Portfolio Presentation

Robins Yadav

About Me

"Apply the skills that I acquired over the past 6 years of my Engineering education to real life problems"

Master of Science in Electrical and Computer Engineering
The University of Arizona, Tucson, AZ
CGPA: 4.0/4.0
Class of 2020

Bachelor of Science in Electrical Engineering
Boise State University, Boise, ID
CGPA: 3.7/4.0
Class of 2018

Data Science Workflow

Data Science Workflow

- Series of steps that you take to complete a data science project

Define

1. Understand and define the problem

Discover

2. Obtain data
3. Clean data
4. Explore data
5. Establish baseline outcomes
6. Hypothesize solutions

Develop

7. Engineer features
8. Create models
9. Test models
10. Select best models

Deploy

11. Automate pipeline
12. Deploy solution
13. Measure efficacy

Note: If the outcome does not meet the customer requirements, go back to step 1.

Human Activity Recognition

Human Activity Recognition

- **Problem Statement:** To understand if a person carrying a smartphone is performing exercises like Downstairs, Jogging, Sitting, Standing, Upstairs, Walking.
- **Solution:** Based on the available WISDM data set, the CNN (Convolution Neural Network) will learn how to differentiate between each of the exercises. We can then show new data to the neural network and it will tell us what the user is doing at any particular point in time.
- **Baseline Accuracy is** 93.32% with 10s window**:** Based on the paper, “Real-time human activity recognition from accelerometer data using Convolutional Neural Networks”

  Reference: Ignatov, A. (2018). Real-time human activity recognition from accelerometer data using Convolutional Neural Networks. *Applied Soft Computing*, *62*, 915-922.

- **WISDM dataset:** Accelerometer data obtained from smartphone

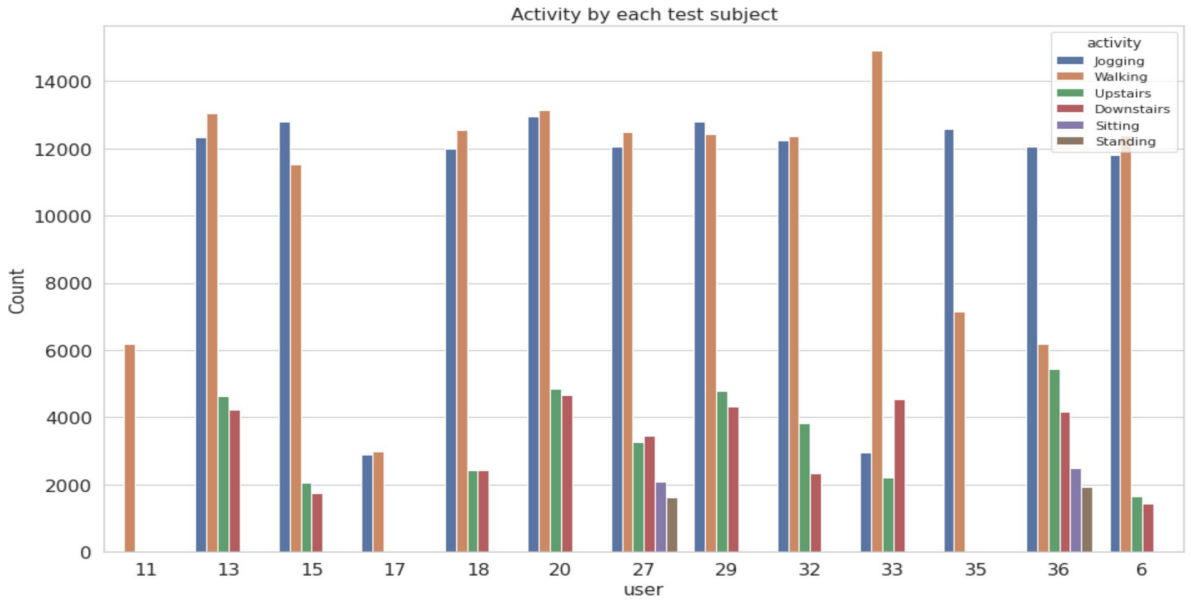
```
33,Jogging,49105962326000,-0.6946377,12.680544,0.50395286;
33,Jogging,49106062271000,5.012288,11.264028,0.95342433;
33,Jogging,49106112167000,4.903325,10.882658,-0.08172209;
33,Jogging,49106222305000,-0.61291564,18.496431,3.0237172;
33,Jogging,49106332290000,-1.1849703,12.108489,7.205164;
33,Jogging,49106442306000,1.3756552,-2.4925237,-6.510526;
```

Raw data – Unstructured Data

	user	activity	time	x	y	z
0	33	Jogging	49105962326000	-0.6946377	12.680544	0.50395286
1	33	Jogging	49106062271000	5.012288	11.264028	0.95342433
2	33	Jogging	49106112167000	4.903325	10.882658	-0.08172209
3	33	Jogging	49106222305000	-0.61291564	18.496431	3.0237172
4	33	Jogging	49106332290000	-1.1849703	12.108489	7.205164

Structured Data

- **Visualization:** Activities by each with the number of data records



Reference: GitHub: Human Activity Recognition

Human Activity Recognition

- **Accelerator Data:** data is recorded at a sampling rate of 20 Hz (20 values per second)

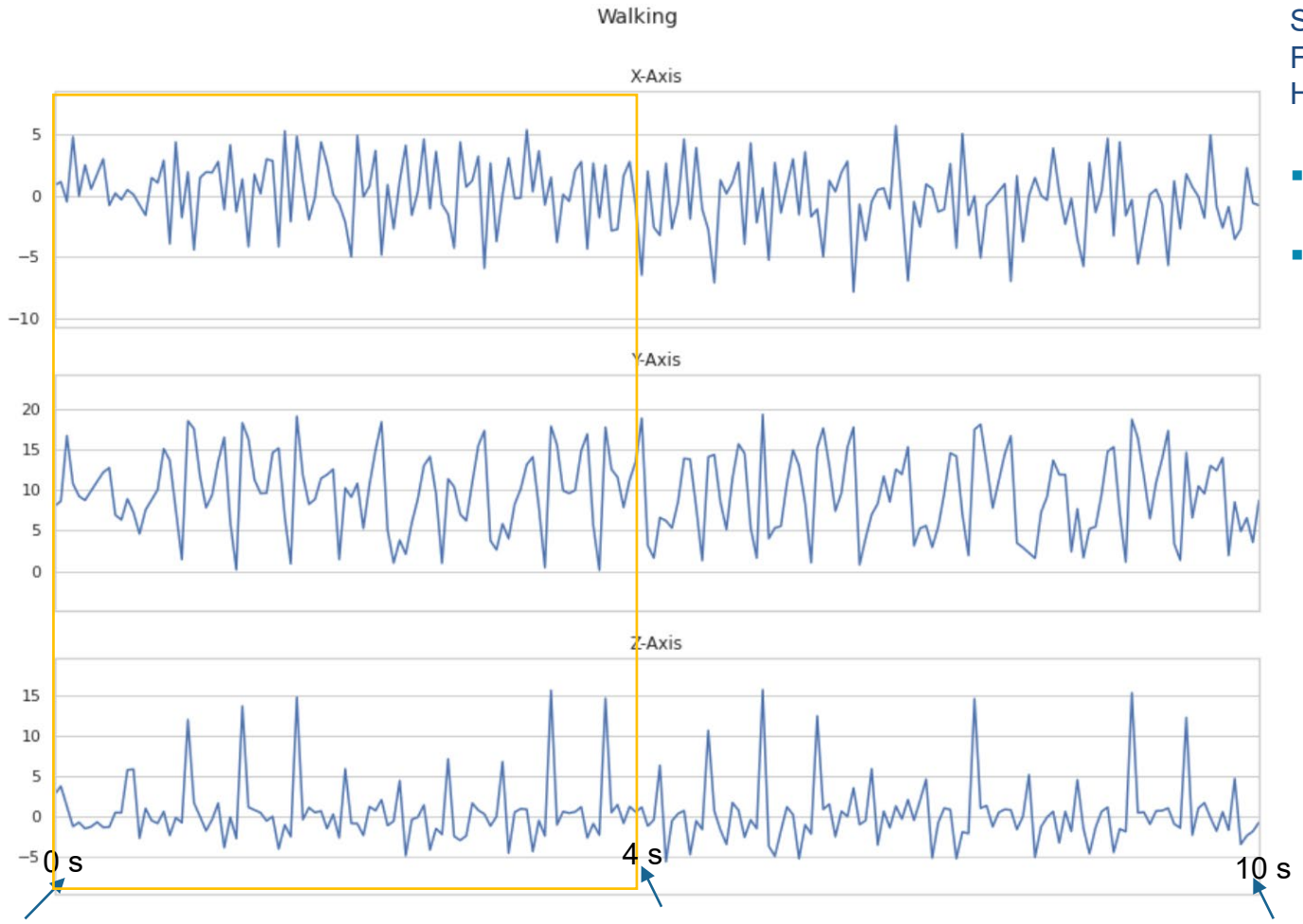


Figure: Graph shows the 200 records (i.e. for 10 seconds) of walking activity

- **Keras CNN model accept 2-D input data**

Sample Freq, Fs = 20
Frame Size = Fs*4 , here 4 seconds → 80
Hop Size = Fs*2, here 2 seconds → 40 records

- **Training Data:** (A, 80, 3), Here A is no. of 2D frame

- **Results:**

	precision	recall	f1-score	support
0	0.85	0.83	0.84	167
1	0.98	0.98	0.98	647
2	1.00	1.00	1.00	23
3	1.00	1.00	1.00	17
4	0.82	0.80	0.81	176
5	0.97	0.98	0.97	687
accuracy			0.95	1717
macro avg	0.94	0.93	0.93	1717
weighted avg	0.95	0.95	0.95	1717

As you can see, the precision and the recall of the model are good for predicting jogging (1), sitting (2), standing (3), and walking (5). The model has problems for clearly identifying upstairs and downstairs activities.

Reference: GitHub: Human Activity Recognition

Hotel Booking Cancellations

Hotel Booking Cancellations

- **Problem Statement:** According to a research conducted in 2008, on an average hotel booking cancellations have reached almost 40% in Europe
- **Solution:** Create a flexible and scalable model to predict the hotel booking cancelations based on the dataset available to minimize the revenue leakage in the industry .
- **Baseline Accuracy is** 72.51% (Based on Majority Class)

- **Exploratory Data Analysis:**

Data Information

- Obtain Data from Kaggle
- 120 K records and 32 features
- Data formatting
- NA or Null values, duplicates, corrupted data

Data Analysis

- Numerical continuous features: bar plot and distribution plot
- Numerical discrete and categorical features: bar plot & value_counts()
- Imputation of missing values

Feature Engineering

- Correlation Matrix and Heat-map
- Recursive Feature Selection
- Feature Importance (ExtraTreesClassifier())
- Chi Square Test & Fisher Score

Reference: GitHub: Hotel Booking Demand – EDA

Reference: GitHub: Hotel Booking Demand - Modeling and Tunning

Hotel Booking Cancellations

- **Preprocessing, Model Training & Tuning:**

01 Encoding and Scaling

- One-hot encoding - Categorical
- Ordinal encoding - Categorical
- Standardization - Numeric

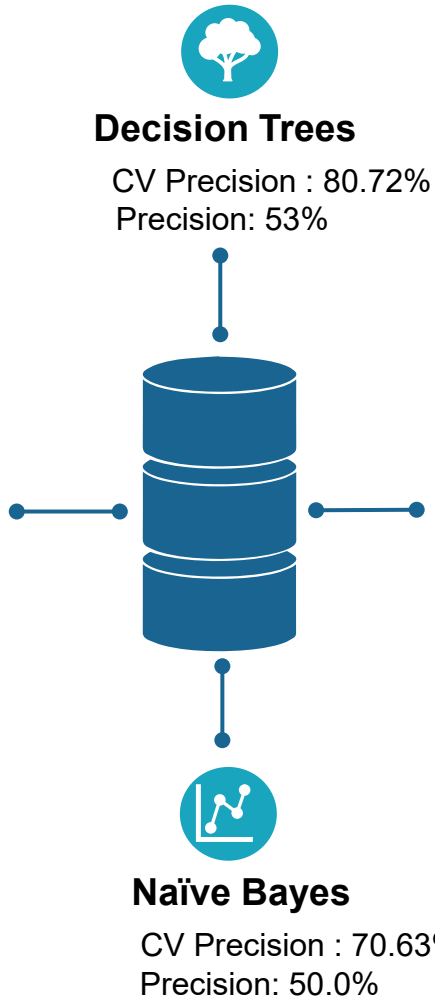
02 Cross-Validation

- Evaluate the ML models:
  Decision Tree,
  Logistic Regression,
  Gaussian Naïve Bayes,
  Multilayer Perceptron

03 Tune the Best ML Model

- Grid Search CV

- **Results:**



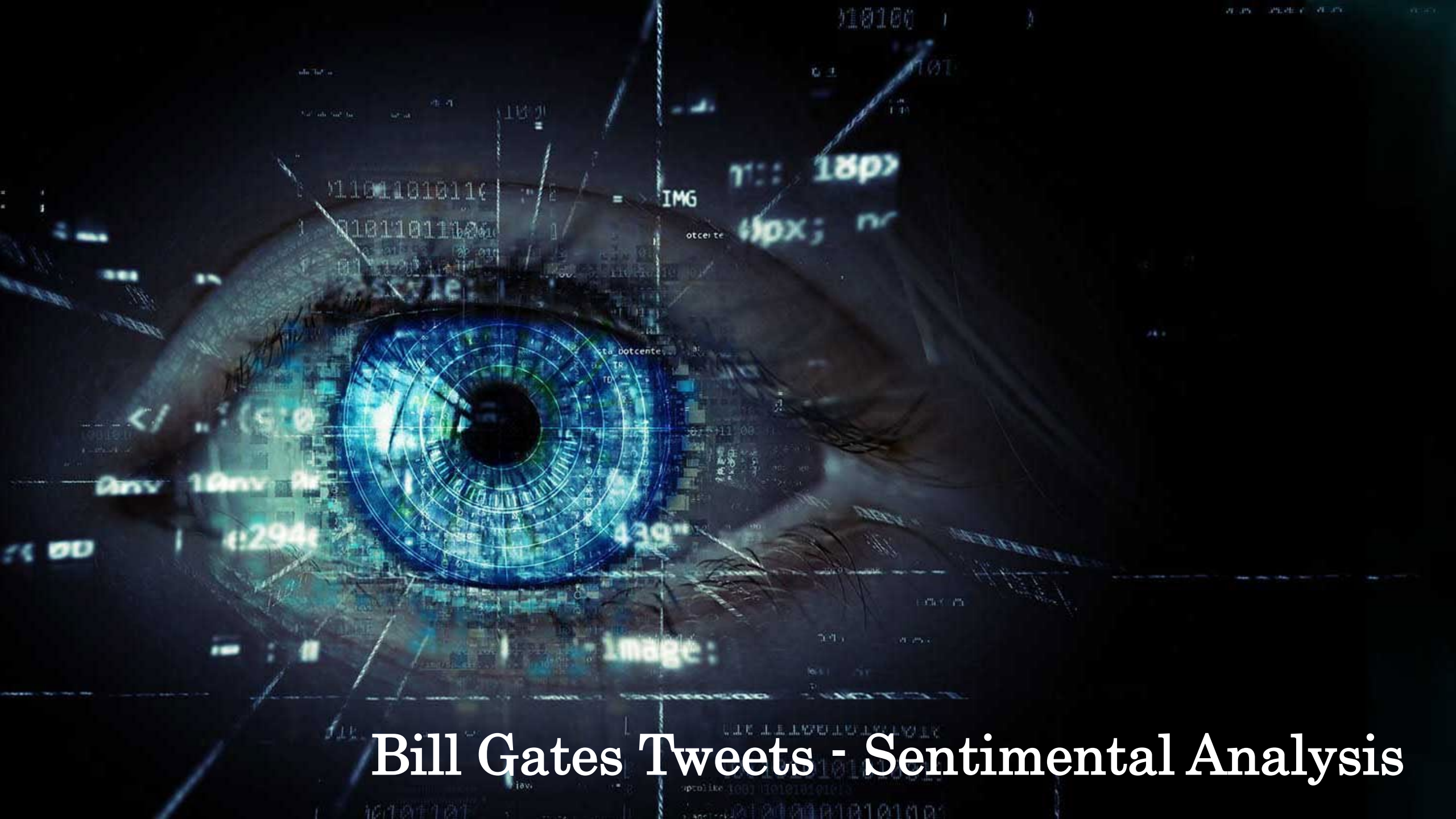
Decision Trees
CV Precision : 80.72%
Precision: 53%

Multi-layer Perceptron
CV Precision : 78.09%
Precision: 71.0%

Logistic Regression
CV Precision : 70.95%
Precision: 70%

Naïve Bayes
CV Precision : 70.63%
Precision: 50.0%

Reference: GitHub: Hotel Booking Demand - Modeling and Tunning



Bill Gates Tweets - Sentimental Analysis

Bill Gate Tweets

- **Problem Statement:** To investigate the Bill Gates tweets, and understand whether his tweets are positive, negative or neutral.
- **Solution:** The sentimental analysis which interprets and classifies the emotions (positive, negative and neutral) within text data using natural language processing.

Workflow

Data: Tweets

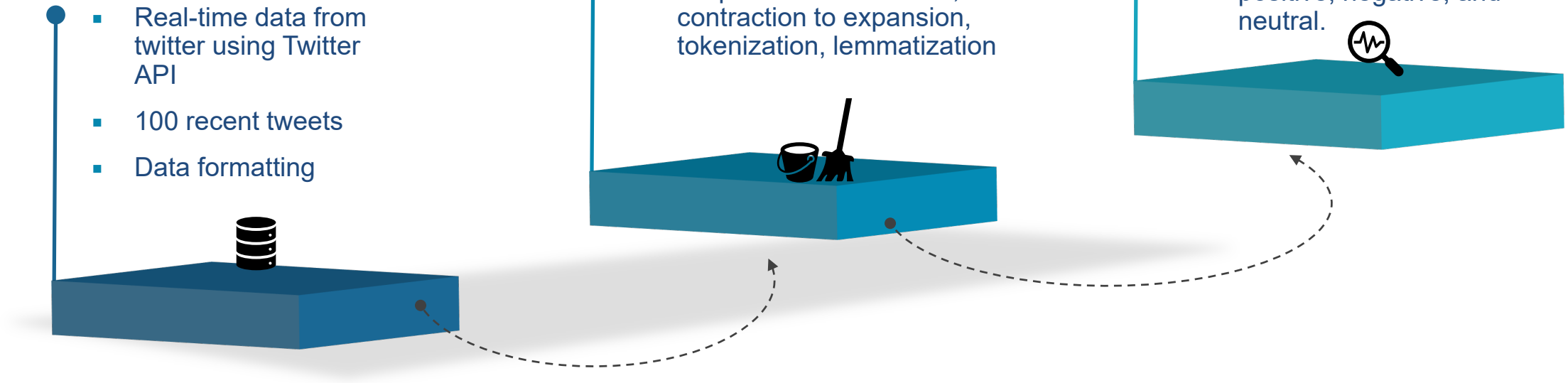
- Real-time data from twitter using Twitter API
- 100 recent tweets
- Data formatting

Tweets Cleaning & Preprocessing

- Removed emails, urls, RT, html tags, accented characters, multiple spaces.
- Preprocess: lower case, contraction to expansion, tokenization, lemmatization

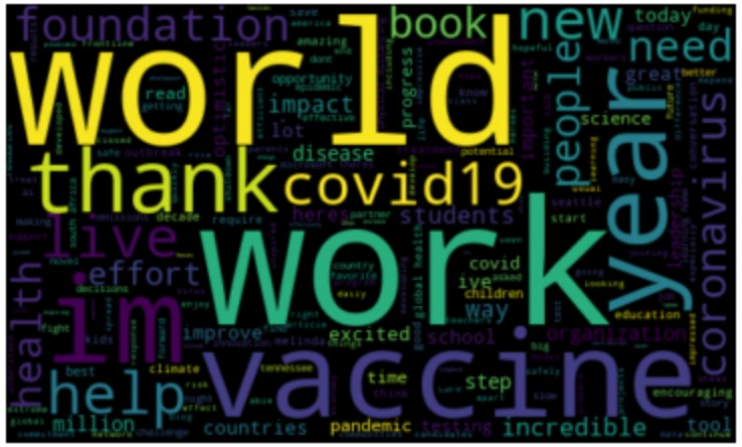
Tweets Analysis

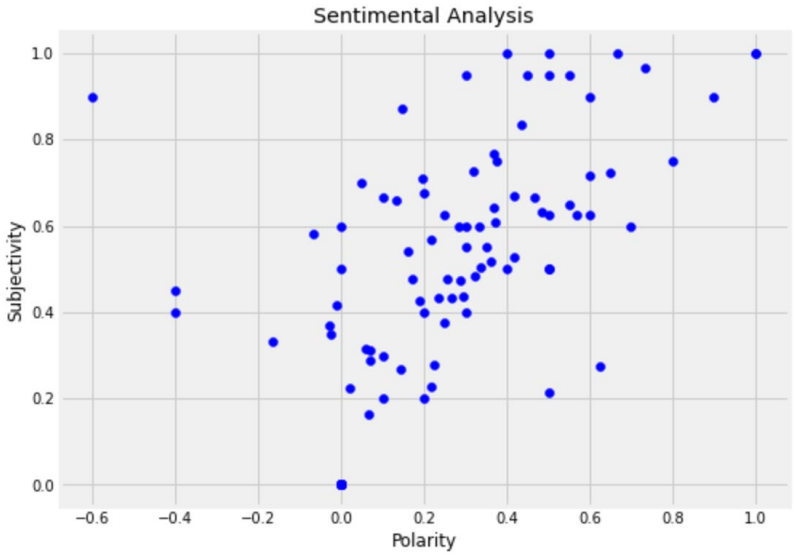
- Subjectivity and Polarity of the tweets using TextBlob library.
- Visualization of word cloud
- Sentimental Analysis: positive, negative, and neutral.

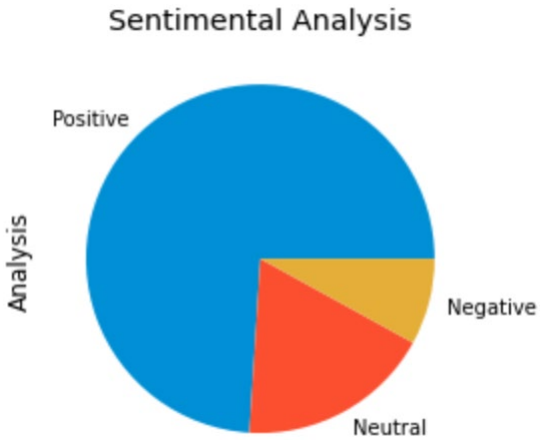


Reference: Github: https://github.com/robinyUArizona/Data-Science-NLP-ML-Projects/blob/master/Bill%20Gates%20Twitter%20Sentimental%20Analysis/2-NLP%20--%20Bill%20Gates%20Twitter%20Sentimental%20Analysis.ipynb

Results







Reference: Github: https://github.com/robinyUArizona/Data-Science-NLP-ML-Projects/blob/master/Bill%20Gates%20Twitter%20Sentimental%20Analysis/2-NLP%20--%20Bill%20Gates%20Twitter%20Sentimental%20Analysis.ipynb

Comedian Ali Wong - Topic Modelling

Comedian Ali Wong

- **Problem Statement:** What makes Ali Wongs' comedy routine stand out?
- **Solution:** By analyzing the sentiment and topic modelling with NLP Techniques.

Sentimental analysis interprets and classifies the emotions (positive, negative and neutral) in words and sentences spoken by Ali Wong using natural language processing.

Topic modelling: find themes across various comedians' routines

Discover

Data

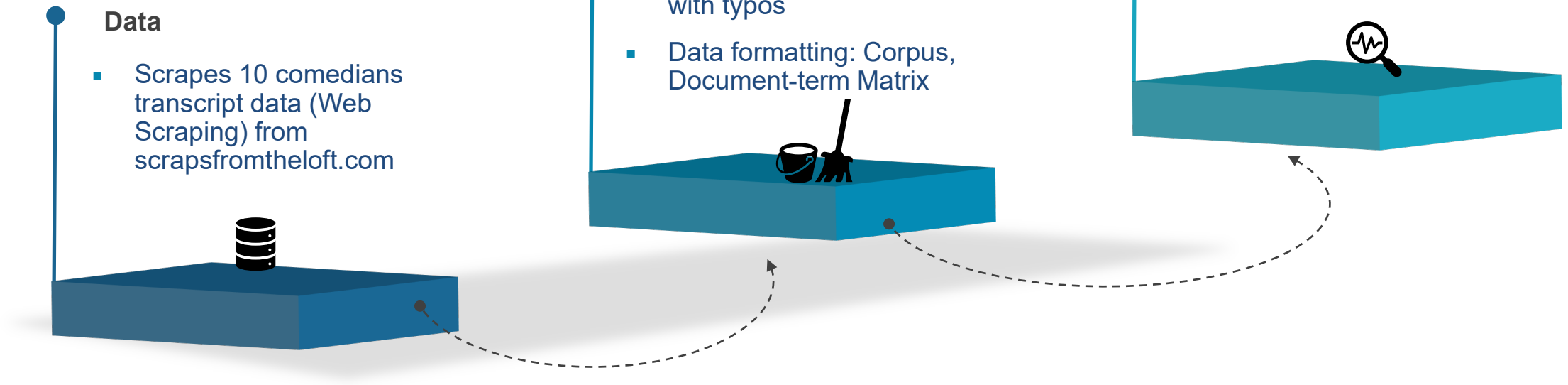
- Scrapes 10 comedians transcript data (Web Scraping) from scrapsfromtheloft.com

Data Cleaning & Preprocessing

- Removed punctuation, numerical values, non-sensical text, stop words
- Made text lowercase, tokenize text, stemming/lemmatization, create bi-grams/tri-grams, deal with typos
- Data formatting: Corpus, Document-term Matrix

Exploratory Data Analysis

- Top 30 words spoken by comedians
- Customized stopwords
- WordCloud Visualization with unique words spoken by comedians
- Number of words by comedians



Reference: Github: https://github.com/robinyUArizona/NLP-Project

Visualization from EDA

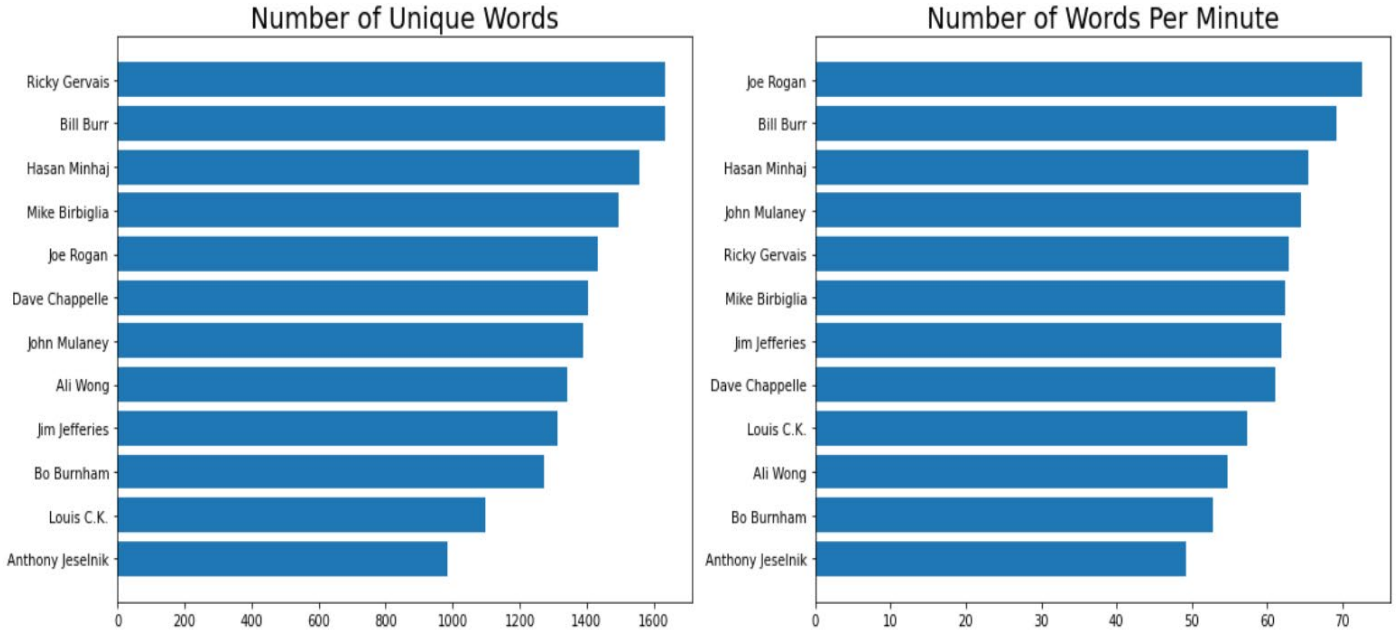


Figure: Vocabulary of comedians

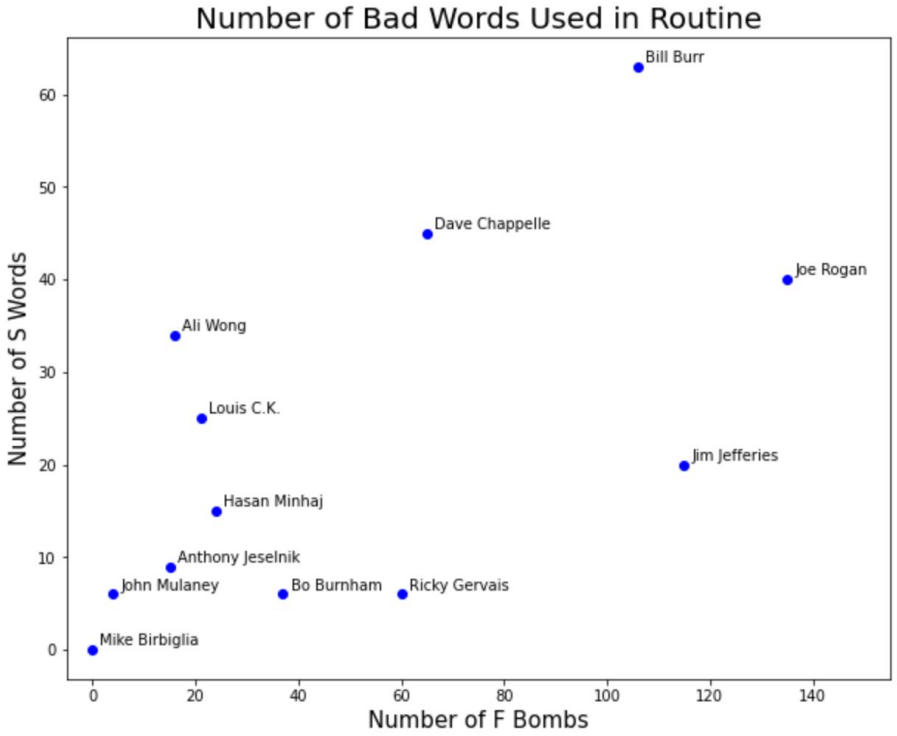


Figure: No. of bad words by comedians

Note: The goal of EDA is to take an initial look at the data and see if the results of basic analysis made sense.

Reference: Github: https://github.com/robinyUArizona/NLP-Project/blob/master/2-Exploratory%20Data%20Analysis.ipynb

NLP Techniques

Sentimental Analysis

- **Input:** A corpus, because order matters. "great" = positive, "not great" = negative.
- **TextBlob:** TextBlob (Python library – nltk) finds all of the words and phrases that it can assign a polarity and subjectivity to, and **averages** all of them together
- **Output:** For each comedian, a sentiment score (how positve/negative are they) and a subjectively score (how opinionated are they).

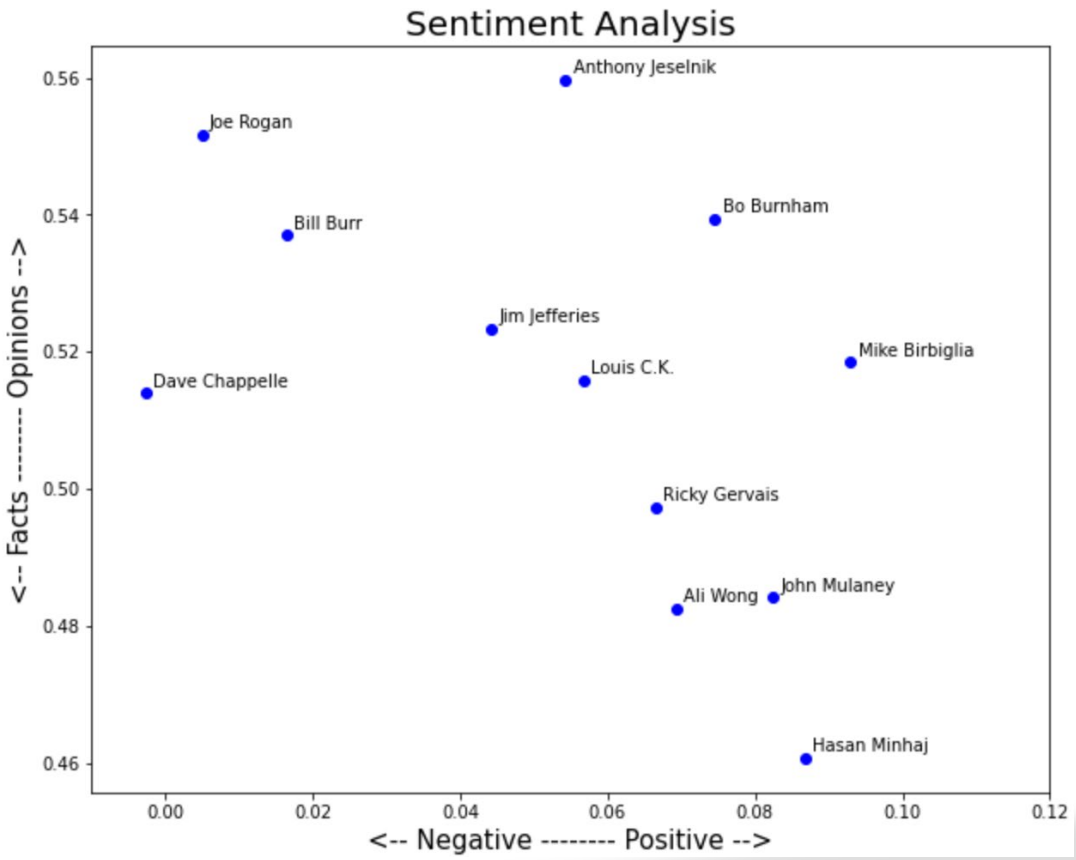
Topic Modelling

- **Input:** A document-term matrix. Each topic will consist of a set of words where order does not matter
- genism: genism is a Python toolkit.
- **Modelling technique:** Latent Dirichlet Allocation (LDA) - Each document can be described by a distribution of topics and each topic can be described by a distribution of words
- **Output:** to find themes across various comedy routines, and see which comedians tend to talk about which themes.

Reference: Github: https://github.com/robinyUArizona/NLP-Project/blob/master/3-Sentiment%20Analysis.ipynb

Reference: Github: https://github.com/robinyUArizona/NLP-Project/blob/master/4-Topic%20Modelling.ipynb

Results



Topic Modelling Result:

For a first pass of LDA, these kind of make sense.
Topic 0: mom, parents [Anthony, Hasan, Louis, Ricky]
Topic 1: husband, wife [Ali, John, Mike]
Topic 2: guns [Bill, Bo, Jim]
Topic 3: profanity [Dave, Joe]

Reference: Github: https://github.com/robinyUArizona/NLP-Project/blob/master/3-Sentiment%20Analysis.ipynb

Reference: Github: https://github.com/robinyUArizona/NLP-Project/blob/master/4-Topic%20Modelling.ipynb

GitHub

https://github.com/robinyUArizona

- **GitHub Repository:** It consist of projects in multiple domains: Time-Series Forecasting, Natural Language Processing, Hospitality industry, Machine Learning Algorithms etc.


THANK YOU