**METHODOLOGY**

A startup is a young company founded by one or more entrepreneurs in order to develop a unique product or service and bring it to market. By its nature, the typical startup tends to be a shoestring operation, with initial funding from the founders or their families.

Startup India Scheme is an initiative by the Government of India for generation of employment and wealth creation. The goal of Startup India is the development and innovation of products and services and increasing the employment rate in India. Benefits of **Startup India Scheme** is Simplification of Work, Finance support, Government tenders, Networking opportunities. Startup India was launched by Prime Minister Shri. Narendra Modi on 16th January 2016. Let us learn more about Benefits and Eligibility of Startup India.

This dataset has funding information of the Indian startups from January 2015 to August 2017. It includes columns with the date funded, the city the startup is based out of, the names of the funders, and the amount invested (in USD).

**SYSTEM ARCHITECTURE**

The following diagram illustrates the working of the process of project

**New Live Data**

**Data Collection**

**Data Cleaning**

**Pre-processing of Data**

**Data Classification**

**Analysis of Result**

1. **Planning and Data collection**

As the name suggests Data Collection you must get all the data needed to solve the problem. Collecting data is not very easy because most of the time you won’t find data lying in a database, waiting for you. Instead, you’ll have to do some research and collect the data or scrape it from the internet. Data used in this project is a set Indian startup funding from Kaggle. Each startup includes the following information:

1. Start-up Name
2. Start-up Date/Month/Year
3. Industry Vertical
4. Sub Vertical
5. City Location
6. Investors Name
7. Investment Type
8. Amount in USD
9. **Pre-processing of data**

The data pre-processing is responsible for preparing a data for further processing. Classification will be based on various reviews and rating of customer. In Data processing the input data. the customer reviews dataset is pre-processed to improve the classification results. The process of data pre-processing includes two main steps. One is Parts of Speech Tagging which is mainly for positive phrases. To find the phrases with negative prefixes we use Negation Phrase Identification algorithm. Data pre-processing is done to eliminate the incomplete, noisy and inconsistent data a stop word. POS tagging is the process of marking a word in a text as corresponding to a particular part of speech as its context i.e. relationship with adjacent and related words in a phrase, sentence or a paragraph Parts of speech include nouns, verbs, adverbs, adjectives, pronouns, conjunction, prepositions and determiners

1. **Data Cleaning**

Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. This data is usually not necessary or helpful when it comes to analysing data because it may hinder the process or provide inaccurate results. There are several methods for cleaning data depending on how it is stored along with the answers being sought.

Data cleaning is not simply about erasing information to make space for new data, but rather finding a way to maximize a data set’s accuracy without necessarily deleting information in order to prevent wrongful predictions, it is important to get rid of any inconsistencies in the data.

1. **Data Classification**

After cleaning of our dataset, we move on to our data processing part or we can use algorithm on our dataset which can further classify our data set on basis of positive and negative comments. We need algorithms to classify our text.

In Data Classification Opinions are given by the user through various sources about the product and their services. Machine learning techniques like Naive Bayes (NB), maximum entropy (ME), and support vector machines (SVM) have achieved great success in sentiment analysis. Sentiment analysis that saves running time and reduces computational complexity is the analysis done.

**Naive Bayes classifier**

Naïve Bayes is an algorithm of probability based on Bayes theorem of learning. It aims to create a model in the form of probability. The advantage of Naïve Bayes is an effective

method which is easy processing. In other words, classiﬁcation features are independent of each other given the class. In the context of text classiﬁcation, NB adopts the Bag-of-Words (BOW) approach. More speciﬁcally, the features of the model are the individual words of the text. The data is typically represented by a 2-dimensional word x document matrix. In the Bernoulli NB model, an entry in the matrix is a binary value that indicates whether the document contains a word or not (i.e., {0,1}). The Multinomial NB, on the other hand, uses normalized frequencies of the words in the text to construct the word x document matrix

**Support Vector Machines (SVM**):

SVM is a supervised machine learning algorithm that is used for classiﬁcation and regression analysis in multidimensional data spaces. SVM attempts to ﬁnd optimal hyperplanes for linearly separable patterns in the data and then maximizes the margins around these hyperplanes. Technically, support vectors are the critical instances of the training set that would change the position of the dividing hyperplane if removed. SVM classiﬁes the data by mapping input vectors into an N-dimensional space, and deciding on which side of the deﬁned hyperplane the data instance lies. SVMs have been empirically shown to be effective in domains where the data is sparse and highly dimensional.

SVM and NB have been found to work well with short text. Short-text is a relatively recent Natural Language Processing (NLP) type of text that has been motivated by the explosive growth of micro-blogs on social media (e.g. Tweets and YouTube and Facebook comments) and the urgent need for effective methods to analyze such large amounts of limited textual data. The main characteristics of short-text include data sparsity, noisy content, and colloquial terminologies. In what follows, we investigate the performance of these two classiﬁers in detecting useful feedback present in comments.

* **K-means cluster**

K-means clustering is a type of unsupervised learning, which is used when you have unlabelled data (data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity.

K-means clustering is used to classify the retrieved dataset through a certain number of clusters.

1. **Analysis of Result: -**

You now have a nice dataset so this is a good way to start exploring it by building graphs. When you’re dealing with large volumes of data, visualization is the best way to explore and communicate your finding. In order to understand data, it is often useful to visualize it. Normally in Big Data applications, the interest relies in finding insight rather than just making beautiful plots

**Data Exploration and Feature Selection**

* **Correlation Analysis**

We plotted the unscaled 'Amount Raised' against all unscaled numerical features, and found no linear correlations. We then plotted the unscaled 'Amount Raised' against all numerical features normalized using boxcox transformation, and again found no linear correlations. We did not find linear correlations for scaled 'Amount Raised' against unscaled features either. We did find some homoscedastic correlations for scaled 'Amount Raised' against scaled features', but unfortunately these slopes were flat. From our correlation analysis, we infer that using SVR may be a better method for predictive modeling than linear regression. Code for correlation analysis is here.

* **PCA**

Principal Component Analysis (PCA) tries to isolate a handful of linear combinations of features that "explain" most variances in the data. This is a descriptive, not predictive, technique, and it operates on the whole dataset without the training/testing division. Moreover, PCA is more informative if all features are suitably normalized, so no single feature can dominate the total variance. We then use Box Cox transformation on each column (the library chooses an appropriate parameter, different for each column, to make the resulting transformation approximately Normal.) The exception is the funding raised, which we use the log transformation (which is also a special case of Box Cox). This is justified because our plot shows that log(funding) looks Normal, and when we predict log(funding), reversing the function to get funding is more expedient. Our PCA shows that only a few (aggregated) features explain most of the variance. 3 top features explain 95% of the variance, while 5 top features explain 98% of the variance. The python notebook for PCA is here.

**Predictive Modeling**

* **Linear Regression**

**SVR**

We suspect that the problem is not linear, so we turn to Support Vector Regression (SVR). We split the data into the training data and the testing data, standardize the numerical features of each of the two datasets separately. We try three choices of kernels - rbf, linear and polynomial. For each choice of kernal, we use GridSearchCV with 5-fold cross validation to find the optimal parameters of the predictor over a reasonable (pre-determined) range of parameters. We then fit the predictor to the training data, predict it on the test data, and evaluate it by computing RMSE on log(funding). We found that rbf predictor with C=100 and gamma=0.01 is the best, with RMSE around 1. This result is comparable to linear regression. The python notebook for PCA is here (same notebook as PCA).

**System Requirement Specification**

A software requirements specification is a description of a software system to be developed it lays out functional and non- functional requirements, and may include a set of use cases that describe the interaction that the software must provide.

software requirements specification establishes the basis for an agreement between customers. and contractors or suppliers (in market-driven projects. These roles may be played by the marketing and development divisions) on what software product is done as well as what is not expected to do, software requirements specifications permit a rigorous assessment of requirements Before the design can begin and reduces later redesign. It should also provide a realistic basis for estimating product costs, risks, and schedules

**System Specification:**

* **R**

R is a free and open source software programming language and software environment for statistical computing and graphics. Distributed under the GNU General Public License version 2, R is an easy language to learn and commonly used for developing data analysis and statistical software. R compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.

R is designed to allow users to add additional functionality by defining new functions. Much of the system is itself written in the R dialect of the S programming language. For computationally-intensive tasks, C, C++ and Fortran code can be linked and called at run time. Advanced users can write C code to manipulate R objects directly. This programming language was named R, based on the first letter of first name of the two R authors (Robert Gentleman and Ross Ihaka), and partly a play on the name of the Bell Labs Language S.

* **Python**

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed. Python has been built with extraordinary Python libraries that are used in Big Data every day for solving problems. which are following

* TensorFlow
* NumPy
* SciPy
* Pandas
* Matplotlib

**Attributes in Data set:**

|  |
| --- |
| Start-up Name |
| Start-up Date/Month/Year |
| Industry Vertical |
| Sub Vertical |
| City Location |
| Investors Name |
| Investment Type |
| Amount in USD |