# AUTOMATED AND SEMI-AUTOMATED ARCHITECTURE FOR GATHERING INTELLIGENCE USING DIFFERENT DATASETS.

**A REPORT ON PROJEC BASED LEARNING**

**(SEMESTER – II)**

*Submitted by: -*


AADEESH BALI (21101)

ADITI MULAY (21103)

ANIKET BHAGWAT (21109)

SAMEER BRAMHECHA (21115)

DEVANSH CHOUDHURY (21122)


**COMPUTER ENGINEERING**



**Society for Computer Technology and Research's**

**PUNE INSTITUTE OF COMPUTER TECHNOLOGY**

**DHANKAWADI, PUNE – 43**


**A.Y. 2021-2022**

## - CERTIFICATE-

This is to certify that the work incorporated in the report entitled **"Automated and Semi-automated Architecture for Gathering Intelligence Using Different Datasets."** is carried out by **AADEESH BALI (21101)**, who is a part of Group -1 under the subject **Project Based Learning** during A.Y. **2021-2022**

Such material has not been submitted to any other University/ Institute for any financial support. The literature related to the problem investigated has been appropriately cited and duly acknowledged wherever facilities and suggestions have been availed of.

**Date:** **Name & Sign of Project Guide**

**Place:** PUNE **Prof. Shruti Kudagi**

**Name & Sign of PBL Coordinator** **Name & Sign of Head of Department**

## ACKNOWLEDGEMENT

On this great occasion of accomplishment of our project on "**Automated and Semi-automated Architecture for Gathering Intelligence Using Different Datasets"**, we would like to sincerely express our gratitude to **Prof. Shruti Kudagi Ma'am** who has supported through the completion of this project. Timely guidance of the ma'am inspired us to work on this project. Ma'am gave many different ideas which proved to be very helpful for this project. Ma'am has also inspected our work and made continuous evaluation of the project. We would like to thank ma'am for his remarkable role in completion of our project.

Finally, as one of the team members, I would like to appreciate all my group members for their support and coordination, I hope we will achieve more in our future endeavors.

**Place: Pune** **AADEESH BALI**

# TABLE OF CONTENTS

# LIST OF FIGURES

## NOMENCLATURE

| NOTATION | MEANING |
|---|---|
| HTML | HyperText Markup Language |
| CSS | Cascading Style Sheets |
| NLP | Natural Language Processing |
| ML | Machine Learning |

## INTRODUCTION

For developing a mechanism to protect and secure the cyber world, our little step is towards beginning a process of developing prototypes for guarding not only assets but also civilians. Here we have first worked on automated, and then semi-automated prototypes. Automation, from the word we can understand it can operate freely without any human intervention. Now same goes with the semi-automation, from the word we can understand it is partially automated and have human interference in it. We will see further in the implementation section how these prototypes will perform. Nowadays, a lot of people use social network sites like Facebook, Twitter, Google Plus, LinkedIn, etc. to express their emotions and share views about their daily lives. Through these online communities, we get an interactive media where consumers inform and influence others through the forums. Social media is generating a large volume of sentiment rich data in the form of tweets, status updates, blog posts, comments, reviews, etc. Moreover, social media provides an opportunity for businesses by giving a platform to connect with their customers for advertising. The amount of content generated by users is too vast for a normal user to analyze. So, there is a need to automate this, various sentiment analysis techniques are widely used.

## MOTIVATION

There are many people in this world who are thinking to end their life, due to their negative thinking. There may be some solution on their problem due to which they are going to end their life, but they are not sharing with anyone else. Some of them are posting such type of negative text on twitter. So, we are aiming to help such people identify their sentiments and accordingly it would help them to improve on it.

## SCOPE OF THE PROJECT

Our team aims to create a tangible model that can be adapted to various data sets and be used in an efficient manner. We mainly focus on tweets as a form of data input to sort out various sentiments- positive or negative. Currently, this being implemented on a small scale, we aim to take it even further with the guidance of our mentor Prof. Shruti Kudagi ma'am. The model can be further trained to analyze and gather the required information from different datasets having different origins.

## INTENDED AUDIENCE

Our project mainly targets twitter users and sorts through tweets that might pose a potential indication of deteriorating mental health. Also, this model, if trained further can be used for finding out the target consumers for particular products by different LLC or Government Bodies. The government can use it to track the social media structure of a particular target posing a threat to our national well-being and can also be used to analyze the sentiments of the public over the decision made by the governments.

## LITERATURE SURVEY

### 1. MACHINE LEARNING: -

**Arthur Samuel**, a pioneer in the field of artificial intelligence and computer gaming, coined the term **“Machine Learning”**. He defined machine learning as – **“Field of study that gives computers the capability to learn without being explicitly programmed”**.

In a very layman manner, Machine Learning (ML) can be explained as automating and improving the learning process of computers based on their experiences without being actually programmed i.e. without any human assistance. The process starts with feeding good quality data and then training our machines(computers) by building machine learning models using the data and different algorithms. The choice of algorithms depends on what type of data do we have and what kind of task we are trying to automate.

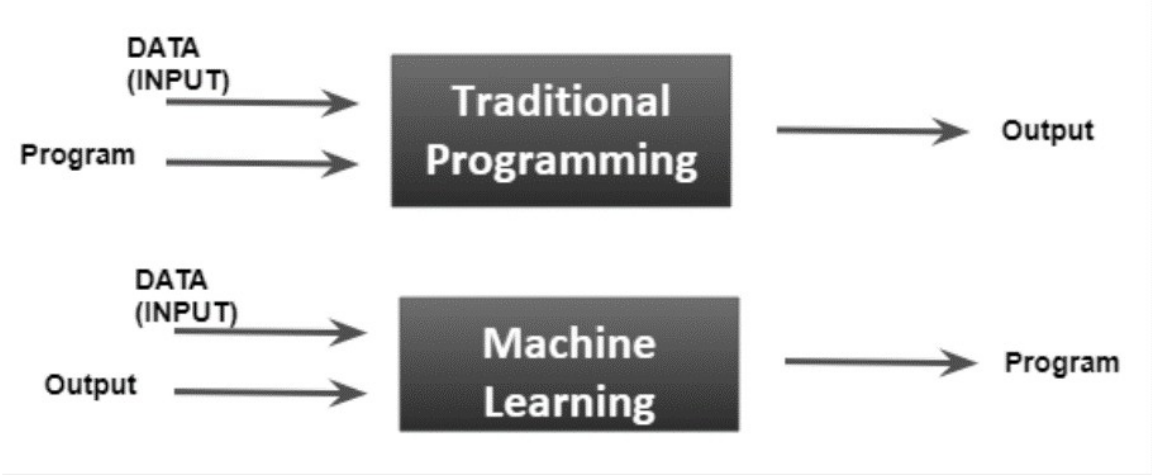


Fig. 1

**How ML works?**

1. Gathering past data in any form suitable for processing. The better the quality of data, the more suitable it will be for modelling.
2. Data Processing – Sometimes, the data collected is in the raw form and it needs to be pre-processed.

3. Divide the input data into training, cross-validation, and test sets. The ratio between the respective sets must be 6:2:2.
4. Building models with suitable algorithms and techniques on the training set.
5. Testing our conceptualized model with data which was not fed to the model at the time of training and evaluating its performance using metrics such as F1 score, precision, and recall.

## 2. NATURAL LANGUAGE PROCESSING: -

Natural language processing (NLP) refers to the branch of computer science —and more specifically, the branch of artificial intelligence or AI— concerned with giving computers the ability to understand text and spoken words in much the same way human beings can. NLP combines computational linguistics—rule-based modeling of human language—with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to ‘understand’ its full meaning, complete with the speaker or writer’s intent and sentiment. NLP drives computer programs that translate text from one language to another, respond to spoken commands, and summarize large volumes of text rapidly—even in real time.

Applications of NLP: -

1. Translating Languages.

2. Speech Recognition.

3. Sentiment Analysis.

4. Chatbots.

5. Automatic Text Classification.

**3. SENTIMENTAL 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. With the recent advances in deep learning, the ability of algorithms to analyse text has improved considerably. Creative use of advanced artificial intelligence techniques can be an effective tool for doing in-depth research.

# OPERATIONAL ENVIRONMENT

**SOFTWATE REQUIREMENTS: -**

**Python 3.9** – Programming Language.

**Anaconda** – A distribution of python programming language for scientific computing.

**PyCharm** – Integrated Development Environment (IDE).

**Jupyter Notebook** – A software that allows to creation and sharing of documents that integrate live code.

**Google Collaboratory** – Google Collaboratory is a free Jupyter notebook environment that runs entirely in the cloud. Most importantly, it does not require a setup and the notebooks that you create can be simultaneously edited by your team members - just the way you edit documents in Google Docs.

**Visual Studio Code** - A code editor to be used for Web Development using HTML and CSS.

Front- End: **-** HTML, CSS.

Back - End: - Flask Framework (Python)

## FLOWCHART

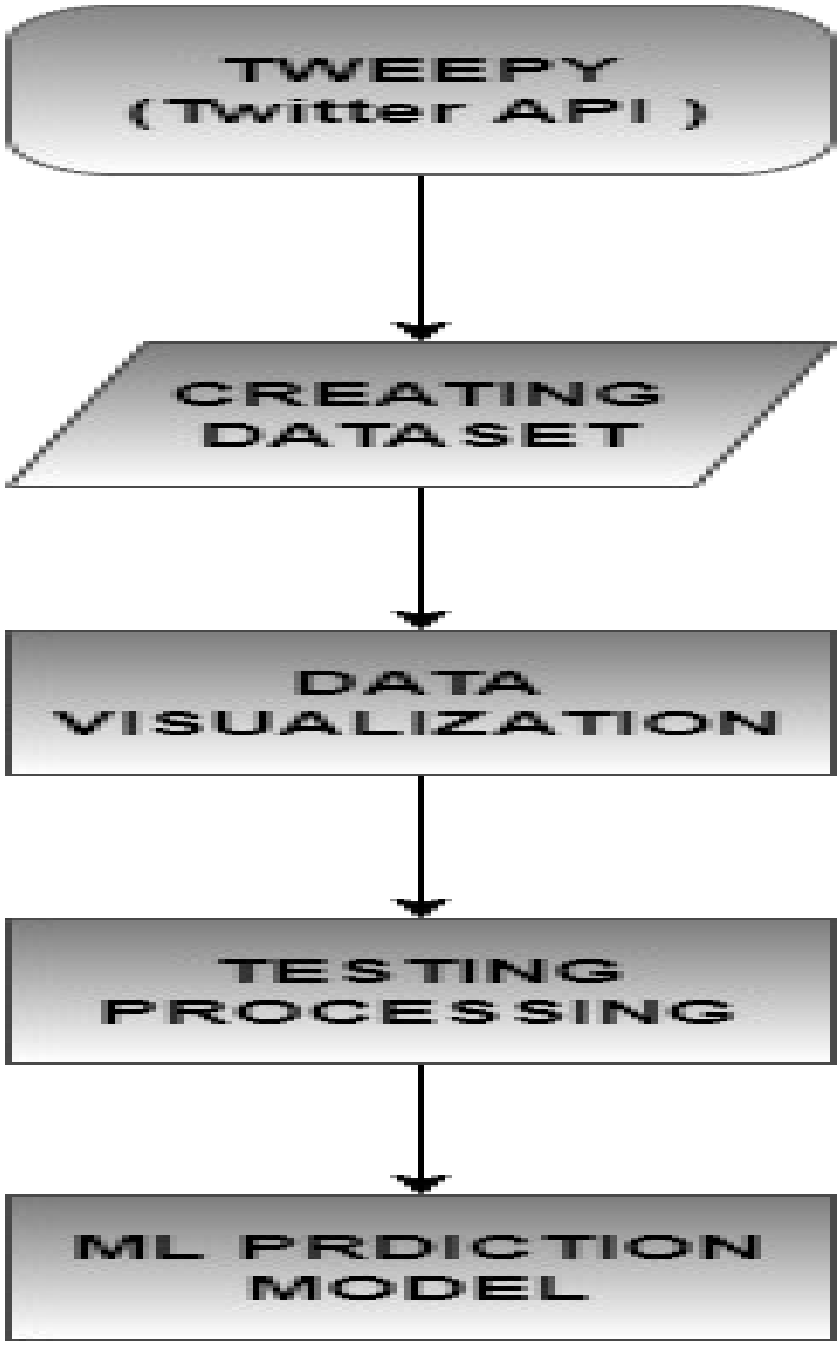


Fig. 2

# UML DIAGRAM

## 1. USE CASE DIAGRAM

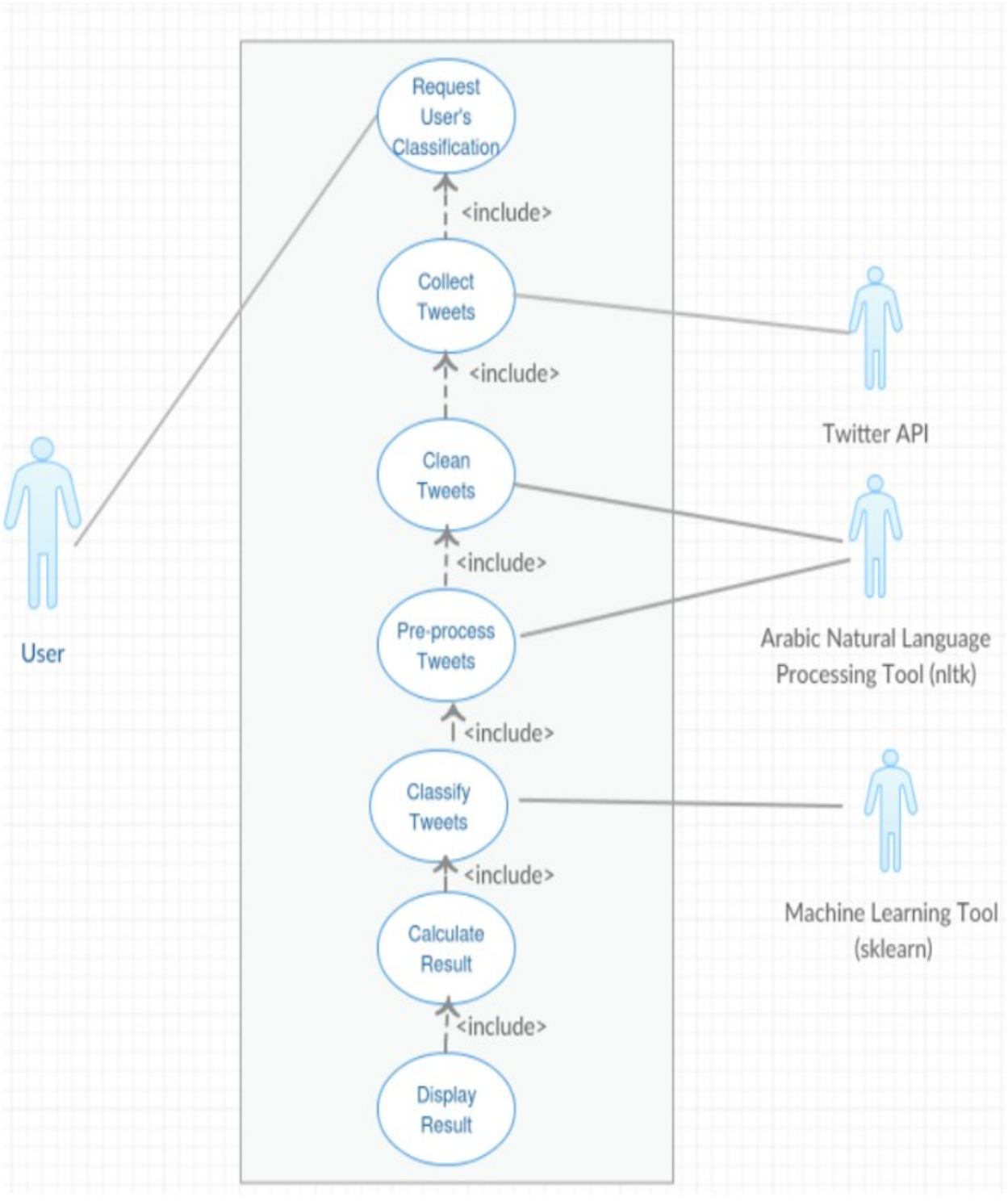


Fig. 3

## 2. SYSTEM ARCHITECTURE

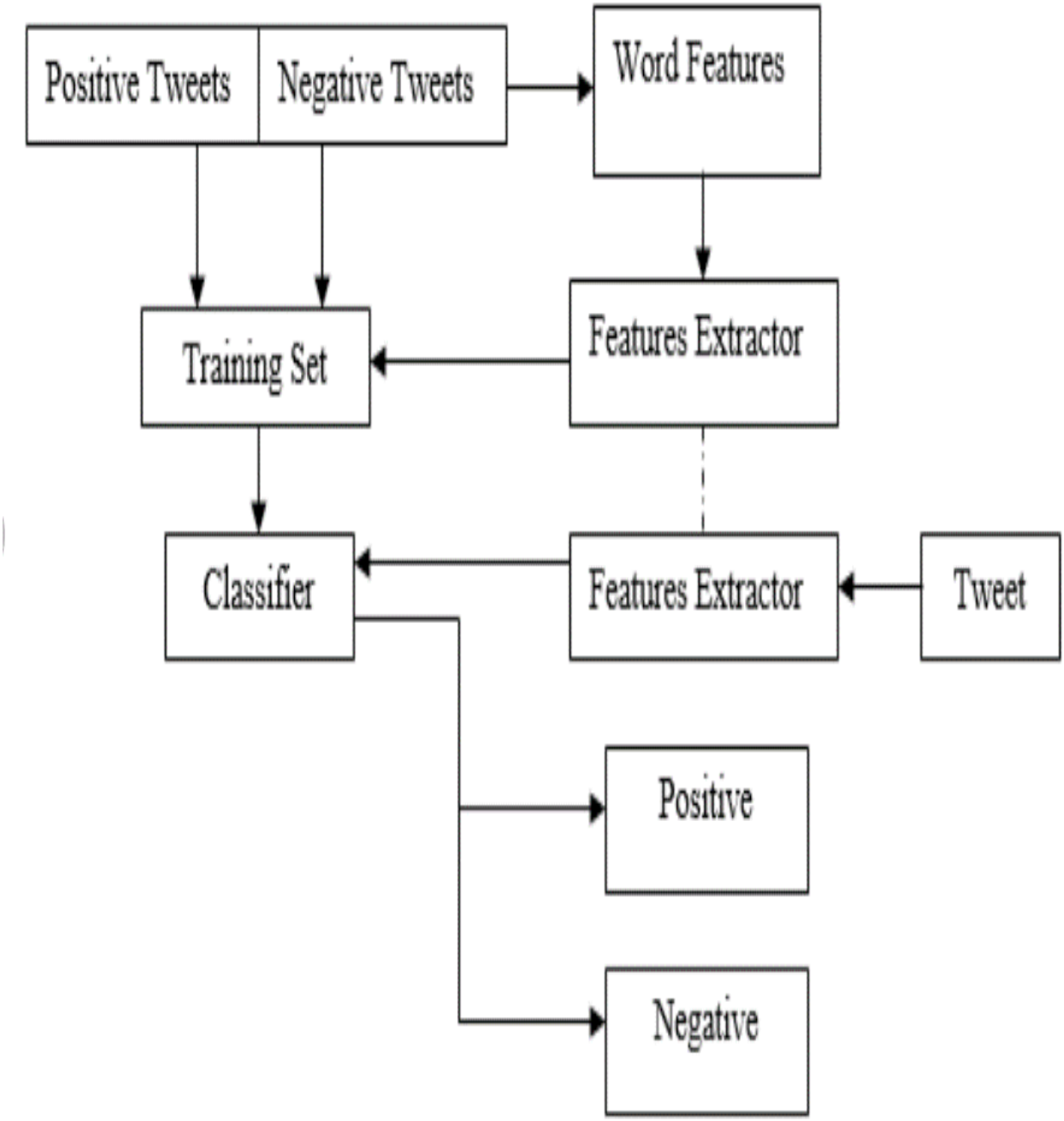


Fig. 4

# IMPLEMENTATION DETAILS

The basic step towards understanding the projects is having basic knowledge of all the Machine-Learning inclined libraries in python.

**Pandas** - Pandas is an open-source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named NumPy, which provides support for multi-dimensional arrays.

**NumPy** - It supplies an enormous library of high-level mathematical functions that operate on these arrays and matrices.

**Spacy** - spaCy is designed specifically for production use and helps you build applications that process and "understand" large volumes of text. It can be used to build information extraction or natural language understanding systems, or to pre-process text for deep learning.

**Beautifulsoup** - Beautiful Soup is a Python library that is used for web scraping purposes to pull the data out of HTML and XML files. It creates a parse tree from page source code that can be used to extract data in a hierarchical and more readable manner.

**Text Preprocessing: -**

```
def get_clean(x):
    x = str(x).lower().replace('\\', '').replace('_', ' ')
    x = ps.cont_exp(x)
    x = ps.remove_emails(x)
    x = ps.remove_urls(x)
    x = ps.remove_html_tags(x)
    x = ps.remove_rt(x)
    x = ps.remove_accented_chars(x)
    x = ps.remove_special_chars(x)
    x = re.sub("(.)\\1{2,}", "\\1", x)
    return x
```

Fig. 5

The get_clean () function: -

It prepares the data and converts it into a model readable form taking out all the special characters, emoticons, emails, links, etc.

Converting the whole data into lowercase for uniformity.

Anonymous/Lambda Function: -

In computer programming, an anonymous function is a function definition that is not bound to an identifier. Anonymous functions are often arguments being passed to higher-order functions or used for constructing the result of a higher-order function that needs to return a function.

**Storing Data in Optimized Form**: -

**Term Frequency-Inverse Document Frequency**: -

The term frequency is the number of occurrences of a specific term in a document. Term frequency indicates how important a specific term in a document. Term frequency represents every text from the data as a matrix whose rows are the number of documents and columns are the number of distinct terms throughout all documents.

Inverse document frequency (IDF) is the weight of a term, it aims to reduce the weight of a term if the term's occurrences are scattered throughout all the documents.

Storing the whole data in the two vectors X & Y

X – Storing Tweets.

Y – Storing Intentions.

Working: -

Using the skLearn library the model is comparing the dataset to the top 20000 words of universal dictionary and filtering out as positive feelings with respect to the user.

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.svm import LinearSVC
from sklearn.metrics import classification_report
```

```
tfidf = TfidfVectorizer(max_features=20000, ngram_range=(1, 3), analyzer='char')
```

Fig. 6

The objective of a Linear SVC (Support Vector Classifier) is to fit to the data you provide, returning a "best fit" hyperplane that divides, or categorizes, your data. From

there, after getting the hyperplane, you can then feed some features to your classifier to see what the "predicted" class is. This makes this specific algorithm rather suitable for our uses, though you can use this for many situations.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
clf = LinearSVC()
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)
```

Fig. 7

## Flask Framework (Python): -

Flask is a web application framework written in Python. Flask is based on the Werkzeug WSGI toolkit and Jinja2 template engine.

WSGI: - The Web Server Gateway Interface (Web Server Gateway Interface, WSGI) has been used as a standard for Python web application development. WSGI is the specification of a common interface between web servers and web applications.

Werkzeug: - Werkzeug is a WSGI toolkit that implements requests, response objects, and utility functions. This enables a web frame to be built on it. The Flask framework uses Werkzeg as one of its bases.

Jinja2: - Jinja2 is a popular template engine for Python. A web template system combines a template with a specific data source to render a dynamic web page.

## RESULTS

Accuracy of the model: -

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.94 | 0.93 | 0.93 | 1060 |
| 1 | 0.91 | 0.91 | 0.91 | 764 |
| accuracy | | | 0.92 | 1824 |
| macro avg | 0.92 | 0.92 | 0.92 | 1824 |
| weighted avg | 0.92 | 0.92 | 0.92 | 1824 |

Fig. 8

Output: -

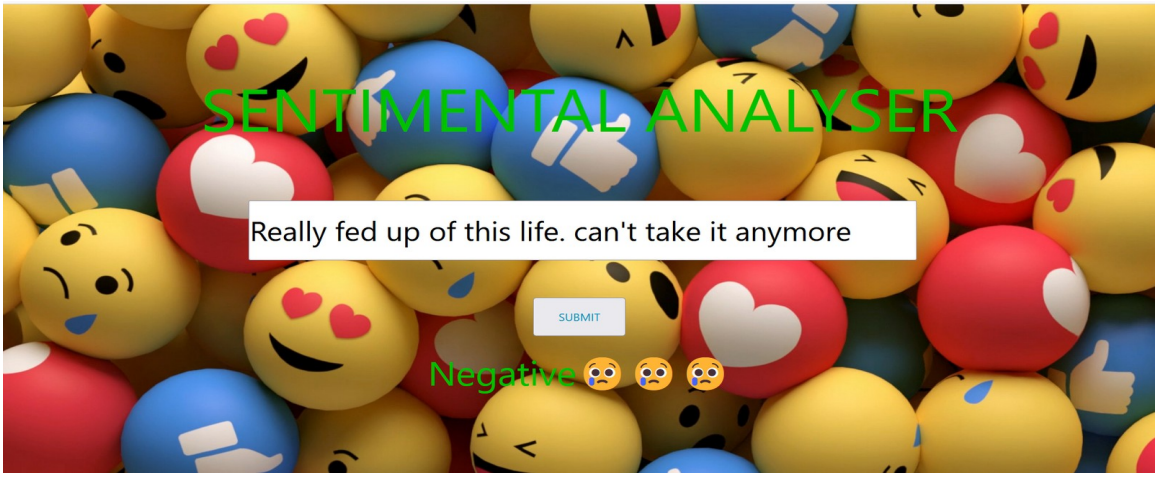


Fig. 9

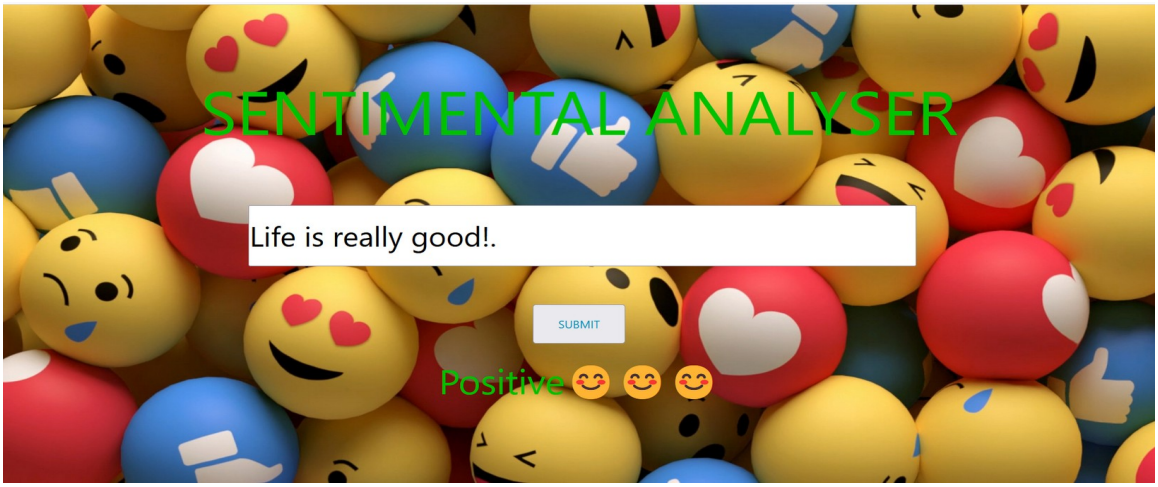


Fig. 10

## CONCLUSION

In our project, we have made a web application where in a user can enter his/her sentiments and our ML model would classify the sentiment as positive or negative. The accuracy of our ML model is 92%. In future, we aim to perform similar analysis in order to increase the accuracy of the model.

## REFERENCES


1. https://www.youtube.com/playlist?list=PLc2rvfiptPSToz3K_ozo7zrMJXqe16YUd
2. https://www.youtube.com/watch?v=NWONeJKn6kc&t=31s
3. https://www.youtube.com/watch?v=dIUTsFT2MeQ&t=35s