompassing specific methodologies like neural networks, decision trees and clustering (Elsevier Inc., 2017)

CRISP DM methodology has 6 steps, starting with basic understanding of the business & the need of data mining project and ends with deployment of solution to meet the specific objective of business need. The key to success of any CRISP DM project is to ensure initial steps are strong & correct as lateral steps provide the outcome on data & assumptions considered in earlier steps.

Figure 6.1: Six Step CRISP-DM Data Mining Process

## 1.3. Model Building:

In this step, we shall use different types of predictive models on the transformed data to conduct experiments to determine best suited model based on model efficiency, & confusion matrix results. For each model, we have suitably optimized relevant dependent parameters from loan data set which has 34 variable to obtain best results for our independent variable ( Loan Status : Funded or not funded). We explored 3 Data Mining & Machine Learning Models on our data set as shown in below figure.

* 1. Regression
  2. Decision Tree
  3. Random Forest

We shall explore topic modeling approaches to determine key features from loan description data. Our independent variable in this model would be Topical prevalence is a vector that sums up to one for each individual text or response k=10 topics including key features from loan description data.

**Topic Modeling**:

Natural language is complicated, unclear and full of skewed understanding, and sometimes trying to clean uncertainty, reduces the language to an unnatural form. Topic Models are a type of machine learning models used for discovering hidden features in a collection of texts. By doing topic modeling, we build clusters of words rather than clusters of texts by capturing “topics” that appear in a collection of records that best symbolizes the information in them A text is thus a combination of all the topics, each having a specific weight. In a practical and more instinctively, it is similar to Dimensionality Reduction, Unsupervised clustering, or tagging. Topic models are built based on term co-occurrence where every document is mixture of topics.

*Source: <https://rpubs.com/chelseyhill/672546>*

* Document 1 could be 30% Topic A, 20% Topic B, 40% Topic C and 10% Topic D.
* Every topic is a mixture of terms.
* Terms have weights associated with topics.
* Terms can be shared across topics.
* Each topic is a distribution over the topical *terms*.
* Each document is a mixture of corpus-wide *topics*.
* Each term is drawn from one of the *topics*.

There are several existing algorithms you can use to perform the topic modeling. The most common of it are, Latent Semantic Analysis (LSA/LSI), Probabilistic Latent Semantic Analysis (pLSA), and Latent Dirichlet Allocation (LDA). LDA modeling assumptions include:

* bag of words (BOW): ordering of terms in documents is unimportant
* documents are exchangeable: document sequencing is unimportant
* topics are independent (uncorrelated)

**LITERATURE REVIEW :**

**KIVA MICROFINANCE:**

**INTRODUCTION:**

In every country millions of people are surviving without legitimate financial services which becomes on of the reasons for their poverty. As there is lack of basic financial services from banks like loans of funds transfer there are only few ways to get to better financial stability. An this is where microfinance institutions come into picture which provides short credit lines for people with a plan to serve the community by legitimate business ideas by which in turn they can make a decent living and also repay the same short credit line debts.

Micro finance institutions like KIVA can also provide insurance easily as these institutions customers credit lines are so short. It might not look like there is no need for insurance for such short loans but it does help in getting things back up for there customers and this whole infrastructure can make the customers actual building blocks of the economy of there own community.

In This literature review we are looking at the facts mentioned by ShinYen based on previous studies on the KIVA’s business model. We are also looking at the methods used and applied like which entities were considered to evaluate how new consumers are being selected to grant the financial services based on their requirements.

**BIAS ON RELATED LITERATURE:**

In his paper ShinYen carried out a data cleaning method that removed the loan data that did not have perfect English worded loan description before loading the data.

According to the dataset only 1% of the loans were rejected hence removing the data entries of the rejected loans the model could focus on approved rate accurately.

He also considered the approval speed as fast approval time takes is fast if it is under 3 days which is least time taken by any organization.

**LIMITATIONS OF RELATED LITERATURE:**

ShenYin in his paper used recurrent neural network model and described it as it has given better results when compared to Logistic regression model.

When the weights of the gradients becomes too small the gradients vanishes and significant information is dis regarded in the results.

Here the RNN model is only used because entity used for calculating the loan approval speed is loan description.

**RESEARCH STUDY:**

Recent advancements in the disciplines of text mining and natural language processing have helped text classification (TC), a task of basic importance, gain traction. Text classification techniques all aim to assign a predefined label to a given input text, though this term can apply to a range of specialized techniques used in many fields. As from the article **on Text Classification Algorithms: From Text to Predictions [1],** the standard preprocessingtechniques includetokenization, stop word and noise removal.

After standard preprocessing the model, it is recommended to apply preprocessing deep models, these include models such as Byte Pair Encoding, which is now regarded as the breakthrough method for sub-word tokenization. The most common vocabulary term is combined into a new vocabulary word via this algorithm, which iteratively calculates the occurrences of subsequent pairs of vocabulary terms. The word Piece tokenizer employs a data-driven strategy for breaking words into sub-words. It was originally developed as a solution to Japanese text segmentation issues. It uses n-gram-based language models to identify recurrent prefixes, word fragments, and syllables in corpora. But unlike Word piece, UnigramLM takes the opposite approach, initializing at a vocabulary size that is significantly greater than the intended number of sub-words and then going on to iteratively delete them. The expectation-maximization procedure is applied in each iteration to eliminate the items with the lowest likelihood, cycling until the vocabulary reaches the required size. We apply the data preprocessing techniques suggested in this article.

We then use topic modelling for classification. A topic model is a model that can identify themes in a document based on the words that appear there. We use Latent Dirichlet Allocation [2] technique which is one the method for topic modelling. LDA assumes that each document is made up of a combination of subjects and words, and that each topic is made up of a combination of topics.

Graphical user interface, diagram, application, Teams

Description automatically generated

The above flow chart shows the steps involved in LDA. It uses three different parameters. The number of topics anticipated in the text is controlled by the α hyperparameter. K determines how many topics we need to extract, and the β hyperparameter determines how many words per topic there are in the document. The major limitation while applying this model is one must manually assign the themes to the given topics, which could lead to mistakes. LDA performs poorly on small texts. Since the description column in our dataset has sufficient information, we feel that LDA is better for applying

older is someone who plays a vital role in a company’s long-term success.

* **Research Questions**
* Can we use NLP (Natural Language Processing) to predict Loan status?
* Can grouping the loans and borrowers into clusters by unsupervised learning techniques and identify top n features (K – means) to get the insights on top 3 and bottom 3 borrowers and lenders be possible?

## References

1. <https://www.kiva.org/about/impact>
2. <https://www.gsb.stanford.edu/gsb-box/route-download/349176>
3. <https://github.com/damitkwr/KivaExpiry>
4. <https://www.linkedin.com/pulse/model-built-predict-whether-loan-applicant-default-david-ifeoluwa>
5. <https://hbr.org/2013/12/analytics-30>
6. <https://www.kiva.global/global-impact-dashboard/>
7. <https://www.youtube.com/watch?v=om02_lChPHc>

1. <https://www.kiva.global/wp-content/uploads/2022/06/AnnualReport_2021.pdf>
2. <https://www.sas.com/en_us/software/on-demand-for-academics.html>
3. <https://www.tableau.com/products/desktop>
4. https://www.pearson.com/en-us/subject-catalog/p/business-intelligence-analytics-and-data-science-a-managerial-perspective/P200000001822/9780137305711
5. <https://www.mdpi.com/2078-2489/13/2/83/htm>
6. <http://www.brac.net>
7. <https://www.zidisha.org>
8. <https://borgenproject.org/4-microlending-organizations-that-empower-the-poor/>
9. [1]  *Andrea, G; Matteo, M; Alessandro, Z; Andrea, A.* A Survey on Text Classification Algorithms: From Text to Predictions **2022**

https://www.mdpi.com/2078-2489/13/2/83/htm

1. [2] Topic Modelling using LDA

<https://www.mdpi.com/2078-2489/13/2/83/html>

1. *Understanding pro-social landing: Prediction of ... - stanford university*. (n.d.). Retrieved November 1, 2022, from <https://cs224d.stanford.edu/reports/ShenYin.pdf>