
**Independent University, Bangladesh**

# Social Network Analysis

An undergraduate internship report submitted by

Md. Nahin Rifath

**ID: 1720178**

**Autumn, 2020**

**Supervisor: Md. Fahad Monir**

**Department of Computer Science & Engineering**

**Independent University, Bangladesh**

**Dissertation submitted in partial fulfillment for the degree of Bachelor of Science in Computer Science**

# Letter of Transmittal

Md. Fahad Monir

School of Computer Science and Engineering

Independent University Bangladesh

Subject: Submission of Internship Report

Dear Sir,

It is a great pleasure for me to present the internship report on “Social Network Analysis” under the Data Science Department of Cramstack Ltd. For the completion of my Bachelor of Computer Science Degree, I got this as my assignment. I am glad to inform you that, I have successfully completed my 12 weeks of Internship at Cramstack Ltd, under the supervision of K M Jawadur Rahman, Lead Data Scientist. It was a great experience for me to work at Cramstack Ltd. I am extremely grateful to you for your guidance and kind operation on this report. I would be grateful if you kindly go through my report and evaluate my performance.

I pray and hope this report will be quite interesting and fulfil your expectations. I have tried my best to avoid my deficiencies and hope that my report will satisfy you. I also would like to thank you again for giving me the opportunity to submit this report.

Sincerely,

Md. Nahin Rifath

ID- 1720178

# Letter of Endorsement

To Whom It May Concern

Subject: Approval of the Report

This letter is to certify that all the information mentioned in this document is true and confidential to the company. The project mentioned here has successfully involvement of Md. Nahin Rifath, Bachelor's in Computer Science, Independent University, Bangladesh (IUB).

I wish him all the best and hope he will lead a successful career.

Internship Supervisor
Signature ……
K M Jawadur Rahman
Lead Data Scientist
Cramstack Ltd

# Approval

This report entitled

Data Science, Intern

by

Md. Nahin Rifath

Has been approved by

The Department of Computer Science and Engineering

Independent University, Bangladesh (IUB)

Supervisor: **Md. Fahad Monir**

The final copy of this report has been examined by the signatory and I find that both the contents and the form meet acceptable presentation standards of scholarly work in the above-mentioned discipline.

Signature ……

Date ……

# Attestation

This is to certify that the report titled “Social Network Analysis” was completed by Md. Nahin Rifath (ID-1720178) submitted in partial fulfillment of the requirement for the Degree of Computer Science from Independent University, Bangladesh (IUB). It has been completed under the guidance of Md. Fahad Monir (Internal Supervisor) and K M Jawadur Rahman (External Supervisor). I also certify that all my work is original and has not been submitted earlier to this university or any other institution. All the sources of information used in this Project Report has been duly acknowledged in it.

Signature

……..

(Md. Nahin Rifath)

# Acknowledgement

At the very beginning, I would like to thank the Almighty Allah for all His blessings which helped me to complete this report successfully.

I would also like to thank everyone who kindly provided me with information and gave me guidance for making this report. At first, I would like to thank Department of Computer Science, Independent University, Bangladesh for enlightening me over the period of my bachelor's in computer science.

I would like to express my gratitude to my honorable supervisor Md. Fahad Monir from the core of my heart for his kind support, supervision, instructions, and advices for the completion of this report.

I am also thankful to the whole team of "Cramstack Ltd" for giving me the opportunity to work with them in their data science department and providing me with the data and insights that were required for making this report. Also, I would like to thank K M Jawadur Rahman, Lead Data Scientist for guiding me all through the program.

Finally, I would like to thank all the faculties and mentors throughout my 4-year bachelor's in computer science program at Independent University, Bangladesh. All these helped me to get a much better view about the present world and to overcome any challenge given to me.

# Abstract

This report reflects the firsthand experience and knowledge I gathered during my tenure at Cramstack Ltd. Though my work was solely based on the data science department, but I got the opportunity to collaborate and work closely with the web development team as well. As a result, I gathered data and tried to analysis the different social media besides Instagram.com. Among these mediums, Instagram.com was one of the crucial one when it came to analyze the data. In this report I analyzed the reason behind the inefficiencies and the back draw we faced while trying to analyze the data collected from Instagram.com. Also, not having similar features like other social media platforms like Facebook, Twitter, or LinkedIn; made it more challenging for us to utilize Instagram to its full potential. I also realized that Instagram.com has only catered to the needs of a niche target audience at Bangladesh. Among these group of people, most of them are either posting their photos as part of their lifestyle or established their position in somewhere which has made them come to the platform in hope of extending their network more. The report mainly deals with overall experience on Instagram and the challenges we faced and some suggestions to overcome those challenges as well. The whole analysis of the report was done using both qualitative and quantitative measures. Before coming to any conclusion from this report, it should be kept in mind that, it was prepared in a very short period, there were difficulties in finding the data as Cramstack Ltd. is a start-up business and last but not the least, the analyze time is small which has been the biggest obstacle for reaching a conclusion. However, the report may still be useful for designing any further study on effectiveness of data analysis on a platform like Instagram.

# Contents

# Chapter 1

# Introduction

## 1.1 Overview

"Social media" is a way for people to communicate and interact online. The evolution of social media has been fueled by the human impulse to communicate and by advances in digital technology. It is called social media because users engage with (and around) it in a social context, which can include conversations, commentary, and other user-generated annotations and engagement interactions.

Recently, social media has become the platform for interaction among the population of Bangladesh in which they generate, share and exchange information and ideas in digital networks and communities. Social media have been defined in various ways.

The world right now is very tight and fast. Businesses must act quickly to take advantage of whatever appropriate they can. The ability to analyze and act on data is increasingly important to businesses. The pace of change requires companies to be able to react quickly to changing demands from customers and environmental conditions. Although prompt action may be required, decisions are increasingly complex as companies compete in a global marketplace.

Data must be dynamic to ease and maximize user experience. Data can do as such a great deal more for us now than it ever could. With the correct information nearly, anything is conceivable. Information from each field possible is being attempted and tried with machine adapting, frequently with marvelous outcomes.

For businesses, the shift in web consumerism and accompanying rise in social media brings both opportunity and responsibility. The sheer amount of data that customers make available through social media alone has web marketers jumping for joy. The real magic, however, lies in the

opportunity to grow lasting and scalable relationships with organization's customer base through social media. This is also where a business's online responsibility to customers begins to take shape. Whether a business is listening and engaging or not, customers are having conversations relevant to its operations.

## 1.2 Origin of the Work

This paper is prepared for fulfillment of the module "Internship" to conclude bachelor's in computer science program. This paper is done under the supervision of Md. Fahad Monir, Lecturer, Independent University, Bangladesh, and external supervision of K M Jawadur Rahman, Lead Data Scientist, Cramstack Ltd. The purpose of this internship report is to relate intern's working experience at Cramstack Ltd as a data science intern.

## 1.3 Background

Over the last several years, there has been an explosion of growth in popular social media platforms like Facebook, Twitter, Google+, LinkedIn, YouTube, Pinterest, and many others. It is safe to say that the era of social media is just getting started, and the need for social media in business will only become stronger over time. The whole world has seen the impact of the expansion and adoption of social media tactics, and the rising stats speak for themselves. The scenario is getting in the same place in Bangladesh too. The numbers of people getting connected to internet are growing day by day here. The rapid growth in online and mobile users in Bangladesh caught the attention of the marketers and the brands. It has not been of much of time when the digital marketing was only limited to email marketing. Today the brands seek to stand out in the market by using social media rigorously. Through this report it will be understood the data collection and data analysis of Instagram.com. However, some fundamental information has been excluded from the report due to organizational policy of Cramstack Ltd.

## 1.4 Objectives

1.4.1 Major Objectives: The primary objective of the study is to analyze the data of Instagram.com. There are some specific objectives also.

1.4.2 Specific Objectives:

- Understand digital media landscape of Bangladesh
- Collect data from Instagram.com
- Process the collected data
- Analysis of processed data

## 1.5 Scopes

Many leading international and local firms have not yet incorporated brand communication through Instagram.com for their brands, or promotion of their products in social media not because of monetary matter, but for having inadequate awareness of the benefits of social media for brand communication. But it is hoped that, in near future more and more firms will start this new era of marketing and brand communication through social media. This report is intended to provide a detailed overview of the data collection, data processing and data visualization and data analysis of Instagram.com for various clients on behalf of Cramstack Ltd.

# Chapter 2

# Company Profile

## 2.1 Company Background

Cramstack is a data mining and data analytics startup organization that provides a platform to look up valuable data and organize it in the way the user desires. Its predictive modelling uses data mining and probability to forecast outcomes with assistance from Artificial Intelligence. The goal of the company is to reduce the bureaucracy around business data and make data access simple and quick. Cramstack has identified a few keys to its success. The first is the need to only offer top-notch solutions which are based on market demand. The second is providing clients with both solutions and value creations. The final one is to ensure that all its offerings are based on economic justifications and helping the clients increase their own growth potential.

## 2.2 Vision and Mission

With implementation of various business technologies, companies are accumulating more and more data every day. One of the most primary goals of these data is to get insights from it and use those insights to make data driven decisions. But in many cases, the data just sits idle. It takes a whole lot of people, skill, expertise, and most importantly time to utilize that data. Even if the company has dedicated Business Intelligence teams, they work hours and days to get the insights executives want. And there are so many ways this process can go wrong, and the executives might miss out on huge business opportunities. Even though, data are accumulated, they are not utilized properly in the current, conventional methods.

Cramstack's Natural Language Query platform works as a 'Google for data.' With this search engine, executives can look for specific information and insights from millions of stacks instantaneously. No technological expertise is needed to use this platform. Anyone with

permission can use it by searching for information in plain English from the data source. The user can have a customized dashboard to share and collaborate with other colleagues.

## 2.3 Services

Cramstack help clients to build and visualize data warehouses so they can generate insights from day one.

- **Extract, Transform, Load (ETL):** To facilitate business analysis, clients need to load data warehouse regularly. Cramstack find and connect client's multiple sources of data in a streamlined solution through the ETL process.

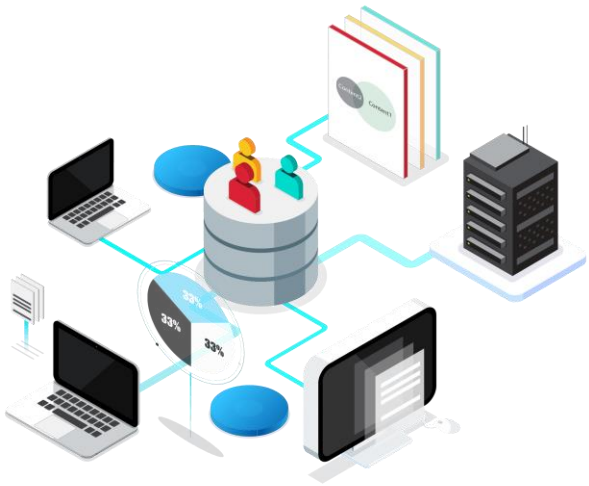


Figure: 2.3.1

- **Visualization:** Visualization leads to the understanding of client's data and adaptation on multiple user levels. Clients will find their business datasets of information as a chart or other images in a meaningful way.

Figure: 2.3.2

- **Advanced Analytics:** Cramstack's advanced analytics uses high- level methods and tools to focus on projecting future trends, events, and behaviors. Cramstack perform data sufficiency pilots to lets clients know where they can go to next with the data.

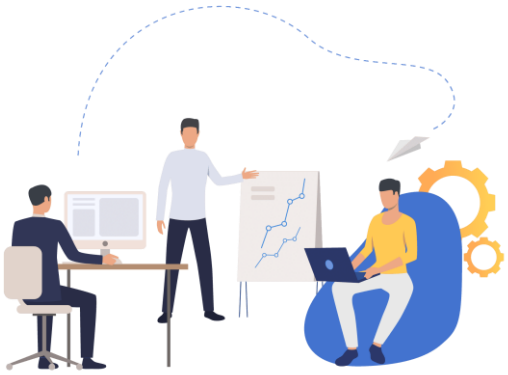


Figure: 2.3.3

## 2.4 Technologies Supported


Figure: 2.4.1

Figure 2.4.1 shows all the technologies supported for development in Cramstack. This includes **Microsoft SQL Server, MySQL, PostgreSQL, Oracle, Sybase,** and **Amazon Aurora.** Cramstack also use **PHP, Laravel, Microsoft .Net Framework, Android, iOS, JavaScript, Node.js, Vue.js, AngularJS, React** and **Python.**

## 2.5 Contact and Address

House 31, Road 6, Block: C, Banani

Dhaka 1213, Bangladesh

Phone Number:

- +88-01844506750
- +88-01844506752
- +88-01844506753

Email: info@cramstack.com

Website: www.cramstack.com

# Chapter 3

# Literature Review

In today's technology driven world, social networking sites have become an avenue where retailers can extend their marketing campaigns to a wider range of consumers. Chi defines social media marketing as a "connection between brands and consumers, [while] offering a personal channel and currency for user centered networking and social interaction" [12]. The tools and approaches for communicating with customers have changed greatly with the emergence of social media; therefore, businesses must learn how to use social media in a way that is consistent with their business plan. This is especially true for companies striving to gain a competitive advantage. This review examines current literature that focuses on a retailer's development and use of social media as an extension of their marketing strategy. This phenomenon has only developed within the last decade, thus social media research has largely focused on defining what it is through the explanation of new terminology and concepts that makeup its foundations and exploring the impact of a company's integration of social media on consumer behavior [13] [14].

## 3.1 Data Science

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from many structural and unstructured data [1][2]. Data science is related to data mining, machine learning and big data.

Data science is a "concept to unify statistics, data analysis and their related methods" to "understand and analyze actual phenomena" with data [3]. It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, domain knowledge and information science. Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational, and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge [4][5].

## 3.2 Data Collection

Data collection is the process of gathering and measuring information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes. Data collection is a research component in all study fields, including physical and social sciences, humanities, and business [6]. While methods vary by discipline, the emphasis on ensuring accurate and honest collection remains the same. The goal for all data collection is to capture quality evidence that allows analysis to lead to the formulation of convincing and credible answers to the questions that have been posed.

## 3.3 Data Processing

Data processing is, generally, “the collection and manipulation of items of data to produce meaningful information” [7]. In this sense it can be considered a subset of information processing, “the change (processing) of information in any manner detectable by an observer”.

## 3.4 Machine Learning

Machine learning (ML) is the study of computer algorithms that improve automatically through experience [8]. It is seen as a subset of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data", to make predictions or decisions without being explicitly programmed to do so [9]. Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

## 3.5 Neural Language Processing

Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, how to program computers to process and analyze large amounts of natural language data. The result is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.

## 3.6 Data Analysis

Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, and is used in different business, science, and social science domains. In today's business world, data analysis plays a role in making decisions more scientific and helping businesses operate more effectively [10].

## 3.7 Social Media Analysis

Social media analytics is the process of gathering and analyzing data from social networks such as Facebook, Instagram, LinkedIn, and Twitter. It is commonly used by marketers to track online conversations about products and companies. One author defined it as "the art and science of extracting valuable hidden insights from vast amounts of semi-structured and unstructured social media data to enable informed and insightful decision making" [11].

## 3.8 Sentiment Analysis

Sentiment analysis (also known as opinion mining or emotion AI) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media.

## 3.9 Relationships with Undergraduate Study

I have attended the Artificial Intelligence, Data Mining and Numerical Methods courses during my study as an undergraduate student. I found some practical implication and similarities from the courses during my internship at Cramstack Ltd. which will be a great experience for me.

## 3.10 Related Works

The method of sentiment analysis can be differentiated in two main strategies: lexicon based and machine learning based technique. Before the analysis of an unknown dataset the machine learning based algorithm must be trained by a training data set. For the lexicon-based approach, the sum of the polarities for each word or phrase is the polarity of the document combine the methods of lexicon as well as machine learning based methods to improve precision and get a high recall [24] [25]. Kaushik and Mishra found a lexicon-based approach for sentiment analysis that works fast [24]. And, finally, Nielsen evaluated a word list for sentiment analysis in microblogs [21]. There are already several studies about sentiment analysis on Twitter posts [22] [24] [25] and product reviews examined if the uploaded media and comments about soccer games on social media are more negative when there is violence during a soccer match [15] [16] [19]. Therefore, they analyzed the emotion of Instagram photos and videos as well as comments of posts. Unlike to our approach, they worked with a machine learning based technique for the comments. Cyber hate was detected on Instagram using a snowball method [24]. They collected data from pictures and videos of 25,000 public Instagram accounts, including the comments of posts. Each post was manually checked for cyberbullying or hyperaggressive behavior and labeled accordingly. As result, they found that users who get bullied in social media gain less likes for the posted media but more frequent comments.

# Chapter 4

# Project Management and Financing

## 4.1 Work Breakdown Structure

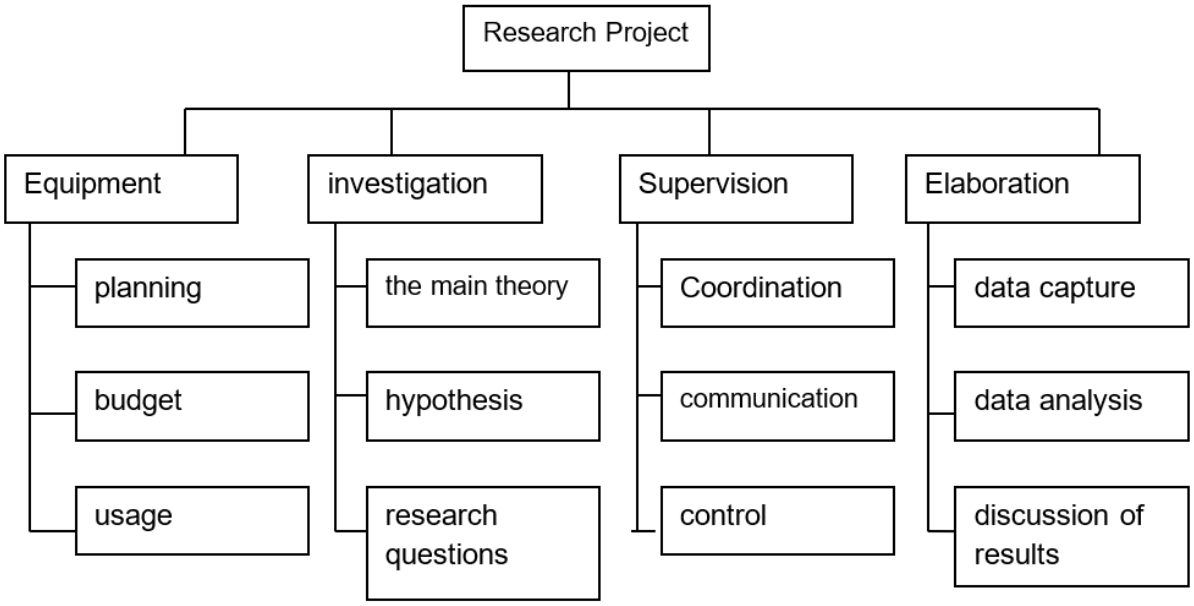



Figure 4.1.1: Work Breakdown Structure

## 4.2 Gantt Chart

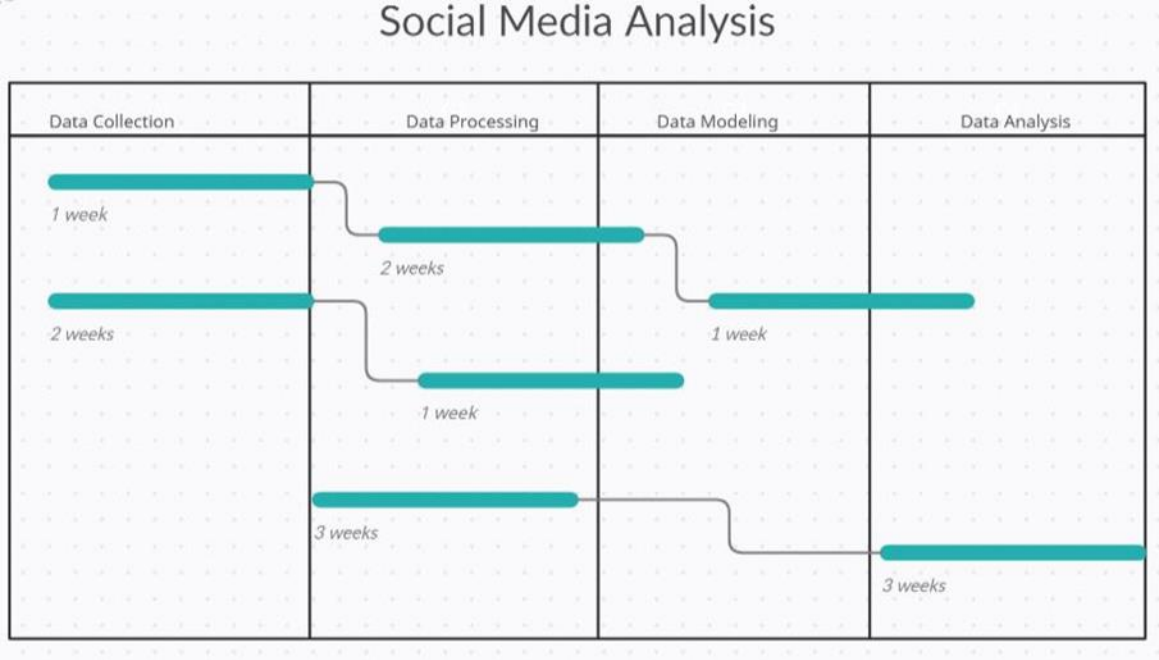



Figure: Gantt Chart

## 4.3 Process Activity Wise Resource Allocation

**1. Prediction error or model's uncertainty:** The impact of wrong predictions could be high. For instance, a wrong scheduling decision might cause violation of Service Level Objective (SLO) and may incur monetary penalties.

**2. Cost of training:** To train these data-driven models, we need to collect enough training data. The cost in terms of time and/or money spent on collecting such data could be non-trivial.

**3. Generalization:** To enable adoption of the models in real system deployments, we need to ensure that we build models that can generalize from benchmark workloads to real user workloads.

## 4.4 Estimated Costing

1. **Base cost:** The base cost of the whole project is around 5 lakh Bangladeshi Taka.

2. **Financial charges during implementation:** The Financial charges during implementation is around two lakh Bangladeshi Taka.

# Chapter 5

# Methodology

Sentiment analysis in social media is different from “classical” sentiment analysis of newspaper articles, for instance [23]. Here, we have text and we have additionally emojis. A sentiment analysis in Instagram is virtually new scientific territory. We were only able to identify very few approaches of sentiment analysis of Instagram hashtags and Instagram texts [20]. We conducted for the first time a lexicon-based sentiment analysis of Instagram post’s comments with a very large data base. First the required data (comments) have been collected and preprocessed. Afterwards, the sentiment analysis could be performed on over 660,000 comments.

## 5.1 Data Collection

The data were collected from the beginning of November 2020 until the beginning of January 2021 via the official Instagram API. It took place before the new Instagram API principles were realized. As a result of the Instagram API’s security measure, it was only possible to obtain the first 150 comments of each picture or video. Since the official account was the largest data source, not all comments of every picture and video could be retrieved. The database consisted of approximately one million records after the data extraction.

## 5.2 Preprocessing

Before analyzing the collected data, they had to be preprocessed by a python script. Spam such as chain mails, advertisements, or comments with limited content like “first” (user expressing one is the first to comment on the picture or video) got deleted. Usernames and links in the comments were reduced to a more general term, namely “USERNAME” and “LINK”, without having an impact on the sentiment. Also, the language of the comments was checked and automatically translated to English. Replacing abbreviations with their actual term was not required in this investigation due to repeating characters having emotionality themselves. After eliminating useless comments and cleaning the data, the sentiment analysis was performed on approximately 660,000 remaining records.

## 5.3 Sentiment Analysis

In our study, the sentiment analysis is used to identify, extract, and analyze the opinions and feelings of the comments written under media relating to a celebrity. The following approach detects the sentiment strength (positive, neutral, and negative) within an interval of -5 to +5 (from negative until positive). Sentiment strength of 0 is considered as a neutral sentiment. Using SentiStrength as a model, the Python based sentiment analysis program consists of an emoticon list, an emotion lexicon, a negation lexicon, a lexicon for booster words like "very" or "totally" as well as a lexicon for phrases. AFINN is a list of English words rated for valence with an integer between -5 and +5. An adapted version of AFINN-111 is used as the emotion lexicon in this sentiment analysis. Two words, "like" and "lie", need a special treatment, because the lexicon itself cannot deal with ambiguity problems. As a solution, the Natural Language Toolkit (http://www.nltk.org/) for Python programs is used. With a POS-Tagger, the right part-of-speech is recognized, which leads to more correct sentiment word values. The sentiment analysis program operates different steps and assigns the final sentiment. Each comment gets a sentiment for the written text as well as one for the emoticons – those were combined to the final sentiment of the comment. First, the comment gets tokenized into sentences and next the sentences into words. To calculate the text sentiment, each word gets a sentiment value from the emotion lexicon. Words in quotation marks are considered as quotes and assessed as neutral because they often do not reflect the users' emotionality. Phrases that are present in the phrase lexicon get the sentiment value of that phrase. If the words of the phrase appear in the emotion lexicon as well, only the phrase value is important for the final comment sentiment. Also, the other lexicons were checked for negotiations (which can change the sentiment of a word from positive to negative, e.g., "not very happy") and booster words like "very". All those sentiment values add up to the final text sentiment of a comment. Because emoticons show a facial impression and therefore an emotion, it is important to include them into an emoticon lexicon. Further included are a few emojis that do not show a face but also express an emotion, for example a heart. The final sentiment is then calculated with all resulting values. Besides the text and emoticon sentiment values, there are also some other aspects considered in the final sentiment like repeated punctuations, repeated characters, number of emoticons and whole sentences or words in uppercase, all of them expressing emotion. After the sentiment analysis, each final sentiment of a comment is normalized to an interval of -5 to 5.

$$Normalized\ value\ = \frac{(5 * sentiment)}{max(|sentiment|)}$$

# Chapter 6

# Body of the Project

## 6.1 Work Description

1. **Using the features from Bag-of-Words for training set**
   We have one dataset with features from the Bag-of-Words model and another dataset with features from TF-IDF model.

2. **Splitting the data into training and validation set**
   First task is to split the dataset into training and validation set so that we can train and test our model before applying it to predict for unseen and unlabeled test data.

3. **Fitting the Logistic Regression Model**
   We are using the logistic regression model.

4. **Importing f1_score from sklearn**
   We are using the F1 Score throughout to analyze our model's performance instead of accuracy.

5. **Predicting the probabilities**
   The output basically provides us with the probabilities of the Instagram comments falling into either of the classes that is Negative or Positive.

6. **Calculating the F1 score**
   We will calculate the F1 Score throughout to analyze our model's performance instead of accuracy.

7. **Fitting the Decision Tree model**
   The last model we are using for a result is the decision tree model.

## 6.2 System Analysis

### 6.2.1 Six Element Analysis

#### 1. Hardware

Several computers were used as the part of the project. The processor of my computer (which was allotted for me during my internship at Cramstack Ltd.) was intel $10^{th}$ gen i5 10600 which was a hexa-core processor with 12 multithreads. The graphics card was NVIDIA GeForce RTX 3080.

#### 2. Software

**i. The Jupyter Notebook:** The Jupyter Notebook is an incredibly powerful tool for interactively developing and presenting data science projects. A notebook integrates code and its output into a single document that combines visualizations, narrative text, mathematical equations, and other rich media. This intuitive workflow promotes iterative and rapid development, making notebooks an increasingly popular choice at the heart of contemporary data science, analysis, and increasingly science at large.

**ii. Google Chrome:** Google Chrome is a cross-platform web browser developed by Google. It was first released in 2008 for Microsoft Windows, and was later ported to Linux, macOS, iOS, and Android where it is the default browser built into the OS.

**iii. Office 365:** Office 365 is a line of subscription services offered by Microsoft as part of the Microsoft Office product line. The brand encompasses plans that allow use of the Microsoft Office software suite over the life of the subscription, as well as cloud-based software-as-a-service products for business environments, such as hosted Exchange Server, Skype for Business Server, and SharePoint, among others.

#### 3. Data

The dataset we are going to use and how the train and test are divided with one being labeled and another being unlabeled respectively with the number of tweets present in each of the dataset.

**4. Procedures**

The software management team, the product management team and the data science department are following standard procedures. As for the collection of data, we are maintaining standard legal procedures and legal Instagram API. The research part had been supervised by my supervisor ensuring that the work is original. All the methods we used in scripts were legal and harmless.

**5. People**

I am working under my supervisor, K M Jawadur Rahman Lead Data Scientist of Cramstack on this project. The lead software engineer is also leading this project along with the software department and the product management team.

**6. Communication**

We used phone calls from the official sim number, skype and other communication tools to communicate ourselves in Cramstack Ltd.

### 6.2.2 Feasibility Analysis

We are proposing that project-specific case studies be used to obtain more detailed information on rigorous study designs, or outcomes. The criteria used to select project would depend on the component to be studied. Where possible, the project that satisfy multiple criteria will be selected (e.g., a project with an effective partnership component that had a rigorous study design). However, we expect that many of the project included in the case study component will only satisfy the criterion for one study component. We recommend that the case studies include interviews with a variety of project staff, as well as program participants/beneficiaries and, as relevant, leaders in the community. We are proposing that multiple data collection strategies be used. The remainder of this section presents greater detail on our recommended approaches for using these data collection strategies to address the four major evaluation questions.

### 6.2.3 Effect and Constraint Analysis

The role of mathematical optimization is to determine a set of decisions or actions that gives rise to the best possible results within the context of the stochastic models of the system of interest and subject to various constraints. More specifically, a general formulation of a single-period decision-

making optimization problem can be expressed in terms of minimizing or maximizing an objective functional of interest subject to various constraint functionals. The objective functional and constraint functionals define the criteria for evaluating the best possible results with respect to the decision variables and other dependent variables, where these and related variables are based on the stochastic models of the system of interest. The relationships among these components of the optimization formulation are critically important and often infused with subtleties and complex interactions.

Hence, mathematical optimization generally renders solutions that identify a set of decisions or actions at the start of the time horizon or identify a set of dynamic decision-making policies for dynamic adjustments to decisions or actions throughout the time horizon adapted to filtrations, in both cases having the goal of achieving the best possible results within the context of the stochastic models of the system of interest and subject to various constraints. Various mathematical methods can be used to obtain these solutions based on the properties of the stochastic models of the system and its underlying decision processes. The most appropriate methods will often depend on the complexity of the underlying stochastic models of uncertainty and the details of the formulation of the optimization problem of interest within the context of the stochastic models. The domain knowledge needed for this area spans stochastic processes, probability theory, optimization theory, control theory, and simulation theory.

### 6.2.4 Problem-Solution Analysis

1. Problem: "*Is the data available to solve this problem?*"

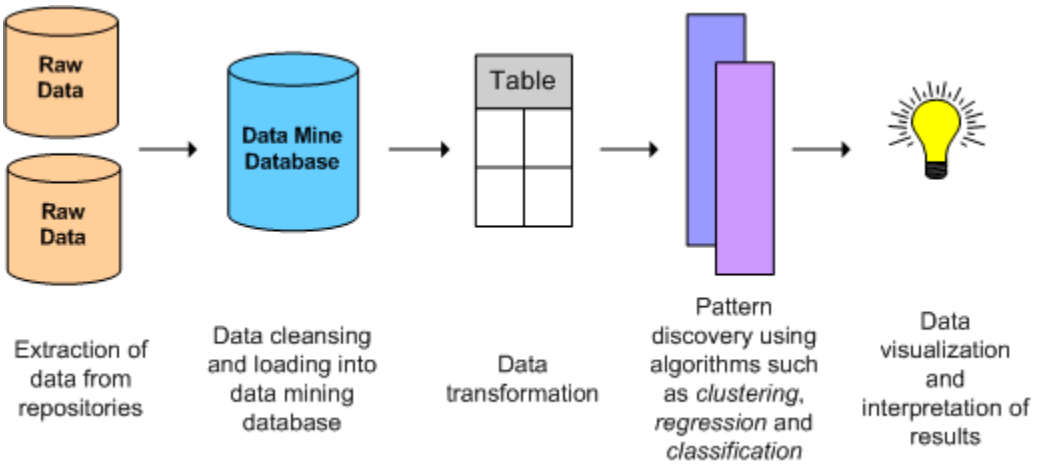
Solution: A data scientist without data is not a very helpful individual. Many of the data science techniques that are highlighted in media today — such as deep learning with artificial neural networks — requires a massive amount of data. A hundred data points is unlikely to provide enough data to train and test a model. If the answer to this question is no, then we can consider acquiring more data and pipelining that data to warehouses, where it can be accessed later.

2. Problem: "*Who are the team members we need in order to solve this problem?*"

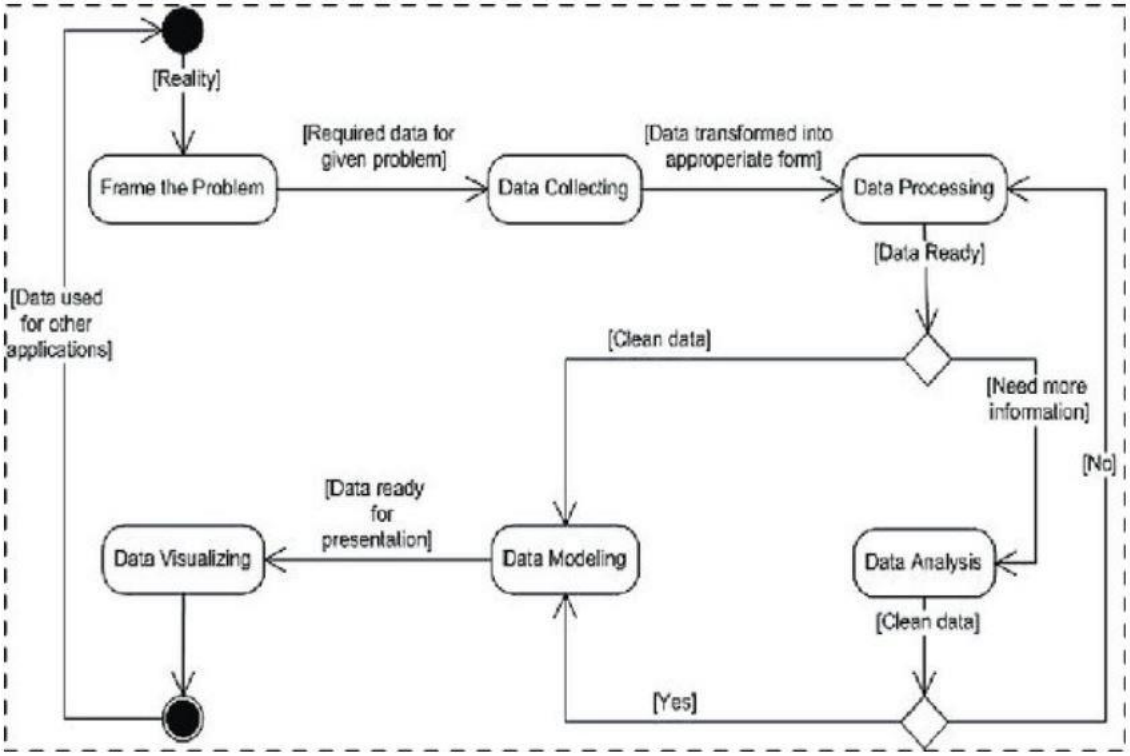
Solution: Our initial answer to this question will be, "The data scientist, of course!" Most of the problems we face at Cramstack Ltd cannot or should not be solved by a lone data science intern because we are solving business problems. Our data scientists team up with developers and project managers and hardware developers to develop digital strategies and solving data science problems is one part of that strategy. Solving our problem and solving your data scientists is not helpful for anyone.

## 6.3 System Design

### 6.3.1 Rich Picture



### 6.3.2 UML Diagram



### 6.3.3 Functional Requirements

- Supports both Python (Jupyter Notebook) and Laravel.
- Send information from Jupyter Notebook to Laravel Application using webhook.
- Execute Laravel Jobs after information is received via webhook.
- Laravel Job to clean data that is modeled after the SQL raw queries.
- The whole thing should be able to operate in cross-platform, in different operating systems.

### 6.3.4 Non-functional Requirements

- Properly timed and actions executed in a sequential pattern.
- Good performance and lightweight on the server.
- Modular, so that it is easy to edit and maintain.

# Chapter 7

# Result & Analysis

F1 Score is needed when we want to seek a balance between Precision and Recall. F1 score might be a better measure to use if we need to seek a balance between Precision and Recall and there is an uneven class distribution (large number of Actual Negatives). In one scenario, for the scale 0 to 1, f the score is nearer to 1, we are calling it positive and if the score is below 0.5 and nearer to 0, we are calling it negative. To be noted scores and parameters were increased during this project.

Our goal of this project is two-fold. First, extract all the words used in the captions of the images uploaded by the user including ‘Hashtags’ and ‘Emoji’ from his Instagram account. Second, to evaluate the effectiveness of the features from Instagram comments and to get the frequency of each distinct word used by the user in all the images. The algorithm used in the Instagram Analysis displays the number of comments over which the analysis has been done, number of words in the captions along with the hashtags and emoji, the words in the sequence in the way they were uploaded and the last the frequency of the words that how many times a word has been used by the user in his profile.

# Chapter 8

# Project as Engineering Problem Analysis

## 8.1 Sustainability of the Project

The sustainability report with this qualitative standard content brings specific analytical difficulties because its textual database. To support the analysis, this study discusses an analysis that is rarely applied in the sustainability report study, namely the choice of the use of sentiment analysis methods in it. Sentiment analysis or commonly referred as opinion mining techniques, is related to diverse and multidisciplinary artificial intelligence problems, to minimize gaps between human and computer. Sentiment analysis will find content and even regulate the client's ideas, likes, hatreds and desires by using complex language. Sentiment analysis will function as management, examination of feelings, sentiments, and intelligence of a writer or speaker in a few different specific texts.

This study finds that sentiment information on construction companies chose the use of words with positive sentiment and sought to show high accountability. That means the company shows more attention on the sustainability report disclosure.

## 8.2 Social and Environmental Analysis and Effects

Our project with focus on social, economic, and environmental issues will gain a competitive advantage and have a credible reputation in the public eyes. This kind of focus is shown on projects' sustainability report, as a concept that becomes important for businesses both at national and global levels.

This research contributes to the form of forecasting financial performance and supporting corporate stakeholder on decision-making processes which also detects fraud, manages risks, and predicts future performance. This research also can assist stakeholders to analyze and to make decisions related to economic, social, and environmental issues.

## 8.3 Addressing Ethics and Ethical Issues

In the contemporary digitalized age, data analytics have enabled organizations to automate and analyze multiple sources of data and information quickly such that it facilitates optimized decision-making process that help in achieving organizational goals. While, from a strategic perspective analyzing of the data for eventual analysis is vital, given the availability of varieties of data that can be accessed from multiple media sources makes big data management highly challenging. Moreover, given that data analyses data very fast, it enables access to data-information which could compromise (either inadvertently or deliberately) individual privacy, be misused, etc. raising ethical issues concerning the sharing and usage of data. To address these concerns on ethicality in big data management, this study proposes to use a simple 'stakeholders-ethics-framework' to develop a 'stakeholder analysis approach framework' suggestive be linked to sustainability guidelines that help towards a sustainable big data industry, is assumed.

# Chapter 9

# Future Works & Conclusion

## 9.1 Future Works

Future studies could take the current study a step further with examining responses from users with uses and gratifications theory. Behaviors, such as rating, saving, sharing, and commenting on content, are meant to fulfill a user's social needs. There is a chance that these interactions will decrease the user's loneliness, depression, isolation, and increase self-acceptance and acceptance by others. Again, socialites and celebrities of this status have millions of followers, but only thousands of comments, so there must be something motivating each user who takes the time to leave a comment. Another approach is to look at these findings from the perspective of identity development.

## 9.2 Conclusion

The internship has been a very fruitful and worthy experience for me. I was able to work, hands-on, in an industry that I had no prior knowledge about. The process of transforming the rich theoretical knowledge with the practical knowledge of the industry has dawned on me and driven to seek excellence in the craft of Data Science.

Interns do not usually get to work on live projects and contribute to the workflow of an ongoing project in the office. But the people at Cramstack Ltd, felt that I was worth giving a chance to and tasked me with such projects that would help me grow in every aspect of my career. Being the youngest there and the least experienced of the bunch, I got a plethora of advice from the people of the offices. I also learned the tools and techniques that were utilized by industry hardened software developers and engineers alike.

On top of that I was taught etiquettes of the corporate life and how to maintain proper rapport with my co-workers. These are the skills that cannot be learned using books and must be applied to assure proper implementation. It was a blessing for me to be in the presence of such good people who were willing to help me at each part of my journey through the internship.

In the end, I would like to thank both my internal and external supervisors whose guidance and motivations have persuaded me to strive for the success in this project and for the endless projects to come in my way in the future.

# Reference


[1] Dhar, V., "Data science and prediction", Communications of the ACM, 2013 pp. 64–73. [Accessed: 28- Nov- 2020]

[2] Jeff Leek, The key word in "Data Science" is not Data, it is Science, 2013. [Accessed: 28- Nov- 2020]

[3] Hayashi, Chikio, "What is Data Science? Fundamental Concepts and a Heuristic Example", 1998. [Accessed: 28- Nov- 2020]

[4] Tony Hey, Stewart Tansley, Kristin Michele Tolle, The Fourth Paradigm: Data-intensive Scientific Discovery, 2009. [Accessed: 28- Nov- 2020]

[5] Bell, G., Hey, T., Szalay, A., COMPUTER SCIENCE: Beyond the Data Deluge. Science, 2009. [Accessed: 28- Nov- 2020]

[6] Vuong, Quan-Hoang; La, Viet-Phuong; Vuong, Thu-Trang; Ho, Manh-Toan; Nguyen, Hong-Kong T.; Nguyen, Viet-Ha; Pham, Hiep-Hung; Ho, Manh-Tung, An open database of productivity in Vietnam's social sciences and humanities for public use, 2018. [Accessed: 28- Nov- 2020]

[7] French, Carl, Data Processing and Information Technology (10th ed.), 1996. [Accessed: 28- Nov- 2020]

[8] McGraw Hill, Machine Learning New York, pp. 23 1997. [Accessed: 28- Nov- 2020]

[9] John R.; Bennett, Forrest H.; Andre, David; Keane, Martin A., Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming, pp. 151–170, 1996. [Accessed: 28- Nov- 2020]

[10] Xia, B. S., & Gong, P., Review of business intelligence through data analysis, pp. 300-311, 2015. [Accessed: 28- Nov- 2020]

[11] Sponder, Marshall; Khan, Gohar F., Digital analytics for marketing, 2017. [Accessed: 28- Nov- 2020]

[12] Cheong, Hyuk Jun, and Margaret A. Morrison., Consumers' Reliance on Product Information and Recommendations Found in UGC, Journal of Interactive Advertising, pp. 8, 38-49. Chi, Hsu-Hsien, 2011. [Accessed: 28- Nov- 2020]

[13] Chu, Shu-Chuan, Interactive Digital Advertising VS. Virtual Brand Community: Exploratory Study of User Motivation and Social Media Marketing Responses in Taiwan., Journal of Interactive Advertising, 12, pp. 44-61, 2011. [Accessed: 28- Nov- 2020]

[14] Viral advertising in social media: Participation in Facebook groups and responses among college-aged users, Journal of Interactive Advertising, pp. 12, 30-43, 2018. [Accessed: 28- Nov- 2020]

[15] Boychuk, V., Sukharev, K., Voloshin, D. and Karbovskii, V., An Exploratory Sentiment and Facial Expressions Analysis of Data from Photo-sharing on Social Media: The Case of Football Violence, Procedia Computer Science, Vol 80, pp 398–406, 2016. [Accessed: 28- Nov- 2020]

[16] Cui, H., Mittal, V. and Datar, M., Comparative Experiments on Sentiment Classification for Online Product Reviews, Proceedings of the 21st National Conference on Artificial Intelligence, Vol. 2, pp. 1265–1270, 2006. [Accessed: 28- Nov- 2020]

[17] Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q. and Mishra, S., Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network",

International Conference on Social Informatics, pp. 49–66, 2015. [Accessed: 28- Nov- 2020]

[18] Dave, K., Lawrence, S. and Pennock, D.M., Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews", Proceedings of the 12th International Conference on World Wide Web, ACM, New York, NY, pp. 519–528, 2013. [Accessed: 28- Nov- 2020]

[19] Mukherjee, S. and Bhattacharyya, P., Feature Specific Sentiment Analysis for Product Reviews, International Conference on Intelligent Text Processing and Computational Linguistics, pp 475–487, 2012. [Accessed: 28- Nov- 2020]

[20] Nam, N., Lee, E. and Shin, J., A Method for User Sentiment Classification Using Instagram Hashtags", Journal of Korea Multimedia Society, Vol 18, No. 11, pp. 1391–1399, 2015. [Accessed: 28- Nov- 2020]

[21] Nielsen, F. A., A New: Evaluation of a Word List for Sentiment Analysis in Microblogs", Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts', pp 93–98, 2011. [Accessed: 28- Nov- 2020]

[22] Pak, A. and Paroubek, P., Twitter as a Corpus for Sentiment Analysis and Opinion Mining", Proceedings of the International Conference on Language Resources and Evaluation 2010, pp 1320–1326, 2010. [Accessed: 28- Nov- 2020]

[23] Pozzi, F. A., Fersini, E., Messina, E. and Liu, B., Sentiment Analysis in Social Networks, Morgan Kaufman, Cambridge, MA, 2017. [Accessed: 28- Nov- 2020]

[24] Kaushik, C. and Mishra, A., A Scalable, Lexicon Based Technique for Sentiment Analysis, International Journal in Foundations of Computer Science & Technology, Vol 4, No. 5, pp 35–43, 2014. [Accessed: 28- Nov- 2020]

[25] Khan, A. Z., Atique, M. and Thakare, V. M., Combining Lexicon-Based and Learning-Based Methods for Twitter Sentiment Analysis, International Journal of Electronics, Communication and Soft Computing Science & Engineering, pp 89–91, 2015. [Accessed: 28- Nov- 2020]