Application of DMAIC Framework on the Sentiment Analysis

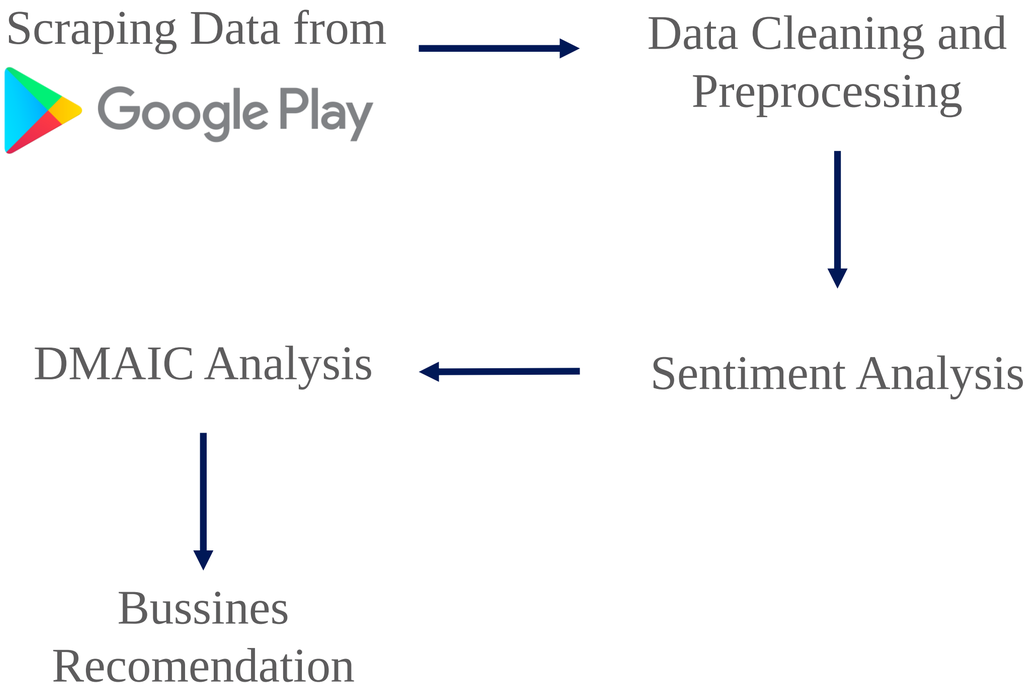
Executive Summary

It is important to pay attention to customer satisfaction. One form of service provided is the application that can be downloaded on google play store. Customer satisfaction can be seen by making sentiments based on opinions on application reviews. Through the opinions of buyers, e-commerce and sellers are able to improve the quality of their services. Good service quality will affect customer engagement and loyalty.

What are the results obtained from this case study??

- Public sentiment level by month.
- SVM classification model with 95% precision score, and 91% f1-score.
- Business recommendation by using DMAIC method.

Background



The customer is an important entity for the running of the company. This app is a service representation in the digital era. Knowing how the public sentiment on application satisfaction. This sentiment information can be used as an input in the analysis of the DMAIC method to get the right business recommendations to improve the application quality.

Dataset

Data Scraping

Scraping data using google play scraper. From the results of data scraping, 100,000 data were obtained for the period May 2020 - August 2021.

You can see the code for scraping in https://github.com/madekrisnaj/

Data Introduction

Review : Customer opinion.

Month : Month comment given.

Year : Year comment given.

Score : application ratings.

Data Preparation

Data Cleaning & Preprocessing

The data provided from google play store are not clean yet. We have to cleaned up and feature selection before going to the sentiment analysis.

Data cleaning and preprocessing procedure including :

- Lowercase
- Remove Number
- Remove Punctuation
- Remove Whitespace
- Remove ASCII and Unicode
- Remove Newline
- Stop words
- Stemming
- Vectorization

Feature Selection

From the variables we already have, we need to add one more variable. The variable is a value which is a label of positive or negative sentiment. So, now we have 5 variables to enter the data analysis.

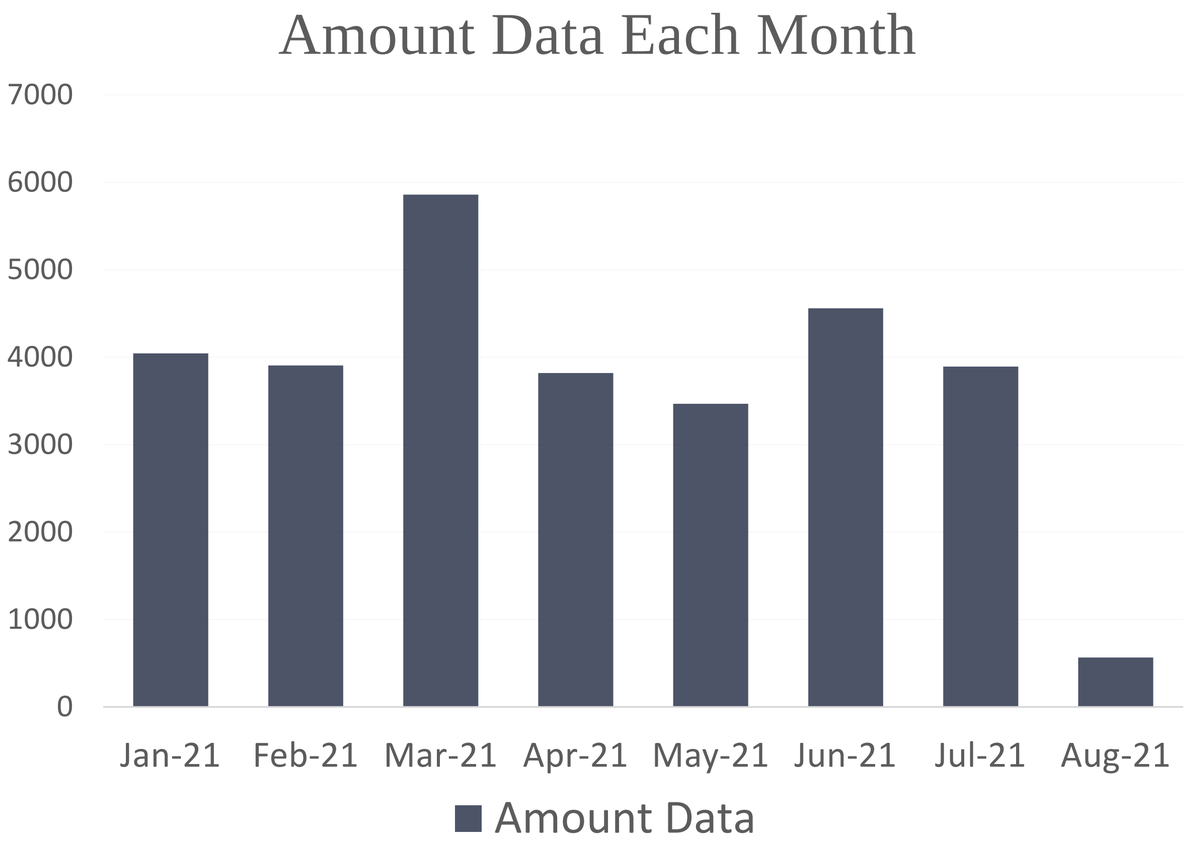
Data Preparation

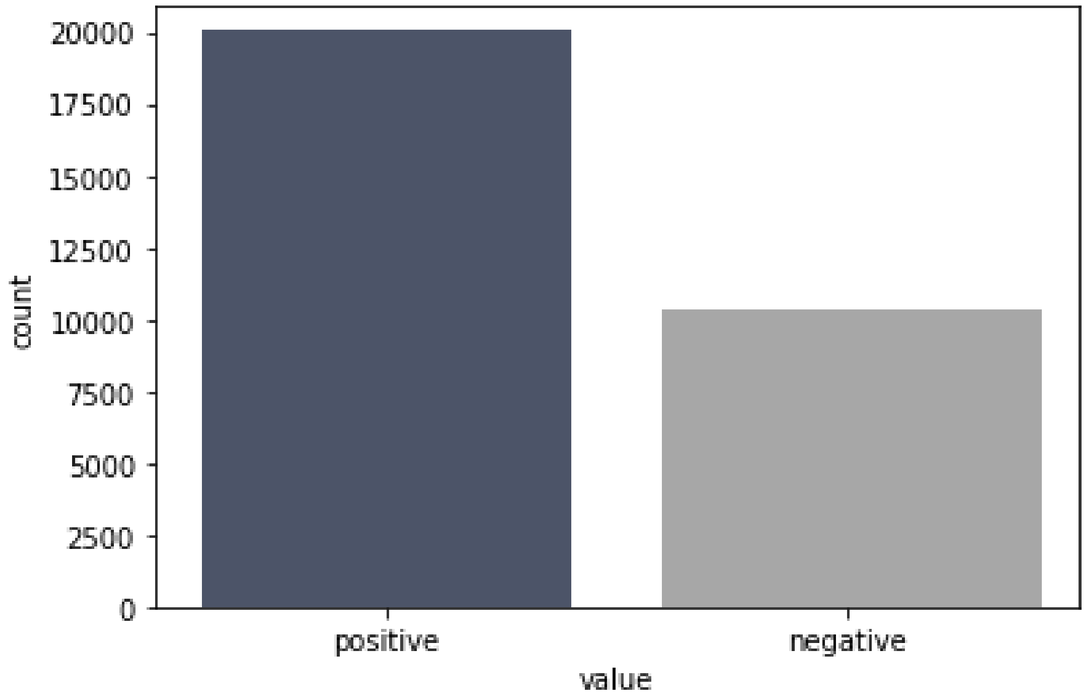
Labeling

- As initial labeling, scores 1-3 are classified as negative label and 4-5 are classified as positive label.
- Only 2 classification classes are used because the need for DMAIC is only 2 classes in the form of defective and non-defective data.

We only use data in **2021 period**. After going through the process of cleaning, stop words, stemming, remove missing values and filter data, the data used is **30,119 data.**

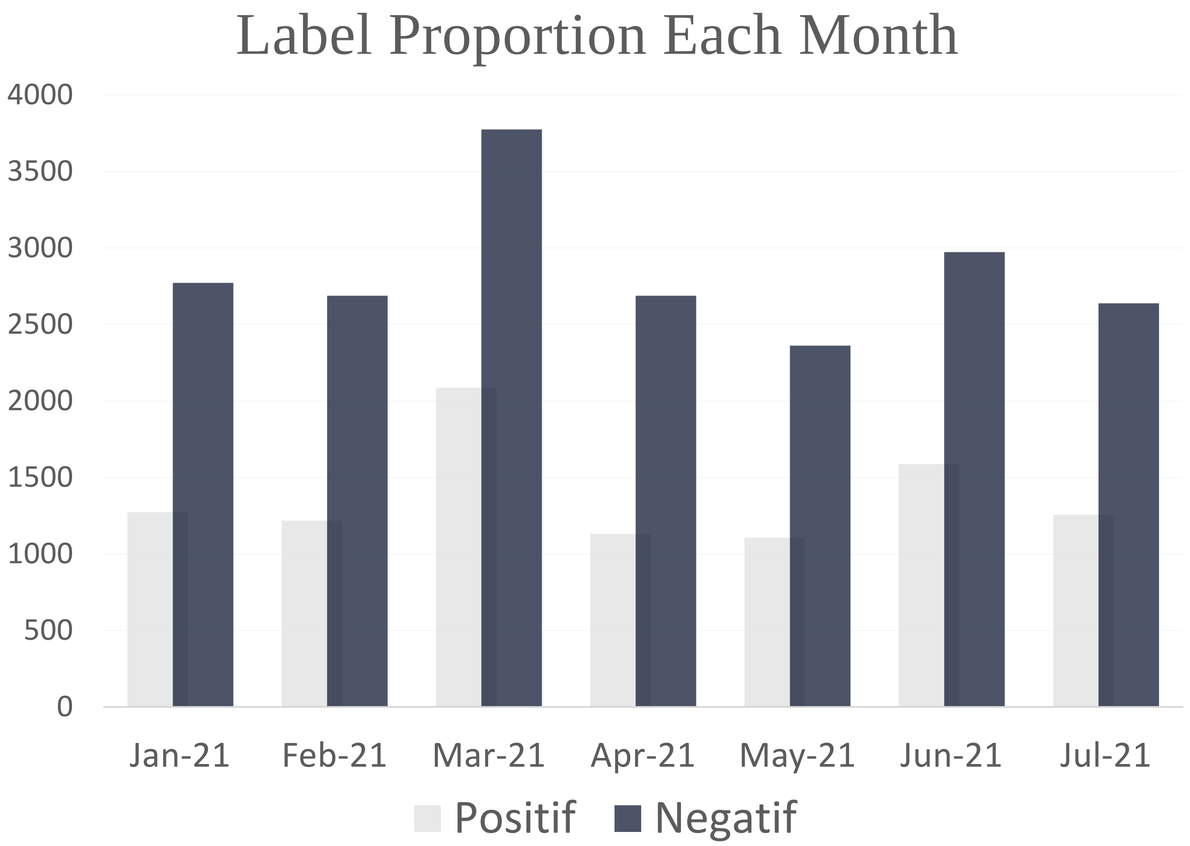
Exploratory Data Analysis





- The data used is only data for the period **January-July 2021.**
- August 2021 is not used because it is too few.

Exploratory Data Analysis



MonthName	Negative	Positive	Negative_proportion
21-Jan	1273	2772	31%
21-Feb	1218	2688	31%
21-Mar	**2086**	**3775**	**36%**
21-Apr	1131	2688	30%
21-May	1106	2362	32%
21-Jun	1587	2973	35%
21-Jul	1256	2638	32%

Sentiment Analysis

Sentiment Analysis

Sentiment analysis is contextual mining of text which identifies and extracts subjective information in source material, and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations. However, analysis of social media streams is usually restricted to just basic sentiment analysis and count based metrics. This is akin to just scratching the surface and missing out on those high value insights that are waiting to be discovered.

You can see the code for scraping in https://github.com/madekrisnaj/

Classification Method

Support Vector Machine (SVM)

Support Vector Machine (SVM) Method is a linear classification method with find the best working hyperplane as a separator of two classes in the input space. The basic principles of SVM is a linear classifier, then developed into a nonlinear classifier by inserting a trick kernel on high dimensional space.

The kernel function can be formulated by the equation:

$$K(x_i . x_j) = \Phi(x_i) . \Phi(x_j)$$

Kernel Name	Function
Polynomial	$K(X_i \cdot X_j) = (x_i . x_j + 1)^P$
Gaussian RBF	$K(X_i \cdot X_j) = \exp\left(-\left(\frac{\|X_i - X_j\|^2}{2\sigma^2}\right)\right)$
Sigmoid	$K(X_i \cdot X_j) = \tanh(\alpha x_i . x_j + \beta$

Modeling

The classification model used is the SVM model and handling imbalance data by resampling.

After resampling, we use 9,500 data for negative and positive label.

Classification Report	Score
F1-score	91%
Accuracy score	91%
Precision score	95%
Recall score	87%

	Positive	Negative
Positive	1804	96
Negative	240	1660

Testing	Label
Boleh juga ni, walau size ringan	Positive
Bagus nih, bagusnya gak usah di download	Negative

DMAIC

DMAIC

DMAIC (Define, Measure, Analyze, Improve, Control) is a broadly structured problem solving procedure that used in quality and quality improvement processes. It is often associated with six sigma activities, and almost all implementations of six sigma using the DMAIC process.

- **DEFINE** the Customer, their Critical to Quality (CTQ) issues, and the Core Business Process involved.
- **MEASURE** the performance of the Core Business Process involved.

  ANALYZE the data collected and process map to determine root causes of defects and opportunities for improvement.
- **IMPROVE** the target process by designing creative solutions to fix and prevent problems.
- **CONTROL** the improvements to keep the process on the new course.

Business Recomendation

DEFINE

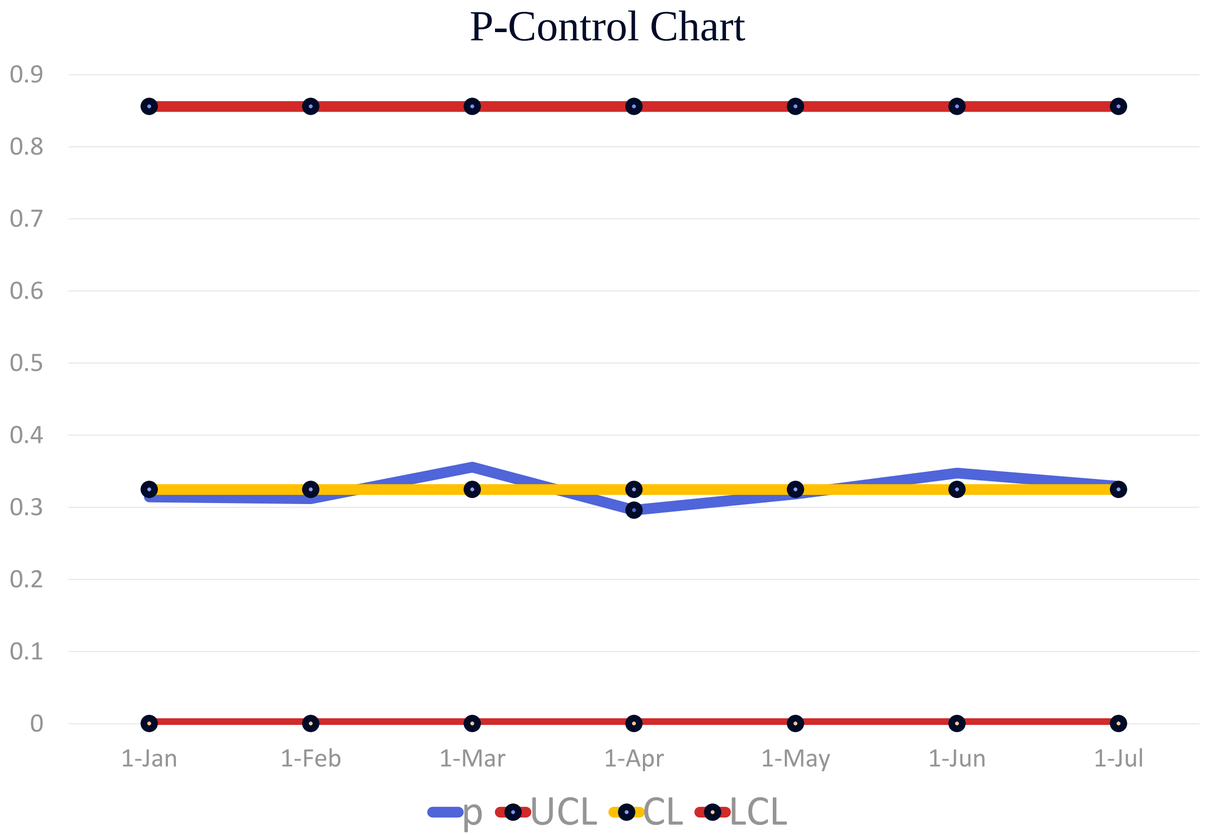
The CTQ (Critical to Quality) that is determined is the user rating score, where a negative value will be defined as a defect.

MEASURE

MonthName	Defect	Total	p	UCL	CL	LCL
Jan21	1273	4048	0.314476	0.855761	0.324771	0
Feb21	1218	3905	0.311908	0.855761	0.324771	0
Mar21	2086	5865	0.355669	0.855761	0.324771	0
Apr21	1131	3820	0.296073	0.855761	0.324771	0
May21	1106	3472	0.318548	0.855761	0.324771	0
Jun21	1587	4567	0.347493	0.855761	0.324771	0
Jul21	1256	3815	0.329227	0.855761	0.324771	0

Business Recomendation

MEASURE



There is no average defect that crosses the upper or lower control limit.

A temporary conclusion can be drawn from the P-control chart, the proportion of defects is still within reasonable limits.

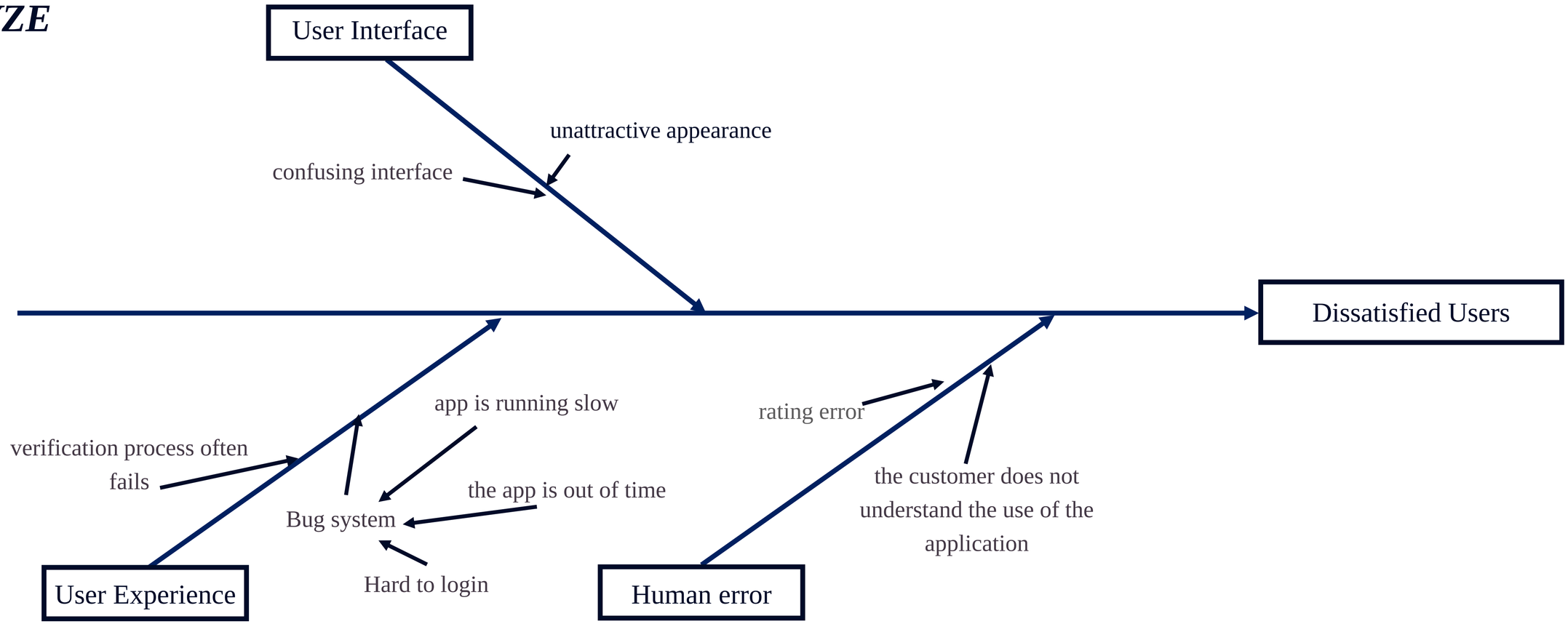
DPMO = 327444.73

Six Sigma values can be seen using the Six Sigma table

with a value of 1.95. Sigma value is **still below** the company standard in Indonesia ($\sigma = 3$).

Business Recomendation

ANALYZE



Business Recomendation

IMPROVE

- User interface improvements, can use A/B testing to choose user prefered user interface.
- Simplify menu choices on the main screen.
- Displaying a tutorial guide for using the application to the first user.
- Fix application bugs.
- Develop a fingerprint scanner system to simplify the verification and login system.

Business Recomendation

CONTROL

The control that is carried out is by paying attention to the p control chart and the sigma value on the DPMO. Incoming review data will be classified first using the classification model that has been created. The classification results are then entered into the calculation of the p and dpmo control charts. The calculation will be updated once a month.

Thank You

LinkedIn
www.linkedin.com/in/imadekrisnajaya

Github
https://github.com/madekrisnaj

Email
madekrisnaj@gmail.com