|  |
| --- |
| **KIVA.ORG** |
| **Analytics 3.0: Examine Machine Learning applications on crowd empowered microfinance lending platform** |
|  |

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

Table of Contents

[1. Abstract: 1](#_Toc119185061)

[2. Introduction 2](#_Toc119185062)

[2.1. Project Objective 2](#_Toc119185063)

[2.2. Research Questions 2](#_Toc119185064)

[2.3. Kiva.org 3](#_Toc119185065)

[2.4. Crowdfunding 3](#_Toc119185066)

[2.5. Loan process 4](#_Toc119185067)

[2.6. Stakeholders 5](#_Toc119185068)

[3. Literature review 5](#_Toc119185069)

[3.1. Related work on Kiva data 5](#_Toc119185070)

[3.2. Machine learning applications on Crowdfunding platforms 6](#_Toc119185071)

[3.3. Making Sense of social media using Machine Learning Models 7](#_Toc119185072)

[3.4. Topic Modeling as a solution to deal with unstructured text data 7](#_Toc119185073)

[4. Methodology 8](#_Toc119185074)

[4.1. Capturing Data 8](#_Toc119185075)

[The Loans 8](#_Toc119185076)

[The Lenders 9](#_Toc119185077)

[The Field/lending Partners 10](#_Toc119185078)

[4.2. Processing Data 10](#_Toc119185079)

[4.3. Methods of Analysis 10](#_Toc119185080)

[4.4. Model Building: 11](#_Toc119185081)

[4.5. Data Transformation & Analysis 13](#_Toc119185082)

[4.6. Data Exploration using Tableau data visualization tool 13](#_Toc119185083)

[4.7. Model Analysis & Validation 16](#_Toc119185084)

[5. Appendices 18](#_Toc119185085)

[6. References 21](#_Toc119185086)

## Abstract:

Micro lending through crowdfunding is the technique of generating investment to fund a project or loan through several investors using an innovative digital platform. *Kiva.org*, an extensively treated crowd-funded micro-financial service, offers students & researchers with an extensive volume of publicly available data consist of a rich set of diverse 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 to build a most appropriate machine learning model to predict whether loan application gets funded or not and examine the application of topic modeling techniques to loan descriptions to identify key features from loan description that attracts lenders on a crowd funding platform like kiva.org. We also tried descriptive and diagnostic analysis using data visualization tool – Tableau to identify factors that are behind Kiva’s success story.

With the implementation of five classification models, we determined that neural network models are best model with an efficiency of 97.13% to predict loan funding. Also, loan application with key features such as sell, family, farming, earning, improve, grow etc. are some of the top thirty dominant words helping lenders to fund the loan application.

**Key Words: Loan, Crowdfunding, Topic Modeling, NLP, Latent Dirichlet Allocation, Clustering, Confusion matrix**

## Introduction

## Project Objective

The goal of this project is to effectively discover common topics among a large data set of loan applications from 2005 to 2021 of P2P crowd funding platform. Each loan story/description will be assigned to some number of topics, and the specific segments of the description which address a given topic will hopefully be specified as well. Thus, not only will the description be classified as covering some set of topics, but the key words themselves will be partitioned into different sub-topics. This work would be useful in identifying relationship between borrowers’ story and lenders interest in funding her/his loan application, may be highly relevant due to a shared liking or sector. Thus, improving the loan funding time & happy borrowers. Additionally, we also explore machine learning models to identify best fit model to predict weather a loan application gets funded or not based on the key features in the loan application.

## Research Questions

## Which is the best fit machine learning model to predict loan funding based on model efficiency?

* + 1. What are the meaningful insights from loan descriptions by using NLP models / text analytics?
    2. What are the top features in loan descriptions that attracts funding?

Founded in 2005, Kiva.org is one of the world’s major online crowdfunding platforms where people can lend money to underserved entrepreneurs across the world. With the mission of “connecting people through lending to alleviate poverty,” ([www.kiva.org](http://www.kiva.org)) Kiva ventures to deliver inexpensive access to financing in areas where traditional banking services is futile to meet the requirement, thereby empowering people to create opportunities to become entrepreneurs (Melina Moleskis-2016). In fact, Kiva acts like *UBER*, an online connection between borrowers and lenders. The profiles of borrowers from various 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 seventy countries and work 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) refugee and c) gender. Till date, the kiva is able to support more than 4.4 million people by lending $1.77 billion, at a repayment rate of 96.3%. (Kiva.org .2022)

A picture containing logo

Description automatically generatedLogo, company name

Description automatically generated

*Source:* [*www.kiva.org/about*](http://www.kiva.org/about)*us*

## Crowdfunding

Crowdfunding is the system of raising money to fund a venture or business 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 states that, 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.

## Loan process

The process begins when a barrower 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 gets 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 begin with a minimum capital amount of $25 to any amount and a loan is funded by more than a few lenders. The lenders then transmit 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 accumulates loan investment from the individual lenders and transfers it to the suitable Field Partners, which distribute 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.

## 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 seventy 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 $1.77 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.

## Literature review

## Related work on Kiva data

Microfinance is a promising, if imperfect, solution to alleviating poverty. It can be difficult to scale, and there is a risk of putting vulnerable populations on a kind of debt “treadmill.” However, historical success of microfinancing, coupled with modern innovation, allow for an optimistic outlook in this socially responsible lending space. Microfinance is not a novel idea, with its roots dating back to the 1800’s, when Lysander Spooner theorized an alleviation of poverty by dispensing microcredits to entrepreneurs and farmers (Garrity & Martin, 2018). It came into focus in 1976 when Muhammed Yunus, a well-known economics professor, felt compelled to put theory to practice after studying how basket weavers were forced to take usurious loans trapping them in a cycle of poverty. With a simple loan of $27 Yunus was able to pay off the obligations of forty-two people that were otherwise unable to rise above their means. From these meager beginnings, the idea of Grameen Bank began smashing the notion that the poorest of the poor were not creditworthy (Gajjala et al., 2011). Yunus had a vision for poverty alleviation that focused on self-help instead of direct income redistribution (Cull et al., 2009).

Several prior works have also shed light on Kiva’s online marketplace. Ly and Mason (2012) studied & described how loan parameters impacts the speed of the loans by building regression models on some of its quantitative characteristics. They took log transformation of the funding time and performed ordinary least square regression. Results produced a model with an R2 of 0.42 and indicated that loan size describes most of the variance. They succeeded to discover several other small but very significant coefficient estimates. However, they have put too many parameters in their OLS equation and hence overfitting. Liu et al. (2012) has studied the supply side of the market using NLP methods. They classified the lenders’ self-stated rationales into ten categories with human coders (who will create the true labels) and machine learning based classifiers. They used text classifiers using lexical features, with social features based on lender activity data on Kiva, to predict the categories of lender rationale. With these the results, they predicted lending activity from lender motivation and team affiliations. Only the category variable produced out of the NLP classifiers was selected toward their final linear regression formula while the potentially informative texts and phrases are discarded.

## Machine learning applications on Crowdfunding platforms

Lately, natural language processing (NLP) has developed as powerful methods for many industries & academia fields due to its ability to identify sentiments and feeling into the text in more effectively. Many industries have started adopting the NLP techniques to give their users better experience (Xing et al., 2018). Though it easy to get the other tabular data with help of many sources but it is tough to access and parse the data through unstructured, multi-lingual having different special characters. So, creating information content from large sentences or stories is tedious and difficult.

There are numerous studies on crowdfunding that target to predict different trends and project success (Ala’raj et al.2021). Some studies have discovered various semantic features, specific patterns, writing styles in loan applications to reveal the impact of language on the success or failure of a loan application on P2P platform. Crowdfunding success prediction is estimated over a text analytics method, where LDA is used to extract semantic features from the texts, along with feature selection, and data mining (Tripati 2021). In a similar work (Yuan.2016) crowdfunding data are analyzed by using LDA to classify into distinct topic classes.

Topic models such as the latent Dirichlet allocation (LDA) have become a well-known tool in machine learning (Blei et al., 2003). They have played an important role in a variety of data mining tasks, both within the scope of computer science (Blei et al., 2003; Liu et al., 2012;) and reaching out to other fields, such healthcare, political science & digital space.

To capture the semantic features of the text, topic models play a vital role. The semantic features are extracted as hidden topics. There have been numerous studies and applications of topic models since this idea was first introduced by Blei et al. in 2003. Then on several studies have conducted many research areas varying from scientific research to mathematical equations. Recently, much research is being done with a combination of multiple deep learning models along with topic models. Regardless of the high advances in machine learning & deep learning models, there are still some challenges in terms of the data cleansing, data size, time to run models with a standalone system specification. Despite having some successes, certain properties of the LDA have become part of the machine learning myth, but never been formally proven; Having said that, project objective is to identify key word patterns in loan applications to determine any relationship between those semantic features attracting volunteers to lend money for their requirements. Though kiva.org loan approval team conduct due diligence on the requirements such as loan term, repayment terms as well as risk assessment, importance of key word features should not be ignored

## Making Sense of social media using Machine Learning Models

Among the several types of communication means, social media stands out as an important medium in facilitating interactions between social actors (Murthy D 2012). To understand the enormous numbers of posts shared on social media, NLP can understand human languages, as programmed for machines, to make predictions based on the observed social phenomena (Hannigan et al., 2019). On the other hand, machine learning models refers to computational methods using the train data to build and train a model for prediction and superior decision making.

## Topic Modeling as a solution to deal with unstructured text data

As our language is multifaceted in its text length, syntactic intricacy, and semantic authenticity have long been deemed pivotal points in both psychology and linguistics (Bradley and Meeds, [2002](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9120935/#B12)). We are witnessing its multifold impact in social media with the use of latest communication applications such as Facebook. Twitter, LinkedIn etc. For instance, scholars have indicated that shorter posts usually lead to a higher engagement rate on Facebook (Sabate et al., [2014](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9120935/#B62)), possibly for the reason that short messages lower the amount of intellectual energy needed for information processing (She et al., [2022](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9120935/#B65)). Across the various available types of platforms, Twitter, in specific, limits each post to a maximum of 280 characters (Queiroz, [2018](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9120935/#B59)), and even though these short and unstructured posts fit with social media way, they enhance the difficulty for algorithms to make point of digital communication. Common tasks arise from using composite words, acronyms, and vague sentences (Ariffin and Tiun, [2020](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9120935/#B8)).

Notwithstanding the reputation of LDA within the social science section, its effectiveness in evaluating social media information has been limited (Egger and Yu, [2021](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9120935/#B22); Sánchez-Franco and Rey-Moreno, [2022](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9120935/#B63)). In the case of Twitter data, Jaradat and Matskin ([2019](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9120935/#B36)) say that, while compound topics can coexist in a document, LDA have a tendency to neglect co-occurrence relationships. Similarly, other researchers underline that inaccurate and sparse datasets are inappropriate for LDA (Chen et al., [2019](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9120935/#B17)) due to a lack of features for statistical learning (Cai et al., [2018](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9120935/#B13))

## Methodology

## Capturing Data

Kiva data available by way of snapshots which are compressed into a simple singular download on their website. As part of Kiva strategy, the archived data is specifically shared for researchers to use the data & help kiva.org to improve their services. The latest data snapshots are available in the format of your choice at the following URLs:

* JSON: http://s3.kiva.org/snapshots/kiva\_ds\_json.zip
* CSV: http://s3.kiva.org/snapshots/kiva\_ds\_csv.zip

However, due to modernization of authentication procedure to access secure content, and we could be able to download data in CSV format amounting to ~ 4.5GB as of Nov-2021. For our project work we will be using loans dataset which has 2,187,118 loan listings with thirty-four variables.

## 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 application 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. |

## The Lenders

Lenders has 157879 listings with fourteen variable 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 chean, plus a number, as in chean2749, 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 s/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 seventy countries on 5 continents. 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, the also provide entrepreneurial training, literacy skills, lending quality seed and farming inputs, and providing access to savings accounts and insurance. The motto of kiva with these lending partners to reach needy people across the globe by enhancing their lives through safe and fair access to credit.

## Processing Data

We initially processed all the metadata available in all three datasets & corelated 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.

## Methods of Analysis

To conduct 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 collective 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) (F. Martínez-Plumed et al., 2021) 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 essential 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

CRISP DM methodology has six 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 provides the outcome on data & assumptions considered in earlier steps.

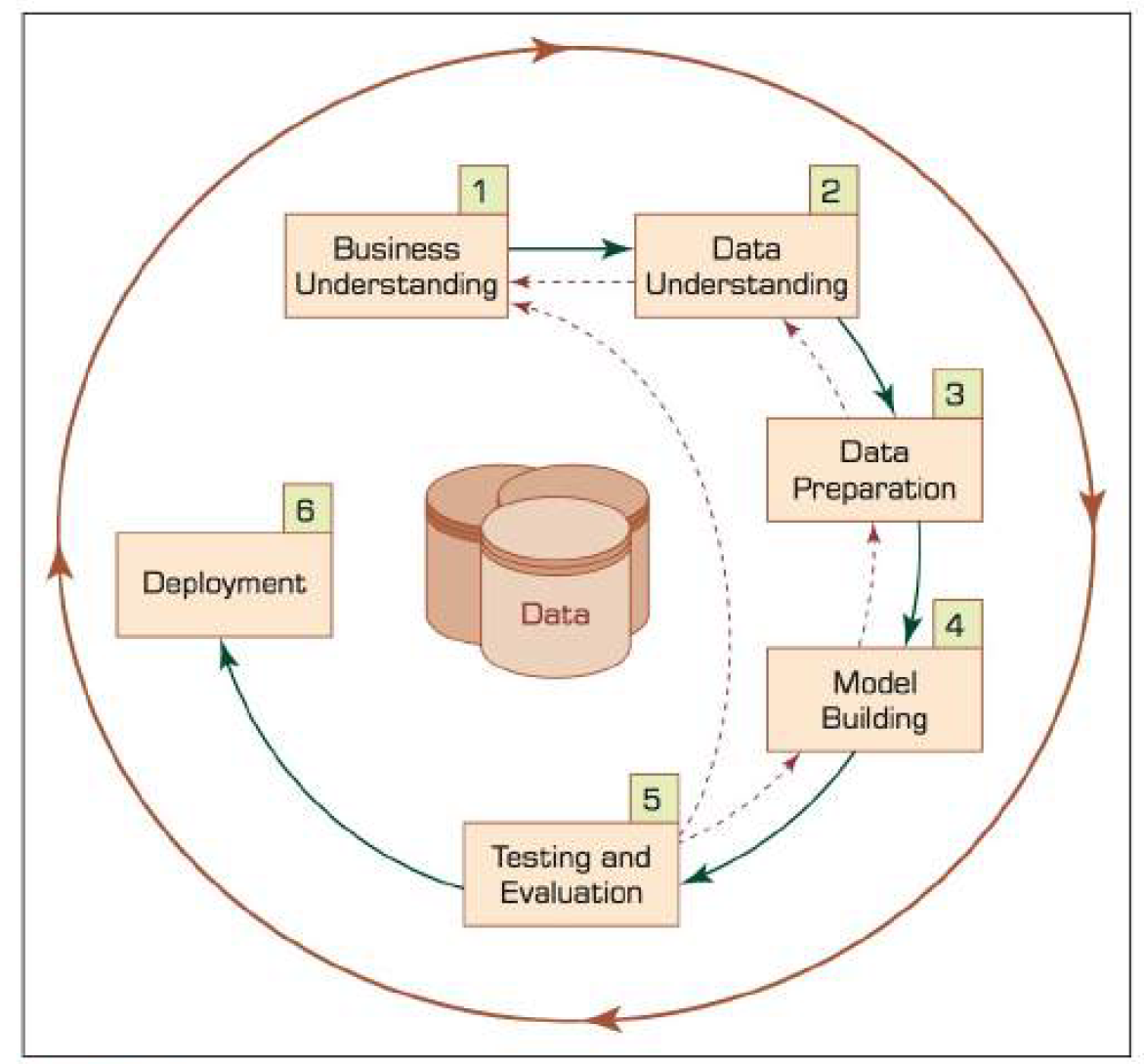


Figure 6.1: Six Step CRISP-DM Data Mining Process

## Model Building:

In this step, we shall use several 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 thirty-four variables 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
  4. Naïve bayes
  5. Neural Network

We shall explore topic modeling approaches to determine key features from loan description data. Our dependent 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 (Patrick Grafe .2010) 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 like Dimensionality Reduction, Unsupervised clustering, or tagging. Topic models are built based on term co-occurrence where every document is mixture of topics.

Diagram

Description automatically generated

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

* Document one 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) (Roman Egger 1 and Joanne Yu .2022). 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)

For tpoic modelling, we used loan description & loan description tarnsalated variale from loan data set to determine top ‘N’ words & topics using Latent Dirichlet Allocation (LDA) method in the python using gensim implementation. The model implementation includes,

1. Loading Data
2. Data Cleaning
3. Phrase Modeling: Bi-grams and Tri-grams
4. Data Transformation: Corpus and Dictionary
5. Base Model
6. Hyper-parameter Tuning
7. Final model
8. Visualize Results

## Data Transformation & Analysis

We used box plot diagram to for Loan amount & Funded amount to identify data distribution & found high variation in data set & has outliers. We removed the outliers in loan amount variable to get more distribution. The data shows high variation & hence we used mathematical transformation to normalize the variation. (Log Transformation)

Table

Description automatically generated with low confidence Chart, box and whisker chart

Description automatically generated

Addition transformation was carried data structure for Image\_ID into binary variable by assing '1'and'0' for image availability Yes or No. We read the policies on Kiva about exchange coverage. For instance, a lender gives twenty-five in US dollars, and it changes to 250 in another currency. But when the money is repaid at 250 in the other currency, it only exchanges to 20 US dollars. The URRENCY\_EXCHANGE\_COVERAGE\_RATE measures how much such loss will be covered by the field partners. Some would cover 10% or 20%. So, it is more likely that the missing values indicating the loan does not have any coverage. So, it is more likely means 0, hence replaced the missing value with 0. Finally, dropping variables which are not required. By understanding the data variables for example there are many more activities than sectors and Sector is higher-level categorization of loans. We choose one of them. We also conducting topic modelling on how to process 'DESCRIPTION' and 'DESCRIPTION\_TRANSLATED' to identify key topics that helps loan funding.

## Data Exploration using Tableau data visualization tool

**Loans by region**:

Asia has the maximum borrowers (more than one third) & rest are in African countries & other nations. Philippines has been the top one 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.

Graphical user interface, text

Description automatically generatedChart, waterfall chart

Description automatically generated

Word cloud by country Loans by Sector & Country

**Loans by Sector:**

The data set includes 2,187,118 loan submissions, grouped into fifteen 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 set & 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 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 three hundred thousand borrowers in 65 countries

## Model Analysis & Validation

The key terminologies used to assess the model efficacy are

**Accuracy** is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

accuracy=correct predictions/all predictions

**Precision** is defined as the fraction of relevant examples (true positives) among all the examples which were predicted to belong in a certain class.

precision=true positives/ (true positives+false positives)

**Sensitivity/Recall** is defined as the fraction of examples which were predicted to belong to a class with respect to all the examples that truly belong in the class.

sensitivity=true positives/ (true positives+false negatives)

**Specificity** is the proportion of samples that test negative using the test in question that are genuinely negative

true negatives/ (true negatives+false positives)

By Comparing Model Accuracies of 5 Models, we can conclude that neural network has better accuracy 97.13%, compared to Random Forest model 97.1 % compared to Logistic regression 95.9% & Decision tree 93.7% compared to Naive Bayes Classifiers 93.7% to predict loan funding status for an application.

Though all models resulted in acceptable efficiency, we chose neural network for the key reason of a neural network is to determine the relationship between features in a data set i.e loan application variables.

**NLP – Topic Modeliing :**

Our intent of this analysis is to accomplish use of NLP topic modeling; hence we used the text data from each loan description, language, loan ID and drop other metadata columns. We performed a simple preprocessing on the content of loan description column to make them more reliable by using a regular expression to remove any punctuation, and then lowercase the text, to tokenize each word by removing special characters. The next step was to remove stop words from the list to make more concise. We used python library spacy & genism to build topic model & to identify top n words. With the LDA model results are profoundly impressive to see top thirty most relevant words in each topic & as well all the topics (term frequency & overall frequency). The model shows that topic\_1, topic\_2 & topic\_3 whereas topic\_4 & topic\_5 are the farthest topics as shown in intertopic distance map.



Chart

Description automatically generated

Outcome topic modelling with relevance to each record or loan application provides a clear relationship between key features used in that application vis-à-vis dominant topic & its topic density as shown below. With this measurement, we classify each loan application into different topics/clusters which can be named into an entity

Graphical user interface, text

Description automatically generated with medium confidence

Now that we have the baseline coherence score for the default LDA model, we can assess the series of coherence score, selected to help determine the following model hyperparameters:

* Number of Topics (K),
* Dirichlet hyperparameter alpha: Document-Topic Density,
* Dirichlet hyperparameter beta: Word-Topic Density

Hyper-parameter tuning Results: with the model parameters such as the coherence score, C\_v, for the number of topics across two validation sets, and a fixed alpha = 0.01 and beta = 0.91, we could be able to determine the optimal number of clusters with highest coherence score as shown in below chart. Thus, optimal number of topics would be eleven with coherence score of 0.425.

The hyper parameter tuning results:

Chart, line chart

Description automatically generated

**Discussion on results:**

**Model needs to be fine tuned to remove inappropriate words & to run the model with higher number records to get better coherence score.**

## Appendices

Data exploration: Histogram on loan amount data

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generatedUnderstanding the relationship between key interval variables to loan amount through pair graphs

Chart, scatter chart

Description automatically generated

Histogram on Loans per sector:

Chart

Description automatically generated

Loan application timeseries over the years:

Chart, line chart

Description automatically generated

**Model evaluation results: Confusion Matrix**

Confusion matrix :Logistic Regression Confusion matrix :Random Forest

Chart

Description automatically generated Chart

Description automatically generated

Confusion matrix :Decision Tree Confusion matrix :Naïve Bayes

Chart

Description automatically generated A picture containing chart

Description automatically generated

**Result of topic modeling with key features & its weights**

[(Topic\_1,

'0.028\*"loan" + 0.019\*"group" + 0.017\*"farm" + 0.016\*"child" + 0.016\*"year" ' '+ 0.016\*"farmer" + 0.013\*"family" + 0.012\*"farming" + 0.011\*"profit" + ' '0.009\*"income"'),

(Topic\_2,

'0.055\*"business" + 0.037\*"loan" + 0.036\*"year" + 0.033\*"child" + ' '0.028\*"php" + 0.024\*"nwtf" + 0.022\*"request" + 0.020\*"philippine" + ' '0.019\*"sell" + 0.018\*"future"'),

(Topic\_3,

'0.055\*"business" + 0.033\*"loan" + 0.030\*"year" + 0.025\*"child" + ' '0.024\*"sell" + 0.016\*"buy" + 0.015\*"old" + 0.013\*"family" + 0.013\*"school" + 0.012\*"marry"'),

(Topic\_4,

'0.024\*"loan" + 0.022\*"work" + 0.019\*"year" + 0.017\*"business" + 0.012\*"buy" ' '+ 0.012\*"sell" + 0.012\*"family" + 0.011\*"child" + 0.011\*"old" + ' '0.010\*"make"'),

(Topic\_5,

'0.026\*"loan" + 0.023\*"family" + 0.018\*"income" + 0.017\*"year" + ' '0.013\*"child" + 0.012\*"old" + 0.011\*"live" + 0.011\*"work" + 0.011\*"dear" + ' '0.010\*"husband"')]

## References

1. Ala’raj et al. J Big Data (2021) 8:69. Modelling customers credit card behaviour using bidirectional LSTM neural networks <https://doi.org/10.1186/s40537-021-00461-7>
2. Annual Report Kiva.org 2021 <https://www.kiva.global/wp-content/uploads/2022/06/AnnualReport_2021.pdf>
3. Ariffin S. N. A. N., Tiun S. (2020). Rule-based text normalization for Malay Social Media Texts. *Int. J. Adv. Comput. Sci. Appl.* 11:21. 10.14569/IJACSA.2020.0111021 [[CrossRef](https://doi.org/10.14569%2FIJACSA.2020.0111021" \t "_blank)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Int.+J.+Adv.+Comput.+Sci.+Appl.&title=Rule-based+text+normalization+for+Malay+Social+Media+Texts&author=S.+N.+A.+N.+Ariffin&author=S.+Tiun&volume=11&publication_year=2020&pages=21&doi=10.14569/IJACSA.2020.0111021&)]
4. Blair S. J., Bi Y., Mulvenna M. D. (2020). Aggregated topic models for increasing social media topic coherence. *Appl. Intellig*. 50, 138–156. 10.1007/s10489-019-01438-z [[CrossRef](https://doi.org/10.1007%2Fs10489-019-01438-z" \t "_blank)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Appl.+Intellig&title=Aggregated+topic+models+for+increasing+social+media+topic+coherence&author=S.+J.+Blair&author=Y.+Bi&author=M.+D.+Mulvenna&volume=50&publication_year=2020&pages=138-156&doi=10.1007/s10489-019-01438-z&)]
5. Blei, D.M.; Ng, A.Y.; Jordan, (2003). Latent dirichlet allocation. J. Mach. Learn. Res. 2003, 3, 993–1022.
6. Bradley S. D., Meeds R. (2002). Surface-structure transformations and advertising slogans: the case for moderate syntactic complexity. Psychol. Market. 19, 595–619. 10.1002/mar.10027
7. Cai G., Sun F., Sha Y. (2018). Interactive Visualization for Topic Model Curation. Tokyo: IUI Workshops.
8. Chen Y., Zhang H., Liu R., Ye Z., Lin J. (2019). Experimental explorations on short text topic mining between LDA and NMF based Schemes. *Knowl. Based Syst*. 163, 1–13. 10.1016/j.knosys.2018.08.011 [[CrossRef](https://doi.org/10.1016%2Fj.knosys.2018.08.011" \t "_blank)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Knowl.+Based+Syst&title=Experimental+explorations+on+short+text+topic+mining+between+LDA+and+NMF+based+Schemes&author=Y.+Chen&author=H.+Zhang&author=R.+Liu&author=Z.+Ye&author=J.+Lin&volume=163&publication_year=2019&pages=1-13&doi=10.1016/j.knosys.2018.08.011&)]
9. Crunchbase (2021) Crowdfunding https://www.crunchbase.com/hub/crowdfunding-companies
10. Cull, R., Demirgüç-Kunt, A., & Morduch, J. (2009). Microfinance meets the market. In T. A., Watkins, & K. Hicks (Eds.), Moving Beyond Storytelling: Emerging Research in Microfinance (Contemporary Studies in Economic and Financial Analysis, Vol. 92), pp. 1-30, Emerald Group Publishing Limited, Bingley.
11. Egger R. (2022b). “Topic modelling. Modelling hidden semantic structures in textual data,” in *Applied Data Science in Tourism. Interdisciplinary Approaches, Methodologies and Applications*, ed R. Egger (Berlin: Springer; ), 18. 10.1007/978-3-030-88389-8\_18 [[CrossRef](https://doi.org/10.1007%2F978-3-030-88389-8_18" \t "_blank)] [[Google Scholar](https://scholar.google.com/scholar_lookup?title=Applied+Data+Science+in+Tourism.+Interdisciplinary+Approaches,+Methodologies+and+Applications&author=R+Egger&publication_year=2022b&)]
12. F. Martínez-Plumed et al., 2021 "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," in IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 8, pp. 3048-3061, 1 Aug. 2021, doi: 10.1109/TKDE.2019.2962680.
13. Gajjala, V., Gajjala, R., Birzescu, A., & Anarbaeva, S. (2011). Microfinance in online space: A visual analysis of Kiva.org. Development in Practice, 21(6), 880–893. <https://doi.org/10.1080/09614524.2011.583637>
14. Garrity, P., & Martin, C. (2018). Developing a microfinance model to break the cycle of poverty. Business Horizons, 61(6), 937–947. <https://doi.org/10.1016/j.bushor.2018.07.002>
15. Hannigan T. R., Haans R. F. J., Vakili K., Tchalian H., Glaser V. L., Wang M. S., et al. (2019). Topic modeling in management research: rendering new theory from textual data. *Acad. Manag. Ann*. 13, 586–632. 10.5465/annals.2017.0099 [[CrossRef](https://doi.org/10.5465%2Fannals.2017.0099" \t "_blank)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Acad.+Manag.+Ann&title=Topic+modeling+in+management+research:+rendering+new+theory+from+textual+data&author=T.+R.+Hannigan&author=R.+F.+J.+Haans&author=K.+Vakili&author=H.+Tchalian&author=V.+L.+Glaser&volume=13&publication_year=2019&pages=586-632&doi=10.5465/annals.2017.0099&)]
16. Hsieh H-I, Lee T-P, Lee T-S. (2010) Data mining in building behavioral scoring models. 2010. https://doi.org/10.1109/cise.2010. 5677005.
17. I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal (2017) [Data Mining-Practical Machine Learning Tools and Techniques](https://www.elsevier.com/books/data-mining/witten/978-0-12-804291-5) (Elsevier Inc., 2017)
18. Jaradat S., Matskin M. (2019). “On dynamic topic models for mining social media,” in *Lecture Notes in Social Networks. Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*, eds N. Agarwal, N. Dokoohaki, and S. Tokdemir (Berlin: Springer;), 209–230. 10.1007/978-3-319-94105-9\_8 [[CrossRef](https://doi.org/10.1007%2F978-3-319-94105-9_8" \t "_blank)] [[Google Scholar](https://scholar.google.com/scholar_lookup?title=Lecture+Notes+in+Social+Networks.+Emerging+Research+Challenges+and+Opportunities+in+Computational+Social+Network+Analysis+and+Mining&author=S.+Jaradat&author=M.+Matskin&publication_year=2019&)]
19. Jian Tang, (2014) Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis PMLR 32(1):190-198, 2014
20. Kiva.org (2022) Global Impact Dashboard <https://www.kiva.global/global-impact-dashboard>[/](https://www.kiva.global/global-impact-dashboard/)
21. Liu et al. (2012).” I Loan Because...”: Understanding Motivations for Pro-Social Lending. WSDM’12.
22. Ly P. and Mason G. (2012). Competition Between Microfinance NGOs: Evidence from Kiva in World Development, 2012, 40(3):643-655 https://doi.org/10.1016/world.dev.2011.09.009
23. Melina Moleskis-(2016) CROWDFUNDING SUCCESS: THE CASE OF KIVA.ORG <https://media.iese.edu/research/pdfs/WP-1137-E.pdf>
24. Murthy D. (2012). Towards a sociological understanding of social media: theorizing twitter. *Sociology* 46, 1059–1073 10.1177/0038038511422553 [[CrossRef](https://doi.org/10.1177%2F0038038511422553" \t "_blank)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Sociology&title=Towards+a+sociological+understanding+of+social+media:+theorizing+twitter&author=D+Murthy&volume=46&publication_year=2012&pages=1059-1073&doi=10.1177/0038038511422553&)]
25. Parhankangas, A.; Renko, (2017) M. Linguistic style and crowdfunding success among social and commercial entrepreneurs. J. Bus. Ventur. 2017, 32, 215–236. [CrossRef]
26. Patrick Grafe (2010) Topic Modeling in Financial Documents <https://cs229.stanford.edu/proj2010/Grafe-opicModelingInFinancialDocuments.pdf>
27. Queiroz M. M. (2018). A framework based on Twitter and big data analytics to enhance sustainability performance. *Environ. Qual. Manag*. 28, 95–100. 10.1002/tqem.21576 [[CrossRef](https://doi.org/10.1002%2Ftqem.21576" \t "_blank)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Environ.+Qual.+Manag&title=A+framework+based+on+Twitter+and+big+data+analytics+to+enhance+sustainability+performance&author=M.+M+Queiroz&volume=28&publication_year=2018&pages=95-100&doi=10.1002/tqem.21576&)]
28. [Roman Egger](https://pubmed.ncbi.nlm.nih.gov/?term=Egger%20R%5BAuthor%5D) 1 and [Joanne Yu](https://pubmed.ncbi.nlm.nih.gov/?term=Yu%20J%5BAuthor%5D) (2022) A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts doi: [10.3389/fsoc.2022.886498](https://doi.org/10.3389%2Ffsoc.2022.886498)
29. Sabate F., Berbegal-Mirabent J., Cañabate A., Lebherz P. R. (2014). Factors influencing popularity of branded content in Facebook fan pages. *Eur. Manag. J*. 32, 1001–1011. 10.1016/j.emj.2014.05.001 [[CrossRef](https://doi.org/10.1016%2Fj.emj.2014.05.001" \t "_blank)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Eur.+Manag.+J&title=Factors+influencing+popularity+of+branded+content+in+Facebook+fan+pages&author=F.+Sabate&author=J.+Berbegal-Mirabent&author=A.+Ca%C3%B1abate&author=P.+R.+Lebherz&volume=32&publication_year=2014&pages=1001-1011&doi=10.1016/j.emj.2014.05.001&)]
30. Sánchez-Franco M. J., Rey-Moreno M. (2022). Do travelers' reviews depend on the destination? An analysis in coastal and urban peer-to-peer lodgings. Psychol. Market. 39, 441–459. 10.1002/mar.21608 [[CrossRef](https://doi.org/10.1002%2Fmar.21608" \t "_blank)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Psychol.+Market&title=Do+travelers%27+reviews+depend+on+the+destination?+An+analysis+in+coastal+and+urban+peer-to-peer+lodgings&author=M.+J.+S%C3%A1nchez-Franco&author=M.+Rey-Moreno&volume=39&publication_year=2022&pages=441-459&doi=10.1002/mar.21608&)]
31. Shafqat W., Byun Y.-C. (2020). A recommendation mechanism for under-emphasized tourist spots using topic modeling and sentiment analysis. *Sustainability* 12:320. 10.3390/su12010320 [[CrossRef](https://doi.org/10.3390%2Fsu12010320" \t "_blank)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Sustainability&title=A+recommendation+mechanism+for+under-emphasized+tourist+spots+using+topic+modeling+and+sentiment+analysis&author=W.+Shafqat&author=Y.-C.+Byun&volume=12&publication_year=2020&pages=320&doi=10.3390/su12010320&)]
32. She J., Zhang T., Chen Q., Zhang J., Fan W., Wang H., et al. (2022). Which social media posts generate the most buzz? Evidence from WeChat. *Internet Res*. 32, 273–291. 10.1108/INTR-12-2019-0534 [[CrossRef](https://doi.org/10.1108%2FINTR-12-2019-0534" \t "_blank)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Internet+Res&title=Which+social+media+posts+generate+the+most+buzz?+Evidence+from+WeChat&author=J.+She&author=T.+Zhang&author=Q.+Chen&author=J.+Zhang&author=W.+Fan&volume=32&publication_year=2022&pages=273-291&doi=10.1108/INTR-12-2019-0534&)]
33. [Thomas H. Davenport](https://hbr.org/search?term=thomas%20h.%20davenport) (2013) Analytics 3.0 <https://hbr.org/2013/12/analytics-30>
34. Tripathi D, Edla DR, Bablani A, Shukla AK, Reddy BR. (2021) Experimental analysis of machine learning methods for credit score classifcation. Prog Artif Intell. 2021; 15:1–27.
35. Xing, F. Z., Cambria, E., & Welsch, R. E. (2018). Natural language based financial forecasting: A survey. Artificial Intelligence Review, 50(1), 49–73. <https://doi.org/10.1007/s10462-017-9588-9>.
36. Yuan, H.; Lau, R.Y.; Xu, W. (2016) The determinants of crowdfunding success: A semantic text analytics approach. Decis. Support Syst. 2016, 91, 67–76. [CrossRef]
37. Zhou L., Pan S., Wang J., Vasilakos A. V. (2017). Machine learning on big data: opportunities and challenges. *Neurocomputing* 237, 350–361. 10.1016/j.neucom.2017.01.026 [[CrossRef](https://doi.org/10.1016%2Fj.neucom.2017.01.026" \t "_blank)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Neurocomputing&title=Machine+learning+on+big+data:+opportunities+and+challenges&author=L.+Zhou&author=S.+Pan&author=J.+Wang&author=A.+V.+Vasilakos&volume=237&publication_year=2017&pages=350-361&doi=10.1016/j.neucom.2017.01.026&)]
38. Zi Yin, Yuanyuan Shen (2014). Understanding pro-social landing: prediction of funding time using loan descriptions on Kiva <https://cs224d.stanford.edu/reports/ShenYin.pdf>