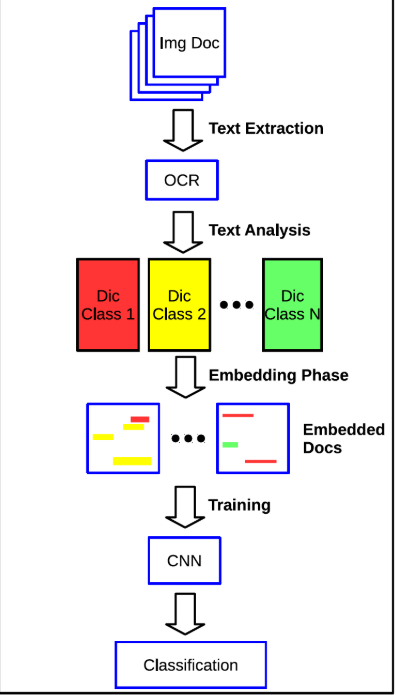
**Predicting Sentiment Analysis and Key Phrases in Documents and Textual Images**

# Aim:

 Most businesses record their data in the form of documents and images. These documents provide useful information about their audience, operations and marketing. Before digitization, the organization used to store information on a piece of paper but now with digitization it has been changed. The data on documents usually consist of natural language, tables, diagram, and so on, or it can be the image of user comments and feedback. Companies are finding it hard to interpret the documents or image containing text. For this purpose, they are moving towards machine learning. Machine learning is a process in which the machine learning in its own and allows the human to solve a complex problem.

The project aims to build an application which finds the insights information of unstructured, semi-structured documents or images and provide sentiment analysis and key-phrases of those documents. The project is divided into two parts.

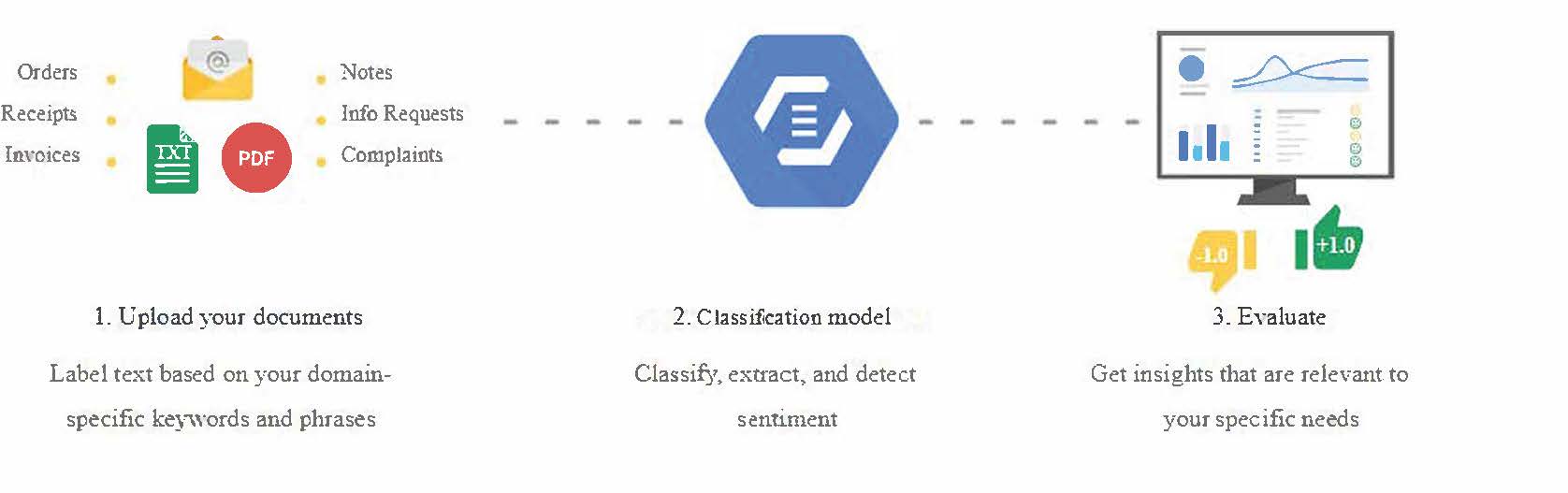
1. To find a way to convert those documents into a machine-readable format.
2. To provide insight information using sentiment analysis.

For the first part, optical character recognition (OCR) will be used to convert the physical document into a machine-readable format. The OCR is a computer vision technique in which a text in a picture is converted into an editable format and provide insights. Later, a conventional neural network to enhance accuracy. A picture can be handwritten, printed document or watermarks. After the image is converted, naïve Bayes classifier will detect the sentiment and keywords in the document. Naïve Bayes is a text classification technique which is widely used in sentiment analysis. For the classification of words, tokenization is used, then a bag of words technique will store words in the model. Then a classifier model will be build using naïve Bayes. Once the model is built, it will test data, and lastly, the system will be deployed. The figure gives the overall information about the project.

# Background:

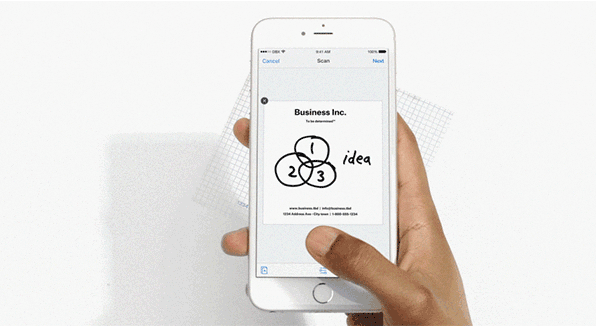
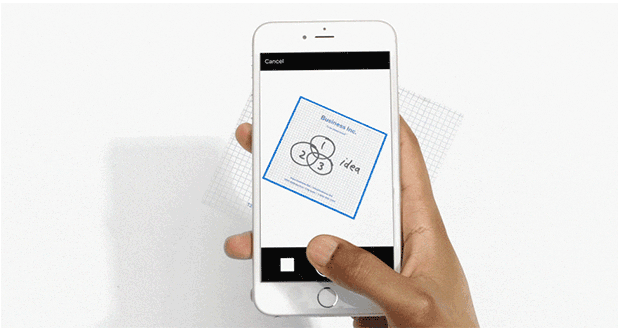
Many research has been done in OCR and sentiment analysis. The OCR is one of the earliest technique in computer vision. According to **Pundlik et al. 2011,**the first OCR for the handheld device was developed by research called Ayatullah Faruk. Later, a new approach was introduced for OCR, which was based on the pattern character algorithm. It used to convert the print document into an editable format. Finally, in 2006, google introduce tesseract, which is an OCR engine, and it was one of the most accurate OCR engines ever. It is able to convert the handwritten and printed document into an editable format. It has the ability to recognize 100 languages, which makes it one of the best OCR engine. On the other hand, the hybrid approach involving machine learning was used to extract sentiment analysis on Facebook messages. They used machine learning classifier to predict the personality of about twenty thousand Facebook user **(Pudaruth et al. 2018)**. The there model was 62% accurate, whereas **bae and lee 2012**, analyzed the impact of the influential people on the twitter followers to understand the emotion of the audience. The proposed project is a combination of OCR and sentiment analysis. The document needs to be converted into a machine readable format before feeding it into the classification model. The existing paper will help us in analysing the tools that we need to solve the proposed problem.

# Research Project:



**Overall process**

Imagine your boss tells you that the company have essential documents for customer feedback which they want to analyse the behaviour of the customer or they have some document which is only in hard copy but not in soft one, and they want to understand the useful information on that particular doc. What will you do? Of course! You will try to read those documents for hours to better understand what is happening and what worse is that most companies have thousands of document and images which people don’t have time to go through each one of them to better understand them. This is where my proposed solution comes in to save the day. The application will convert the single or multiple documents image into an editable format using OCR which user can edit or highlight the important information, and after the conversion is done, it will provide the key phrases and sentiment for that particular document. A lot of application which deals in OCR don’t provide the sentiment and key phrases for the document, which is essential for the companies to know because it gives the insight on the document. The innovation in this project is it provides the insight into the document, which can be anything as long as it is in picture format, and it will provide the key phrases and sentiment.

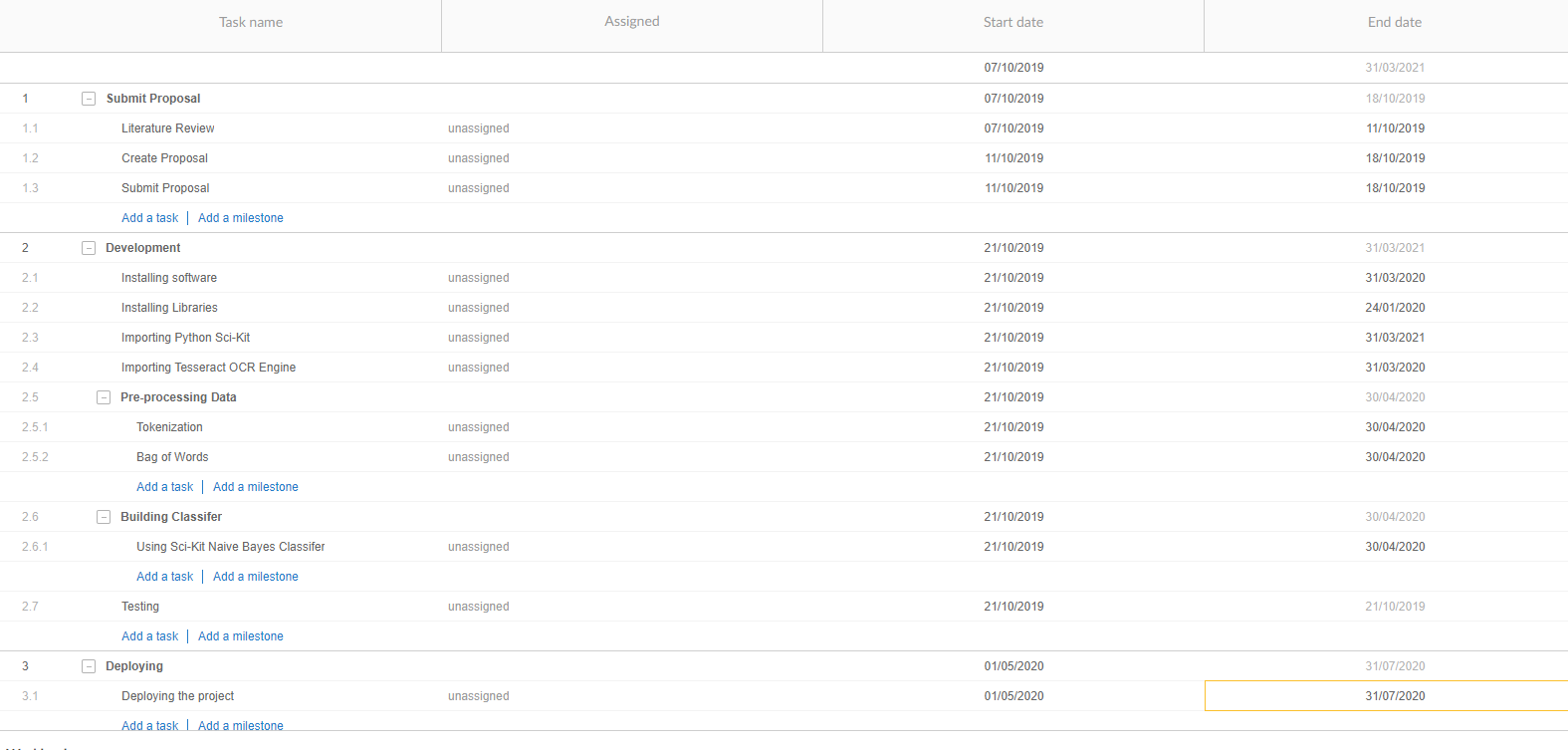


**Before**

**After**

The solution is broken down into serval task:

1. The application will be built for the mobile or desktop using python.
2. Creating datasets which contain textual images.
3. The OCR tesseract engine will be used for the conversion of the image.
4. Pre-processing the documents using tokenization and bag of words technique.
5. Building a classification model using python and Sci-kit learn libraries.
6. Evaluation
7. Sentiment and key phrases.



The Gant chart shows the time it will take to complete this project. The project will be completed within a year. Once the project is approved, it will go towards the development stage, then testing and lastly the project will be deployed. Ideally, the project will convert the image into the editable document, and then with the help of the classification model, it will give insight information such as key phrases and sentiment analysis. The benefits of this project are that it will benefit companies to understand the receipt and invoice, insights of the customers by reading email, feedbacks, and comments. It will reduce the operation cost by 60 per cent. Thus increasing the profit for the company. It is expected that the investor will gain **AUD 114600** profit within a year.

# Budget:

Following is the budget for this project. This is the approximate budget within the project will be build. The budget is for the whole year.

|  |  |
| --- | --- |
| **Type** | **Cost** |
| Data Scientists **(TWO)** | 260,000 AUD => **130,000 each** |
| Developer **(ONE)** | 90,000 AUD |
| Project Manager | 100,000 AUD |
| Software | 30,000 AUD |
| Hardware | 50,000 AUD |
| Office Expense | 35,000 AUD |
| Office 365 Tools | 25,000 AUD |
| Total | 590,000 AUD |

# Personnel:

Following are the people which are required to complete this project.

* **Project Manager:** The PM will make sure that the deliverable is completed on time and within budget. He/she will make the team work together and sort any hurdles.
* **Data Scientists:** The data scientist are the one who will create algorithm for the classification techniques. They are the guru of data science.
* **Developer:** The developer will build the application for the mobile or desktop.

# Reference:

1. McAteer, C. 2019, *Revolutionising Redaction- My Final Year Project*, <<https://towardsdatascience.com/revolutionising-redaction-my-final-year-project-fe664e28ef84>>.
2. Sumitra Pundlik, Varun Rambal, Sonu Tayade, Sonali Ramteke & Suri, P. 2014, 'Towards SENTIEXTRACT: A Combination of OCR and

Sentiment Analysis', *International Journal of Computer Applications*, vol. 94, p. 12.

1. González-Fierro, M. 2017, *A Gentle Introduction To Text Classification And Sentiment Analysis*, <<https://miguelgfierro.com/blog/2017/a-gentle-introduction-to-text-classification-and-sentiment-analysis/>>.
2. Sameerchand Pudaruth, Sharmila Moheeputh, Narmeen Permessur & Chamroo, A. 2018, 'Sentiment Analysis from Facebook Comments using Automatic Coding in NVivo 11', *Advances in Distributed Computing and Articial Intelligence Journal Regular*, vol. 7, pp. 41-8
3. Uzun, Y. 2016, *Keyword Extraction Using Naive Bayes* <<https://pdfs.semanticscholar.org/ad17/49c2f1f386a557d690ca4b3dd1221ea7aefe.pdf>>.

# Video Pitch:

<https://www.youtube.com/watch?v=9muI1eokwTs>