

Identifying the Impact of Tweets on Kickstarter Campaign's Funding Using Sentiment Analysis and Machine Learning Approach

MSc Research Project
Data Analytics

Nilesh Ramesh Bukane
Student ID: x01628517

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
MSc Project Submission Sheet



School of Computing

Student Name:	Nilesh Ramesh Bukane
Student ID:	x01628517
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Lecturer:	Dr. Catherine Mulwa
Submission Due Date:	12/08/2019
Project Title:	Identifying the Impact of Tweets on Kickstarter Campaign's Funding Using Sentiment Analysis and Machine Learning Approach.
Word Count:	8767

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	12 th August 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	<input type="checkbox"/>
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	<input type="checkbox"/>
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	<input type="checkbox"/>

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Identifying the Impact of Tweets on Kickstarter Campaign's Funding Using Sentiment Analysis and Machine Learning Approach

Nilesh Ramesh Bukane
x01628517

Abstract

Kickstarter is the popular crowdsourcing platform, which is used for pitching new ideas and products and campaigning for the same. The campaigners usually invest their significant amount of money, time and resources for pitching their products and for putting things in operations. If the information about the project's successful funding's is available to campaigners during the campaign, it will help them significantly to make the necessary arrangements and take crucial decisions about their products being campaigned. This will help them to save lot of their money and resources. Also, these predictions will allow potential funders to take decisions wisely related to their investments. The tweets on social media such as Twitter are the form of people's opinions and thoughts and has the potential of making an impact. Therefore, in this research, the impact of sentiments within the tweets on project funding is been identified. For this analysis, the lexicon-based sentiment classification technique was implemented for classifying the tweets. After classification, it was identified that there is a significant association between the high funding's of the project and the positive tweets, and between the low funding's of the project and the negative tweets. Based on these findings, the machine learning models were trained using SVN, Random Forest, KNN, Naïve Bayes and Decision Tree algorithms. These trained models can be used to classify the textual information into sentiments and may further help in predicting whether the particular project has a tendency to be successfully funded.

1 Introduction

In today's world, technology plays a very crucial role in day-to-day life. Huge percentage of the population in the world is bounded to the technology. One of the rapidly growing and widely used product of the technology is social media. A significantly large number of people uses Twitter, Facebook, YouTube. These social platforms are the best medium for common people to express their opinions and thoughts. These thoughts can be about the products and services, which they have previously used or currently using or these opinions can be about any random or trending topics. The social media platforms are considered by the industries, technology giants, and organizations as the best medium to interact with the common people. People usually posts, shares, likes their thoughts about any product or services using these social platforms, and this can be a piece of very crucial

information for the respective industries. This makes the social media platforms very powerful and informative medium (Chiarello et al., 2018).

Among these social platforms, Twitter is considered to be one of the most popular medium for expressing one's thoughts. It is the medium where people can write messages in the form of tweets. One can even interact the celebrities, giant business and technological tycoons, and even with the politicians through Twitter. It is believed to be the best medium for expressing yourself on anything or about anyone. Twitter has around 350 million users and is a widely used social platform. Also, during the US presidential elections 2016, it had broken several records of being the highly used platform. There were records of around 40 million tweets in one day by general people (En.wikipedia.org, 2019).

As Twitter is one of the most trending social platform, the tweets from the users acts as the form of the data, which can be analysed. These tweets data act as the vital aspects for the widely known technique of data science, which is text analytics. The text analytics usually acts as the best approach for predicting the success of any available products or service in the markets (Chiarello et al., 2018). Within text analytics, the technique of sentiment analysis is widely used for understanding the customer's thoughts about any particular products. Sentiment analysis is a process in which the sentiment's prediction of the textual data (tweets) is done. This technique helps in getting more insights about the consumer's behaviour or the expectations and is crucial for understanding the market behaviour (Xu et al., 2014).

1.1 Motivation and Background

Sentiment analysis is computationally intelligent machine learning technique for performing the analysis on the textual data. It helps in turning the corpus of natural words into the source of informative intelligence. As sentiment analysis is a powerful technique, it is useful for getting insights on the people's opinion and helps the industries in efficient decision-making about their products. The application of sentiment analysis can be done in almost all domains, where the information about the end-user's opinions are required to be analysed (Chiarello et al., 2018).

The Kickstarter the organization based in the United States is one of the popular online crowdsourcing platform, which helps the artists and entrepreneurs to campaign for their new ideas, innovations, and products. The campaigning is done by the product designers and developers for getting the funding is for their products and innovations (projects) online on Kickstarter. The projects on Kickstarter are available on eight different categories like technology, films, arts, etc. This platform is crowdsourced and allows all the people around the world to fund and support any project if they are interested. Till now around 167,652 projects are successfully funded and around billions of the amount is been funded (Help.kickstarter.com, 2019).

Any aspirant artist or the entrepreneur can start the campaign on the Kickstarter. However, it is not always the case that the campaigns are always successful. The success of the campaign on the Kickstarter depends on lots of factors, but understanding the progression of the campaign and its ability to succeed is important from the perspective of respective campaign initiator (Chen et al., 2013). In this research, the analysis is done on identifying whether or

not the twitter tweets are having any significant impact on projects ability to be successfully funded. This analysis may, therefore allow us to predict whether a particular project have an ability to succeed.

From the perspective of campaign starters on Kickstarter, it will be helpful for them to understand whether their project is on the verge of success or failure. This information about the project status in terms of success, if available to the campaigners during the campaign will help them in a significant manner. Predictions of campaign progression will allow the campaigners to make the necessary arrangements and take crucial decisions about their products being campaigned. These predictions about their projects will help the campaigners to take an important decision regarding manufacturing and campaigning of their products, which will allow them to save their significant amount of money and resources. If this information about the project tendency to succeed is available during the campaign, the other group of people who would be benefited are the people who have funded and supported this project. This prediction will allow them to take accurate decisions about their further investments on that particular project as well as about their future investments on Kickstarter projects. The two other groups of people who will be benefited from these predictions are the ones who are potential funders of the projects and the others are the potential entrepreneurs or artists. These predictions will allow potential funders to take decisions wisely and invest smartly in the future, as well as these predictions will allow the potential entrepreneurs or artists to identify the area or domain of getting success.

1.2 Research Question

RQ: “How can sentiment classification of the tweets related to Kickstarter projects using lexicon based approach and machine learning techniques (Decision Trees, Random Forest, Support Vector Machine, and Naive Bayes) help in predicting the successful funding of Kickstarter projects campaigns?”

The above research problem is addressed by setting the following objectives. These objectives are implemented, evaluated and outcome of the implementation is presented.

1.3 Research Objective

Objective 1: Identify the existing work done on predicting the Kickstarter projects success using Machine Learning approach.

Objective 2: Gathering tweets related to the Kickstarter projects and classifying the tweets into sentiments using the lexicon based approach. Followed by the implementing and evaluating the machine learning models on these tweets.

Objective 2(a): Gathering the tweets, pre-processing them and the classifying the tweets into three sentimental categories Positive, Negative and Neutral using lexicon based approach. (Unsupervised Machine learning)

Objective 2(b): Further processing the categorized tweets using the text analytics techniques like implementing Data Pipeline processing, TF-IDF (Term Frequency and Inverse Document frequency), N-grams, SVD (Singular Value Decomposition). After implementing,

evaluating the machine learning models like (Decision Tree, Random Forest, SVN, Naïve Bayes, KNN) on processed tweets.

2 Related Work

2.1 Introduction

In this chapter, the literature survey is carried out about the work which is done in the field of textual analysis. It is focused towards the research work which is done considering the successful prediction of Kickstarter projects or the campaigns. The primary consideration is given to the researches which are targeted towards the use of machine learning and textual analysis in predicting Kickstarter projects success.

2.2 Review of the Sentiment Analysis and its Implementation for Predictions Related to Kickstarter Projects

A project by Desai, Gupta, and Truong, 2015 evaluates 26,000 projects of Kickstarter for determining the success rate of a project. For this data from project description is extracted in HTML format, N-grams, sentiment analysis and, SVM with Gaussian kernels is performed for predicting. It is found that language does play an important role in the progress of a project. Similarly, in another research, the sentiment analysis was implemented on the comments of the Kickstarter project backers. It was identified that the project backer's sentiment and their interaction, act as an important source of information for start-ups in Kickstarter (Courtney, Dutta and Li, 2016).

Chiarello et al., 2018 acknowledge that performing sentiment analysis is the most appreciated method for social media mining. The model generated is for the consumer buying a new technology which can detect the pros and cons with the use of lexicon methodology. Another research was focused on quality of a post is crucial to derive to a judgemental conclusion. So, research by RM and Halkarnikar, 2018 used Lexicon based sentiment analysis approach to classifying comments into positive, negative or neutral. This gives an overview of a post, which is liked or disliked. An almost similar approach was implemented in another research in which a binary classifier (Sentiment analysis and latent Dirichlet allocation) that tests linguistic, campaign content for predicting the success of a campaign, achieves more than 70% efficiency. It is pointed out that language and project properties alone can be used to predict whether a project will benefit or not (Sawhney, Tran and, Tuason, 2016)

2.3 Review of the Machine Learning Implementation on Kickstarter Campaigns

Hou, Zhang and Zhang, 2019 have considered title images for predicting the success or failure of crowdfunded projects. They have retrieved title images from HTML pages and extracted aesthetic features. Thereafter they used emotions from the title images by using Deep learning analysis. Features used in deep learning analysis were the number of backers, amount and pre-set goals. Finally, it was identified that title images have a significant impact

on a campaign's success or failure. In another research by Sharma and Lee, 2018 they compared products on Kickstarter with products on Amazon the largest e-commerce seller. For this dataset of successful and unsuccessful products from Kickstarter was used along with attributes such as number of images, number of creator, product price, etc. After several experimental models, Random Forest Model classifier gave more accuracy. The machine learning approach was utilized by Tran et al., 2018 to identify whether these relaunched projects were a success. As the number of projects failing on Kickstarter because of less funding, some project developers relaunched their projects to be successful, Tran et al., 2018 developed few predictive frameworks, of which deep neural classifier achieved higher accuracy of 82.5%. Dataset of failed-to-succeeded pairs and failed-to-failed pairs were chosen.

An extraordinary research by Pan, Guo and Chen, 2018, integrated categorical, numerical, goal, duration and textual features together in deep neural network and obtained 72% accuracy. Usually, traditional methods in ML use only two features together, for example either goal and duration etc. Therefore, their claim of implementation all three features together in deep learning is unique. Another unique form of study was done by Venugopal and Bagadia, 2015 in which they pointed out that when projects on Kickstarter have at least one edit, the probability of gaining profits is greater. In addition, it is clearly mentioned that only a few edits are better than larger ones. In this case, supervised ML methods is implemented, Adaboost displayed 64% of prediction accuracy.

2.4 Review of the Textual Analysis Implementation on Kickstarter Campaigns

Kickstarter is a platform that connects new projects directly to the investors within a time constraint. The project is termed successful only if the amount required is gathered within the time frame. The prediction of success of Kickstarter was carried out by Chen et al., 2013 by training the ML model using support vector machine and this model was then implemented on the chrome extension as well as an android application. For this research, they considered the factors such as the number of views of YouTube, the number of times the project link was tweeted, etc. Thus, the data gathered from these social media sites were then trained into the SVM model for predicting the project usefulness. Another text mining implementation was done by Xu et al., 2014 because of the number of projects failing on Kickstarter. They carried out the research by focusing to identify the impact of Kickstarter project updates on the Kickstarter campaigns. To carry out this research, the extraction of the project updates data was done for each project such as the content published during updating the project status, frequency of updates, description added during updates. Along with this data the basic project statistics like data about project funding's, backers were also set. The textual analysis implementing LDA, a bag of words model was done on this updates data such as updates description, usage. To find out the association between the project updates and success the logistic regression was implemented. They found out that, there is an association between the update description(date) and campaign outcome, also the update details are crucial to be considered by the campaign runner in order to retain the potential funders.

The use of Tone Analyzer was done by Kim, Cho and Lee, 2016 in their research. As it has been proved that text, comments, a number of backers are crucial indicators for predicting the success of a project but speech and linguistic styles also are as important as text narratives. They used various tools have been used to extract emotions, IBM Watson Tone Analyzer for text mining, Tone Analyzer for emotions. In a similar context, an interesting Exploratory result by Lai, Lo and Hwang, 2017 prove that prediction success rate depends from features such as comments from backers, updates by creators. This, in turn, is favourable for project creators. Classification models such as Gradient boosting, XgBoost, GLM, and Random Forest were engaged to identify funding success of Kickstarter projects. XgBoost results were more accurate while using feature sets like dynamic, text-mining and static features. To conclude the final unique study was by (Song and Zeng, 2015) by identifying that communicating through social media platforms is an important source for Kickstarter projects to seek funds from potential investors. A performance study performed on 100 Kickstarter entrepreneurs along with their behavioural activity on Twitter. Considering content on Twitter, quantitative results classified into behavioural patterns. This research concludes that focusing on social media metrics gained higher crowdfunding.

2.5 Gaps in Identifying the Impact of Sentiments within the Tweets on Kickstarter Project Funding's

From the literature survey, it was found that there is a scope of research in identifying the impact of tweets on the Kickstarter project's fundings. There has not been a significant or direct work done on identifying the presence of correlation between sentiments within the tweets and projects funding's goal. This idea is not unique but the implementation approach and the dataset used are unique.

2.6 Conclusion

Based on the findings from the literature survey, it can be said that there is ample opportunity to build a research focused towards identifying the relation between tweets and project fundings. This correlation can be found by Sentiment Analysis approach and in addition, after figuring out the relation, the machine learning models can be trained to classify the sentiments with the textual data.

3 Technical Design & Research Methodology

In this chapter, the technical design and methodology used for this research is presented. CRISP-DM is the popular methodology in the data-mining field and is widely used in most of the researches. In this research, also the CRISP-DM is used. (Wirth, and Hipp, 2000) However, in order to serve the goal of this research, some modification has been done to this methodology.

3.1 Customized CRISP-DM Methodology Implemented

The popular CRISP-DM (Cross-industry Standard Process for Data Mining) methodology is customized and used in this research.

The diagrammatic representation of this modified methodology which is used for carrying out textual analysis of the Kickstarter projects tweets is explained in Fig. 1.

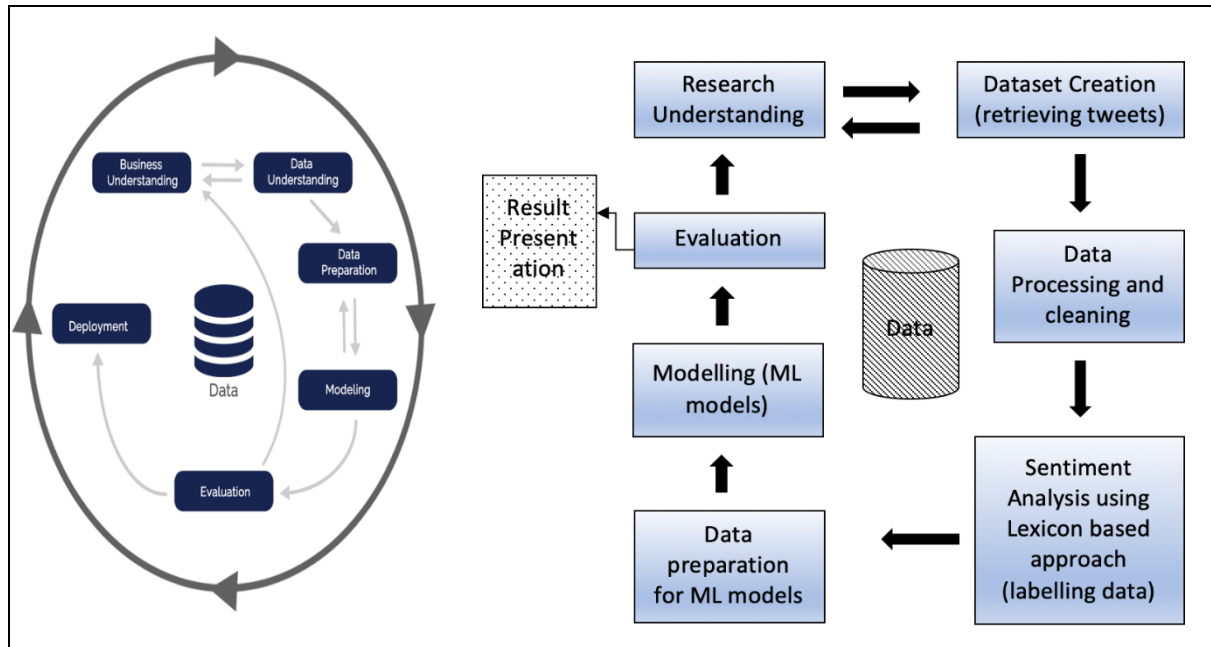


Figure 1: Modified Methodology Implemented (Wirth, and Hipp, 2000)

- **Research objective (Project) understanding:** This phase is focused on understanding the research objectives of this project. In this stage, the technique of textual analysis on the twitter tweets is studied and the implementation of this text mining technique on the Kickstarter projects tweets were inferred.
- **Data Creation:** In this phase, the data is been created for this project. This is done by retrieving the tweets using the Twitter API and by using the Python 3 library titled getOldtweets3.
- **Data Processing and Cleaning:** This phase is focused on processing and cleaning the data. The textual data are one of the most unclean data and are not ready for performing any analysis. Therefore, in this phase, this data is clean and processed and made suitable for further processing.
- **Sentiment Analysis (Data Labelling):** As the tweets retrieved from the twitter is not labelled into (positive, negative and neutral tweets) and hence is not suitable for machine learning models. This phase is focussed on labelling or categorizing each tweet into one of the three categories (positive, negative and neutral). This is done by implementing a Lexicon based approach.
- **Data preparation:** Once the data is labelled in the above section, it is further processed and prepared for being suitable for the machine learning models.
- **Modelling:** In this phase, as of now, data is prepared and labelled; the implementation of machine learning models is done. The supervised machine learning models, which are well known for classification problems like Decision trees, Random Forest, Naïve Bayes, SVN, and KNN were implemented.

- Evaluation: In this phase, the results of the implemented machine learning models are evaluated using the different performance metrics like accuracy, precision, recall, etc.
- Results Presentation: In this phase, the presentation of the results is done in the tabular format.

4 Implementation, Evaluation and Results of Text Mining and Machine Learning based Sentiment Analysis

This chapter focuses on the core implementation of the methodology mentioned in chapter 3. In this chapter, the details about the technical implementation of Lexicon based sentiment analysis on Kickstarter projects tweets is explained. The process of data retrieval and dataset creation is elaborated. The text analytics techniques used for processing, cleaning the data and preparing the data for machine learning models are discussed in detail. The performance metrics which is used for evaluating and analysing the results of the implemented machine learning models will be explained. The implementation of the machine learning models is elaborated and discussed along with the details about specific text analytics technique used for improving the ML model's performance like TF_IDF, SVD, Cosine similarities. This will be followed by evaluations and analysis of the results, which will be done based on performance metrics which is discussed in above subsection 3.3. To conclude this chapter, the results of the models implemented will be compared in tabular format.

4.1 Customized Methodology Approach

The diagrammatical representation of the process of implementation is shown in the above Fig. 2. The entire process flow of implementation is followed using text analytics approach. As it is seen from the below Fig. 2, in the first step the tweets associated with the particular Kickstarter project are retrieved. After the retrieval of the tweets, the dataset is created. This dataset is cleaned and processed. Once the data is processed, the sentiment analysis is performed and each tweet is labelled. This labelled data is then prepared by applying several text analytics techniques and made ready for the machine learning models implementation. Finally, the evaluations and results of the models are discussed.

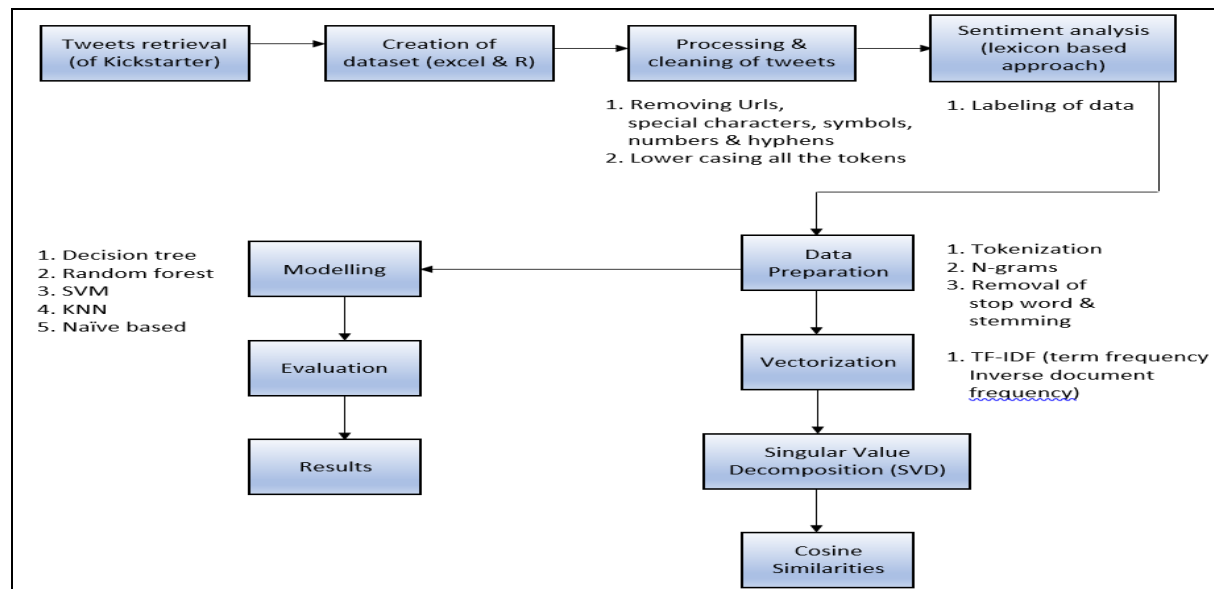


Figure 2: Customized Methodology Workflow

4.1.1 Data Collection

The data is collected by retrieving the tweets from Twitter. For this, research 160 projects where selected. This selection was done for 20 projects from each category. In total eight categories were selected namely Arts, Comics & Illustrations, Design & Tech, Film, Food & Craft, Games, Music, and Publishing. For each of these categories, 20 projects were selected. In addition to this, for each project selected around 30 tweets were retrieved for it. Therefore, in total, the 4800 tweets were retrieved. This retrieval of the tweets was done using twitter API (Developer.twitter.com, 2019) and by using a new Python 3 library, called GetOldTweets3 0.0.10. The library GetOldTweets was used because it has the functionality of retrieving the old tweets as well, which is a limitation of the Twitter API (PyPI, 2019). To retrieve the tweets, the name of the projects was used as the keyword in the python code. Along with the tweets, the information about the number of times the particular tweet has been retweeted was retrieved. In addition to this, the additional retrieval of information was done manually from kickstarter.com for every project. The information included the funding goals (in Dollars), amount funded (in dollars), and a number of people who have backed/supported the project (backers).

4.1.2 Dataset Creation

The data retrieved in the above stage of data collection was then used for the creation of the dataset. The MS -Excel and R studio were used for the dataset creation. The tweets retrieved and number of retweets represented first two columns, the rest columns were represented by the information gathered manually from the kickstarter.com such as project name, number of backers, amount funded (in dollars), project category and funding goal (in dollars). The dataset created can be seen in the Fig. 3.

Tweets	Retweets	ProjectName	No. of Backer	Amount Pledges in \$	Project Categr	Amount Goal in dollars
"Five of Pentacles" Nowhe	6	Cosmic Cycles Tarot	396	22846	Arts	1593
"Page of Swords" Characte	1	Cosmic Cycles Tarot	396	22846	Arts	1593
"Knight of Wands" One of	3	Cosmic Cycles Tarot	396	22846	Arts	1593
The Cosmic Cycles deck is :	0	Cosmic Cycles Tarot	396	22846	Arts	1593
"Ace of Swords" A little vic	4	Cosmic Cycles Tarot	396	22846	Arts	1593
"Seven of Cups" Many opt	4	Cosmic Cycles Tarot	396	22846	Arts	1593
I'm glad to announce that	2	Cosmic Cycles Tarot	396	22846	Arts	1593
I just backed The Cosmic C	1	Cosmic Cycles Tarot	396	22846	Arts	1593
OMG, we made it! The kic	1	Cosmic Cycles Tarot	396	22846	Arts	1593
"Ace of Wands" Inspiratio	3	Cosmic Cycles Tarot	396	22846	Arts	1593
The Kickstarter for the pro	20	Cosmic Cycles Tarot	396	22846	Arts	1593
Tomorrow is the release d	1	Cosmic Cycles Tarot	396	22846	Arts	1593
On the Kickstarter we are j	4	Cosmic Cycles Tarot	396	22846	Arts	1593
The project I have been wi	6	Cosmic Cycles Tarot	396	22846	Arts	1593
I have exciting news of the	0	Cosmic Cycles Tarot	396	22846	Arts	1593

Figure 3: Dataset Created after data collection

4.1.3 Processing and Cleaning

The textual data is considered to be the most unclean data and require lots of processing to be done. The data set created in the above step of data creation, especially the tweets column required a lot of cleaning to make the data available for further step. In this stage, the tweets data was cleaned by removing the URLs, special characters, symbols, question mark, exclamation, apostrophes, commas, and numbers. In the textual analysis, the above features are removed because they are considered to be of least importance, as these features do not play any role in getting the insights from the data. After the removal of these features, the white spaces were removed and the remaining tweets sentence was converted to the lower case (RM and Halkarnikar, 2018).

4.1.4 Sentiment Analysis

After the processing and cleaning of the dataset is done, the data is now ready for further processing. In this phase, the sentiment analysis is performed on every tweet using the Bing lexicon based approach. This approach classifies every tweet into one of the three sentiment categories (Positive, Negative and Neutral). Lexicon based approach the popular and widely used unsupervised technique for labelling the data into sentiments groups. It also has the evidence of being a high-performance model in case of unsupervised classification of tweets. (Taboada et al., 2014). Also, as the tweets which were retrieved and processed were not labelled, this makes these tweets data not suitable for the implementation of well-known machine learning classification models. Therefore, the implementation of this technique also serves the purpose of labelling and categorizing the data.

Thus, to get the sentiment related insights from the tweet data, this technique was applied. The application of this unsupervised technique will give a significant understanding of tweets sentiment and its association with funding's of the Kickstarter project campaign. This understanding will further help in gaining insightful information about the particular Kickstarter project. So, this implementation served the two main crucial purposes of this research. One is getting the sentimental insights within the tweet and identifying the correlation between sentiments within tweets and findings of the particular Kickstarter project. Second, is labelling of the data into three main sentiments categories (positive, negative, neutral) and to make the dataset suitable for implementation of supervised machine learning models.

The lexicon-based approach was therefore used to identify the sentiments within each tweet. The words within the tweets are used to identify the sentiments within them based on the lexicons and this is done using the score associated with them. In this approach, every word within the tweet is assigned with the score and these scores are then used for classifying the tweets (Taboada et al., 2014). For this research, the technique of BING lexicons was used. In this technique, from the BING notations, the list of positive (2006 words) and negative words (4783 words) was retrieved and stored in the two different text files. The words within this both positive and negative text files were used for the calculation of the sentiments within the tweets (Liu, B., 2015).

Implementation : This sentiment analysis was done in R studio by writing a function which accepts positive words list , negative words list , and the once tweet. This function will then return the sentiment for that particular tweet. This was done for all the tweets. The R packages which were used for the implementation of this technique were (data.table, stringr, plyr).

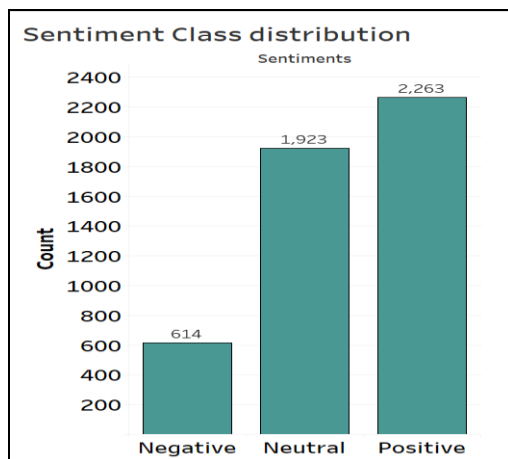


Figure 4: Sentiment Distribution

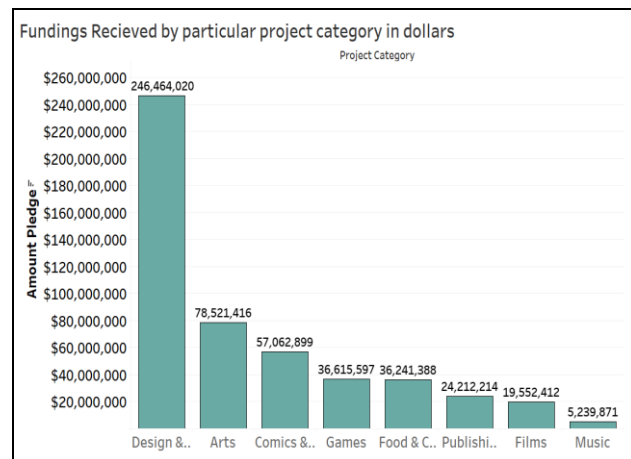


Figure 5: Total amount of funding received

Evaluation & Results : After the implementation of the above algorithm in R studio, the tweets data was successfully labelled into three sentimental categories (Positive, Negative and Neutral). This labelled data is now suitable for further processing and implementation of supervised machine learning models. Fig. 4 shows the distribution of the sentiments after labelling the data. There seems to be a class imbalance. Therefore, to reduce the effect of class imbalance the 10 folds cross-validation has been repeated 3 or 5 times.

The bar graph in Fig. 5 shows the total amount of funding received by the projects in each project category. It can be seen that the projects of the Design & Technology category have the highest funding's which is followed by Arts, Comics & Illustrations, Games, Food & Craft, Publishing, Films, Music categories.

I. Correlation of top-funded project categories and positive tweets

The below Fig. 6 shows that the highest funded project categories have the high number of total positive tweets. For instance, as shown in Fig. 5, the project categories like Design & Technology, Arts, and Food & Craft are among the highest funded project categories. Further analysis was done using Tableau software. One project from each of these three top-funded categories was chosen, and then a total number of positive sentiments for each project was calculated. The results of this analysis are shown in Table 1.

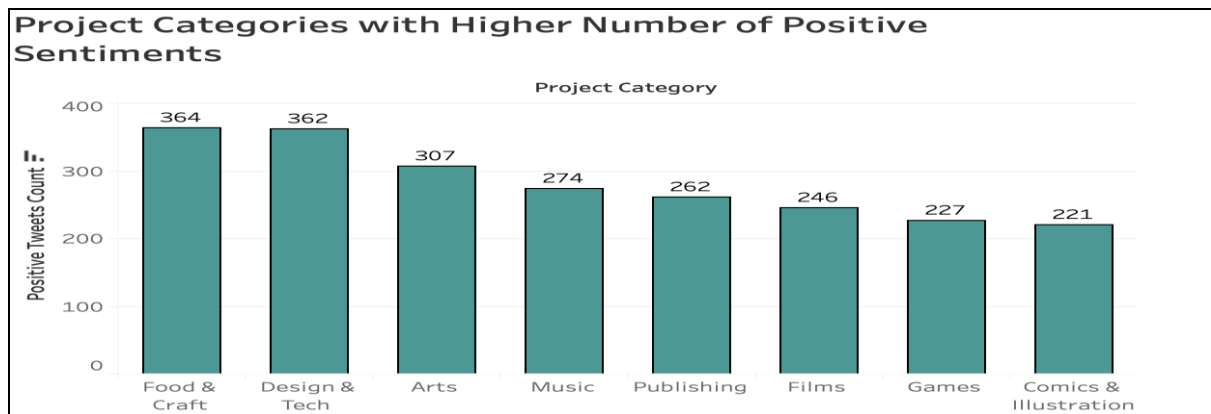


Figure 6: Association of Positive Sentiment Tweets and Top funded project categories.

Table 1: Top Funded projects and the number of positive tweets associated with them

Project Name	Project Category	Number of Positive Tweets
Goose that Lays Golden Egg	Food & Craft	16 out of 30 positives (7 neutral & 7 negative)
Go Sun Sport: Portable High Efficiency Solar Cooker	Design & Tech	28 out of 30 positives
Seb Mckinnon Magic the Gathering	Arts	29 out of 30 positives

Thus, based on the results mentioned in above Table 1 and Fig. 6, it can be said that there is a significant association between the high funding's of the project and the positive tweets.

II. Correlation of lowest funded project categories and negative tweets

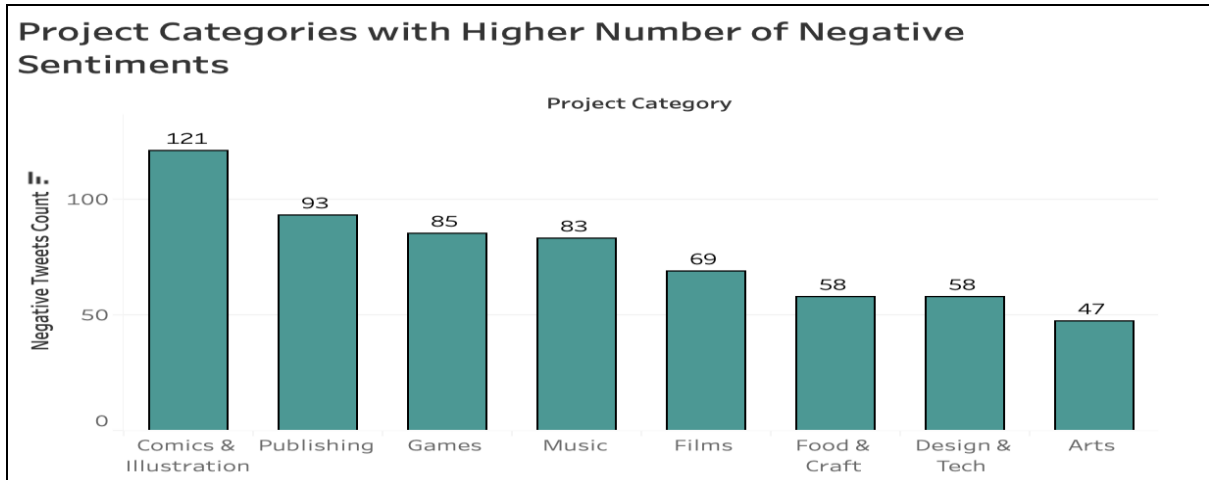


Figure 7: Association of Negative Sentiment Tweets and Lowest funded project categories.

The above Fig. 7 shows that the lowest funded project categories have a high number of total negative tweets. For instance, as shown in Fig. 5, the project categories like Comics & Illustration, Publishing, and Games are among the lowest funded project categories. Therefore, further analysis was again done using Tableau software. One project from each of these three lowest funded categories was chosen, and then the total number of negative sentiments for each project was calculated. The results of this analysis are shown in Table 2.

Table 2: Lowest Funded projects and number of negative tweets associated with them

Project Name	Project Category	Number of Negative Tweets
Wynonna Earp: Bad day at black rock	Comics / Illustration	15 out of 30 negatives (11 neutral & 4 positive)
World of Kanji	Publishing	12 negatives (11 neutral, 7 pos)
Court of Dead: Mourners call board game	Games	20 negatives out of 30

Thus, based on the results mentioned in above Table 2 and Fig.7, it can be said that there is a significant association between the low funding's of the project and the negative tweets.

4.1.5 Data Preparation for Implementation of Supervised Machine Learning Models

After the implementation of sentiment analysis on the tweets data, the dataset is now labelled and three different categories are available (Positive, Negative, Neutral). Before implementing the machine learning models, the textual data has to be processed using textual analytics techniques, to make it properly suitable for the implementation. The key aspects of textual analysis are to pre-process the data to get significant insights from it. The following are the steps and the techniques, which were implemented for processing the data.

I. Tokenization

One of the crucial steps of textual analysis is to tokenize the text data, in which every word within the text will be represented as the token and every word is considered separate from another. In this research, every tweet was tokenized and every word within the tweet was

then represented as tokens. This was done using the functions in the R library titled as (quanteda) (Lai, Lo and Hwang, 2017).

II. N-Grams

In the recent research by Desai, Gupta and Truong, 2015 they suggested that tokenizing the word based on the white space is not efficient approach and there is a high probability of losing the meaning of the word. Therefore, they suggested the approach of Bi-Grams, N-Grams which helps in detecting the more insightful meaning from the text. These techniques also consider the context of the text. Therefore, after the implementation of TF-IDF basic tokenization, the technique of N-Grams was implemented (Desai, Gupta and Truong, 2015).

III. Removal of Stop Words

After the tokenization of the tweets, the process of removing the stop words from the tokenized tweets was carried out. The stops words are the basic English language functional words which are bound to appear almost everywhere, and these words are not supposed to have any predictive power in gaining insights from the text. Thus, these words were removed from the tokenized tweets using the function in R library (quanteda) (Lai, Lo and Hwang, 2017).

IV. Stemming

After the removal of stop words, the technique of stemming was implemented using the stemming function in R library (quanteda). The stemming is the technique in which the root form of the words is detected and this is done by removing the endings of the words. This technique helps in reducing the dimensionality of the textual data. The problem with the dimensionality is very evident in textual analysis, and any technique for helping to reduce dimensionality is crucial (Lai, Lo and Hwang, 2017).

4.1.6 Vectorization

I. TF-IDF (Term Frequency and Inverse Document Frequency)

Term Frequency and Inverse Document Frequency is a widely used technique in text analytics. In this technique, the tokenized words in the tweets which are frequently occurring in the entire dataset are penalized. This is done because, if the particular word is frequently occurring it means that particular word is having less importance and less predictive power (Paik, J.H., 2013).

The formula for TF (Term Frequency):

$$TF(t, d) = \frac{freq(t, d)}{\sum_i^n freq(t_i, d)}$$

Formula 1: TF (Term Frequency Formula) (Tf-Idf, 2019)

Formula Explanation: where: ‘t’ – is term, ‘d’ – is document, and ‘n’- is the number of distinct term in document ‘d’.

frequency(t,d) - the number of times term ‘t’ appeared in one document(one tweet).

TF(t,d) – the frequency of ‘t’ term in document ‘d’ (one tweet) divided by sum of frequencies of all ‘t’ in the document. This calculation will give the percentage of a particular term in one document (in one tweet). This calculation will also normalize the term frequency across documents (tweets) of different length (Paik, J.H., 2013).

The formula for TF (Term Frequency):

$$IDF(t) = \log\left(\frac{N}{count(t)}\right)$$

Formula 2: IDF (Inverse Document Frequency Formula) (Tf-Idf, 2019)

Formula Explanation: where: ‘t’- is term, ‘N’ – is count of distinct document in the corpus, and count(t) – count of document in the corpus in which term ‘t’ is present.

In the current Tweets scenario, the above-mentioned formula in Fig. 9 will work in the following way. Assume, ‘t’ is word ‘Kickstarter’ and N is the 4800 tweets. If, the word ‘t’ is present in every tweet, then the above formula in figure will return 0. This is because, if the particular word is present in every tweet, then it is of least importance and therefore that word is penalized (Paik, J.H., 2013)

Therefore, the formula for TF-IDF is the product of term frequency and inverse document frequency (Paik, J.H., 2013)

Implementation of TF-IDF: The TF-IDF technique was implemented in this research in order to get more insights from the textual data. In order to implement the TF-IDF, the normalization of all the tweets was done to make all tweets of the same length. This was done for implementing Term Frequency and IDF. The separate functions for TF and IDF were created in R. After normalization, both TF and IDF calculations was performed on normalized tweets using functions created. And finally, calculations for TF-IDF function were implemented on the tweet dataset using the results from TF and IDF functions (Desai, Gupta and Truong, 2015).

II. Latent Semantic Analysis

LSA is the process of extracting the relationship between the tweets and words within the tweets assuming that are close in meaning will appear in a correlated piece of text. This technique basically uses the SVD Singular Value Decomposition factorization of term document-matrix to extract this relationship. To implement SVD, the document (tweet) term (word within tweet) matrix was transposed to term (word within tweet) document (tweet) matrix. (Gong, and Liu, 2001).

$$SVD \text{ of } X = Xt = U\Sigma V^T$$

Formula 3: SVD (Singular Value Decomposition Formula) (SVD, 2019)

Formula Explanation: where: ‘U’- contains Eigen vectors of term (words in the tweets) correlation, ‘V’ – contains Eigen vectors of document(tweet) correlations, Sigma contains singular values of factorization. ‘X’ – Matrix before transposition and ‘Xt’ – transposed matrix.

The above formula in Fig.10, is basically extracting the semantic relationships out of the documents at high level of abstraction. For example, this formula will bring the terms like ‘loan’, ‘debt’, ‘credit’ under a single concept of ‘Debt’ (Gong, and Liu, 2001).

As mentioned in the above section of stemming, the big problem with the textual analysis is the dimension, also known as the curse of dimensionality. As of now, we have around 30,424 columns; this technique of LSA will help with the dimensionality problem. It allowed to reduce the columns to 300. The matrix factorization that is carried out in LSA will help in combining the columns and enriching the signal in the data.

Implementation of LSA (SVD): In order to make LSA work the implementation of SVD (matrix factorization) was done on the dataset after implementing the TF-IDF. This was done by using the R package (irlba). In addition, after implementing the SVD on the tokenized tweet dataset, the columns were reduced from 30,424 to 300(Desai, Gupta and Truong, 2015) (Gong, and Liu, 2001)

III. Cosine Similarity

Cosine Similarity is the process of calculating the similarity between documents (tweets) in the vector space. Using cosine between the document vectors is an improvement over the dot product. The benefits of using this technique are that, considering our representation the cosine will be between [0, 1]. Here the cosine of 1 means there is a perfect similarity. Another benefit of using this technique is that metrics work well in high dimensional spaces. The formula for cosine similarity is as follows and this formula will give the similarity between the tweets on the vector space (Sidorov et al., 2014).

$$\cos \theta = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

Formula 4: Cosine Similarity Formula (Cosine Similarity, 2019)

Formula Explanation: where: ‘A’ – Document1 (Tweet 1), ‘B’ - Document (Tweet 2).

Implementation of Cosine Similarity: The implementation of Cosine Similarity was done on the dataset after implementing the SVD. This was done by using the R package (lsa). This was the final process of data processing text analytics pipeline (Gomaa, and Fahmy, 2013).

4.1.7 Performance Evaluation Metrics

The machine learning models which are implemented in this research are evaluated on the basis of certain evaluation metrics. Every trained machine learning model in this project will

produce the results in four categories which include True Positive(TP), False Positive(FP), True Negative(TN) and False Negative(FN). The TP indicates that the labels which are positive sentiment are predicted as positive sentiment, TN indicates that the labels which are negative sentiment are predicted as negative sentiment, whereas the FP indicate that the labels which are negative but is been predicted as positive and FN indicates that the labels are positive but is been predicted as negative (Machine Learning Plus, 2019). For this research, the performance evaluation metrics which are used are accuracy, Sensitivity or Recall, Specificity.

Accuracy: What percentage of the predictions were correct?

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Formula 5: Accuracy Formula(Performance Metrics, 2019)

In the above formula in Fig.12 the numerator denotes the sum of correctly predicted labels, whereas denominator denotes every prediction (Machine Learning Plus, 2019).

Sensitivity or Recall: If in one case the label belongs to one class (for example: positive) then what is the probability of correctly predicting this label as positive.

$$Sensitivity = \frac{TP}{TP + FN}$$

Formula 6: Sensitivity Formula (Performance Metrics, 2019)

The sensitivity of each class (positive, negative and neutral) can be predicted from the above-mentioned formula in Fig. 13 (Machine Learning Plus, 2019)

Specificity: If in one case the label belongs to one class (for example: negative) then what is the probability of correctly predicting this label as negative.

$$Specificity = \frac{TN}{FP + TN}$$

Formula 7: Specificity Formula (Performance Metrics, 2019)

The specificity of each class (positive, negative and neutral) can be predicted from the above-mentioned formula (Machine Learning Plus, 2019).

4.1.8 Implementation, Evaluation and Results of Machine Learning Models Decision Tree, Random Forest, SVM, Naïve Bayes, KNN

I. Implementation, Evaluation and Results of Decision Tree model

Decision trees are the less complex but highly efficient algorithm which is used in both classification and regression problems. Decision tree constructs a model into a tree structure. The dataset provided is broken down into small sets and simultaneously the decision tree is developed (Kotsiantis et al., 2007).

Implementation: Decision trees were implemented because this algorithm is the first preference in any classification problems, because of its less complexity. In this project for the implementation of Decision trees, the 10-fold cross validation was repeated 3 times. This cross validation will then create 30 random stratified samples. This cross-validation strategy is used because of the imbalance in the class labels. The decision tree model was then implemented using this 10-fold stratified cross validation created. For implementing this model and for performing the cross validation the functions from caret package in R is used. (TextAnalytics, 2019)

Evaluation and Results: Based on the results obtained after implementing the Decision Tree algorithm, the accuracy achieved by the Decision Tree Model was 59.64%. This accuracy was achieved at the complexity parameter of 0.0112. This complexity parameter is used for controlling the size of the tree and for selecting optimal tree size. After testing the model on unseen (Test data) was the accuracy achieved was 67.94%. The model has predicted for Negative class to be truly negative (specificity) as 91.22%, and truly positive (sensitivity) as 56.52%. Whereas model has predicted for Positive class to be truly negative (specificity) as 71.32%, and truly positive (sensitivity) as 76.84%. This shows that the model is good at predicting the negative class as compared to the positive class.

II. Implementation, Evaluation and Results of Random Forest model

Random forest is the most powerful algorithm and can be used for both classification and regression problems. It internally uses the bagging technique and is an ensemble learning. In this algorithm, multiple decision trees are created and bagged and can be termed as a bagged classifier. Among these multiple decision trees which are grown in random forest algorithm, each tree is been created using decision tree algorithm (CART) separately. As it is the ensemble of multiple decision trees, the classifier will be more accurate and more precise. In addition to this random forest also helps in getting the information about variable importance, which can be used, for further tuning of the model (Liaw, and Wiener, 2002).

Implementation: In this project, the random forest algorithm was implemented because this algorithm is best suitable for classification problem and is well known for training highly accurate models. To improve the accuracy of the decision tree algorithm implemented before,

the random forest is the best option because of it ensembling of multiple decision trees feature. For this implementation performing 10-fold CV repeated 3 times. Therefore, 30 models were built, and for ‘mtry’ parameter 7 different values were tried. Also, considering random forest algorithms, it builds by default 500 trees. Therefore, the total trees built here after implementation of random forest algorithm are $(10 * 3 * 7 * 500) + 500 = 105,500$ trees. For this implementation, the caret package in R is used (TextAnalytics, 2019).

Evaluation and Results: Implementation of the Random Forest algorithm was done by training the model three times. At first, the model was trained without the tweet length and without implementing the cosine similarity feature. For training second model the tweet length feature was introduced and finally for the third time the model was trained after introducing both text length and cosine similarities features. The performance increased respectively in for every model implemented. Table 3 shows the accuracy achieved by the models.

Table 3: Performance of Models Implemented using Random Forest

Random Forest Models Trained	Accuracy Achieved
Model1 (Without Cosine and Tweet Length)	75.55% (mtry- 300)
Model2 (With Tweet lengths added as feature)	75.58% (mtry -51)
Model3(With both Tweet length and cosine)	77.07% (mtry -102)

The accuracy achieved by the best Random Forest Model was 77.07%. This accuracy was at 102columns per tree (mtry). Therefore, the model3 is considered for the evaluation and further testing. The accuracy achieved after testing the Model 3 on unseen (Test data) was 76.15%. The model has predicted for Negative class to be truly negative (specificity) as 96%, and truly positive (sensitivity) as 60%. Whereas model has predicted for Positive class to be truly negative (specificity) as 81%, and truly positive (sensitivity) as 80%. This shows that the model is good at predicting the negative class as compared to the positive class. Also from this implementation, it can be said that the tweet length and cosine similarity features they play a vital role in training the model.

III. Implementation, Evaluation and Results of SVM model

SVM is the algorithm which uses a linear hyperplane or the sets of hyperplane on high dimensional space for separating the classes available in the training dataset. The maximization of the training point is done for accurately classifying the classes. The data point which is at the margins they play a key role in classification. If in the existing space which is used for classification is not able to separate the classes, then the space is transformed into higher dimensional space. This transformation can be done by the usage of the kernel function in R (Liaw, and Wiener, 2002).

Implementation: SVN was considered for this project because, after the decision tree, SVN is considered to be the next best and accurate model for classification. The SVN was implemented using the two kernels ‘radial’ and ‘polynomial’. The model was trained using

the 10-fold cross validation was repeated 5 times. For this implementation, the caret and e1071 packages in R is used.

Implementation of SVM was done using the two different kernels, ‘radial basis’ and ‘polynomial’. As mentioned above in implementation, these kernels are used for making the hyperplane decision boundary between the classes and for mapping the dataset into higher dimensional space. Results obtained after implementing the SVM can be seen in Table 4.

Table 4: Performance of Models Implemented using Different Kernels (SVM)

Trained SVM Model testing on Test Data	Accuracy
Model implemented using ‘radial basis’ kernel	76.70%
Model implemented using ‘polynomial’ kernel	79.55%

Evaluation and Results: Based on the results obtained after implementing SVM algorithm, the accuracy achieved by the SVM model was 76.70% using radial basis kernel whereas using the polynomial kernel the accuracy achieved was 79.55%. This accuracy was achieved after testing the model on unseen (Test data). This accuracy was the best accuracy achieved among other models implemented in this research.

The model has predicted for Negative class to be truly negative (specificity) as 95.77%, and truly positive (sensitivity) as 59.24%. Whereas, the model has predicted for the Positive class to be truly negative (specificity) as 83.95%, and truly positive (sensitivity) as 85.84%. This shows that the model is good at predicting the negative class as well as the positive class. The performance of this model in predicting the positive class is higher than the Random forest model performance by 5.84%.

IV. Implementation, Evaluation and Results of Naïve Bayes model

Naïve Bayes algorithm is very well known for sentiment based classification. This algorithm is based on the probabilistic approach and is also easy to tune. It helps in detecting the posterior probabilities. This algorithm implements the naïve assumption for classification of the sentiments, that in every tweet all words are used in a way that they are not dependent on each other. For the classification of every tweet (document), this algorithm calculates the probability of the tweet belonging to a particular class. It Naïve Bayes theorem for calculation of these probabilities (Liaw, and Wiener, 2002).

Implementation: Naïve Bayes was implemented in this project because it works well with the sentiment classification problem, also to train the model on the probabilistic perspective. The model was trained using the 10-fold cross validation was repeated 5 times. For this implementation, the caret and e1071 packages in R is used.

Evaluation and Results: According to the results obtained after implementing the Naïve Bayes algorithm, the accuracy achieved by the training the Naïve Bayes Model was 63.84%. This accuracy was achieved after testing the trained model on unseen (Test data). The model has predicted for Negative class to be truly negative (specificity) as 88.35%, and truly positive (sensitivity) as 48.37%. Whereas model has predicted for Positive class to be truly

negative (specificity) as 82.89%, and truly positive (sensitivity) as 61.21%. This shows that this model is also good at predicting the negative class as compared to the positive class.

V. Implementation, Evaluation and Results of KNN model

This algorithm classifies the classes based on nearest neighbour's concept. After performing several parametric approaches, the k value of has to be decided. This value also depends upon data used for the training the models (Liaw, and Wiener, 2002).

Implementation: KNN model was trained using the 10-fold cross validation was repeated 3 times. The K value which was used was 5. For this implementation, the caret, ISLR, class, model metrics packages in R is used.

Evaluation and Results: According to the results obtained after implementing the KNN algorithm, the accuracy achieved by the training the KNN Model was 69.19%. The optimal k value, which was used, was 5. After testing this trained model on unseen test data, the accuracy achieved was 70.45%. The model has predicted for Negative class to be truly negative (specificity) as 94.25%, and truly positive (sensitivity) as 52.71%. Whereas model has predicted for Positive class to be truly negative (specificity) as 76.45%, and truly positive (sensitivity) as 76.99%. This shows that this model is also good at predicting the negative class as compared to the positive class.

5 Conclusion and Future Work

Twitter is the widely used social media, and it plays a crucial role in day-to-day activity. It has been ranked as the most visited social media platform. People use Twitter for expressing their opinions and thought about any upcoming product, ideas, campaigns, and thought. Being the most popular social media platform, significant insights can be retrieved from twitter data by performing the analysis on it. These insights can be crucial from the business point of view as well in making significant predictions.

In this research, we used twitter data of Kickstarter projects and performed sentiment analysis of this data. This analysis was done to identify the impact of sentiment within the tweets on the successful funding of the project. To carry out this analysis the text mining, sentiment classification, and machine learning approach were utilized. After carrying out the sentiment analysis on the tweets, it was identified that there is a significant association between the high funding's of the project and the positive tweets and also there is a significant association between the low funding's of the project and the negative tweets. These findings of the correlation between tweets sentiments and project funding's were then further utilized as the base for training the machine learning models. As there was evidence of an association between sentiments within the tweets and project funding's, the machine learning models were trained to classify the tweets. After implementing the five machine-learning models, it can be seen from Table 5, SVM algorithm was high performing model as it achieved the highest accuracy of 79.55%, which was then followed by Random Forest model.

Table 5: Performance of Models Implemented.

Models Implemented	Accuracy Achieved
SVM (Polynomial Kernel)	79.55%
Random Forest	76.15%
KNN	70.45%
Decision Trees	67.94%
Naïve Bayes	63.84%

These models can then be used to classify the comments or tweets associated with a particular project and help in predicting the information about the project funding's. Also, these models will, therefore help in predicting whether the particular project have a tendency to be successfully funded. This predictive ability about project successful funding's, if available to the campaigners during the campaign will help them in a significant manner as mentioned in the motivation section above. This information about the project tendency to succeed if is available during the campaign, the other group of people who would be benefited are the people who have funded and supported this project. The two other groups of people who will be benefited from these predictions are the ones who are potential funders of the projects and the others are the potential entrepreneurs or artists. Therefore, the research question was successfully answered and all the research objective mentioned in chapter 1 were achieved.

Future Work: In this research, 20 projects from each of the eight project categories, therefore in total 4800 tweets were considered. In the future, the textual analysis can be carried out on more number tweets. The information about the comments mentioned on the Kickstarter website, which is associated with each project, can be used in the future to perform textual or sentiment analysis. This analysis of comments might give further insights. For this research only the tweets from the Twitter platform were considered for analysis, in future other textual contents such as comments related to the respective project can be retrieved from other social media platforms and used for analysis. Also, in textual analysis, there is always a problem with dimensionality, which is computationally expensive. Thus, in the future, the approach of Big Data analysis can be used for dealing with huge data.

Acknowledgement: I would specially like to thank my Supervisor Dr. Catherine Mulwa for her crucial guidance and her continuous support throughout the research project. Catherine was always available with her comments and suggestions in every phase of this project. I would also like to acknowledge my friend Shweta Sahu for her help and support during this research. Finally, I would like to acknowledge my mother for encouraging me throughout my journey of Masters.

References

Bing Liu, Minqing Hu and Junsheng Cheng. "Opinion Observer: Analyzing and Comparing Opinions on the Web." Proceedings of the 14th International World Wide Web conference (WWW-2005), May 10-14, 2005, Chiba, Japan.

Chen, K., Jones, B., Kim, I. and Schlamp, B. (2013). KickPredict: Predicting Kickstarter Success.

Chiarello, F., Bonaccorsi, A., Fantoni, G., Ossola, G., Cimino, A. and Dell'Orletta, F., 2018, September. Technical Sentiment Analysis: Measuring Advantages and Drawbacks of New Products Using Social Media. In 2nd International Conference on Advanced Reserach Methods and Analytics (CARMA 2018) (pp. 145-156). Editorial Universitat Politècnica de València.

Cosine Similarity. (2019). [image] Available at: https://github.com/datasciencedojo/IntroToTextAnalyticsWithR/blob/master/IntroToTextAnalyticsWithR_Part10.pdf [Accessed 10 July 2019].

Courtney, C., Dutta, S. and Li, Y., 2016. Resolving Information Asymmetries: Influence of Media and Crowd Sentiment on Crowdfunding Success. In Academy of Management Proceedings (Vol. 2016, No. 1, p. 14663). Briarcliff Manor, NY 10510: Academy of Management.

Desai, N., Gupta, R. and Truong, K., 2015. Plead or pitch? The role of language in kickstarter project success.

Developer.twitter.com. (2019). Docs. [online] Available at: <https://developer.twitter.com/en/docs.html> [Accessed 8 July. 2019].

En.wikipedia.org. (2019). Twitter. [online] Available at: <https://en.wikipedia.org/wiki/Twitter> [Accessed 2 July 2019].

Gomaa, W.H. and Fahmy, A.A., 2013. A survey of text similarity approaches. International Journal of Computer Applications, 68(13), pp.13-18.

Gong, Y. and Liu, X., 2001, September. Generic text summarization using relevance measure and latent semantic analysis. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 19-25). ACM.

Help.kickstarter.com. (2019). [online] Available at: <https://help.kickstarter.com/hc/en-us> [Accessed 4 July 2019].

Hou, J.R., Zhang, J. and Zhang, K., 2019, January. Can Title Images Predict the Emotions and the Performance of Crowdfunding Projects? In Proceedings of the 52nd Hawaii International Conference on System Sciences.

Kamath, R.S. and Kamat, R.K., 2016. SUPERVISED LEARNING MODEL FOR KICKSTARTER CAMPAIGNS WITH RMining. International Journal of Information Technology, Modeling and Computing (IJITMC), 4(1).

Kim, J., Cho, D. and Lee, B., 2016. The Mind Behind Crowdfunding: An Empirical Study of Speech Emotion in Fundraising Success.

Kotsiantis, S.B., Zaharakis, I. and Pintelas, P., 2007. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160, pp.3-24.

Lai, C.Y., Lo, P.C. and Hwang, S.Y., 2017. Incorporating Comment Text into Success Prediction of Crowdfunding Campaigns. In *PACIS* (p. 156).

Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. *R news*, 2(3), pp.18-22.

Liu, B., 2015. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Machine Learning Plus. (2019). Top 15 Evaluation Metrics for Machine Learning with Examples. [online] Available at: <https://www.machinelearningplus.com/machine-learning/evaluation-metrics-classification-models-r/> [Accessed 10 July 2019].

Paik, J.H., 2013, July. A novel TF-IDF weighting scheme for effective ranking. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (pp. 343-352). ACM.

Pan, C., Guo, Y. and Chen, Y. (2018). Predicting the success of Crowdfunding. *CS230: Deep Learning*, Winter 2018, Stanford University, CA.

Performance Metrics. (2019). [image] Available at: https://github.com/datasciencedojo/IntroToTextAnalyticsWithR/blob/master/IntroToTextAnalyticsWithR_Part9.pdf [Accessed 10 July 2019].

PyPI. (2019). GetOldTweets3. [online] Available at: <https://pypi.org/project/GetOldTweets3/> [Accessed 10 February 2019].

RM, M.N. and Halkarnikar, P.P., 2018. Text Analytics of Web Posts' Comments Using Sentiment Analysis.

Sawhney, K., Tran, C. and Tuason, R., 2016. Using Language to Predict Kickstarter Success. Sharma, V. and Lee, K., 2018, August. Predicting Highly Rated Crowdfunded Products. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 357-362). IEEE.

Sidorov, G., Gelbukh, A., Gómez-Adorno, H. and Pinto, D., 2014. Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas*, 18(3), pp.491-504.

Song, Y. and Zeng, G., 2015. UNCOVER SUCCESSFUL ENTREPRENEURS' CROWDFUNDING BEHAVIORS THROUGH TWITTER.

SVD. (2019). [image] Available at: https://github.com/datasciencedojo/IntroToTextAnalyticsWithR/blob/master/IntroToTextAnalyticsWithR_Part7.pdf [Accessed 10 July 2019].

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M., 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2), pp.267-307.

TextAnalytics. (2019). *datasciencedojo/IntroToTextAnalyticsWithR*. [online] Available at: <https://github.com/datasciencedojo/IntroToTextAnalyticswithR> [Accessed 11 Mar. 2019].

Tf-Idf. (2019). [image]Available at: https://github.com/datasciencedojo/IntroToTextAnalyticsWithR/blob/master/IntroToTextAnalyticsWithR_Part5.pdf [Accessed 10 Jul 2019].

Tran, Thanh & Lee, Kyumin & Dontham, Madhavi & Chung, Jinwook. (2018). Learning from Failure: How to Make a Failed Crowdfunding Project Successful. 10.13140/RG.2.2.18362.03526.

Venugopal, V. and Bagadia, S. 2015, Understanding the Dynamics of Crowdfunding: Kickstarter Edits CS224N Final Project.

Wirth, R. and Hipp, J., 2000, April. CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (pp. 29-39).

Xu, A., Yang, X., Rao, H., Fu, W.T., Huang, S.W. and Bailey, B.P., 2014, April. Show me the money! an analysis of project updates during crowdfunding campaigns. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 591-600). ACM.